Causal Inference using Multivariate Generalized Linear Mixed-Effects Models with Longitudinal Data

Yizhen Xu, Jisoo Kim, Ami Shah, Laura Hummers, Scott Zeger

February 6, 2023

Abstract

Dynamic prediction of causal effects under different treatment regimes conditional on individual's characteristics and longitudinal history is an essential problem in precision medicine. This is challenging in practice because outcomes and treatment assignment mechanisms are unknown in observational studies, an individual's treatment efficacy is a counterfactual, and the existence of selection bias is often unavoidable.

We propose a Bayesian framework for identifying subgroup counterfactual benefits of dynamic treatment regimes by adapting Bayesian g-computation algorithm[26;41] to incorporate multivariate generalized linear mixed-effects models. Unmeasured time-invariant factors are identified as subject-specific random effects in the assumed joint distribution of outcomes, time-varying confounders, and treatment assignments. Existing methods mostly assume no unmeasured confounding and focus on balancing the observed confounder distributions between different treatments, while our method allows the presence of time-invariant unmeasured confounding. We propose a sequential ignorability assumption based on treatment assignment heterogeneity, which is analogous to balancing the latent tendency toward each treatment due to unmeasured time-invariant factors beyond the observables. We use simulation studies to assess the sensitivity of the proposed method's performance to various model assumptions. The method is applied to observational clinical data to investigate the efficacy of continuously using mychophenolate in different subgroups of scleroderma patients who were treated with the drug.

Longitudinal causal inference; latent variable modeling; random effects models; g-computation

1 Introduction

Precision medicine[16;28] is a clinical decision-making process that uses a patient's medical history, current and previous health statuses, and observational data from a large population to make individualized treatment and care recommendations throughout the progression of a disease. For example, Wang et al.[36] predicted individual future biomarker trajectories and major clinical events for improving COVID-19 care and Coley et al.[6] utilized longitudinal biomarker measurements to improve clinical decisions about whether to remove or irradiate a patient's prostate cancer. Studying the heterogeneity in an individual's treatment effect in longitudinal settings is one of the many questions of interest in precision medicine. This involves mapping patient's current information to biomarker trajectories under potential actions such as the selection and timing of therapy. We are particularly interested in using observational data to answer the causal question "what would have happened after τ days if a specific dynamic treatment regime had been implemented, given the patient's history of h days? ", where dynamic treatment regimes are defined as treatment that may change based on observed patient history. We may then determine which treatment option is the best for a patient by assessing the average treatment effect (ATE) under different regimes for a subgroup of patients who share similar characteristics or history.

Our motivating application is to study the effectiveness of an immunosuppressant medication, mycophenolate (MMF)[22;38], on scleroderma patients using clinically observed data from the Johns Hopkins Precision Medicine Analytics Platform (PMAP) Registry. Scleroderma is a group of rare chronic autoimmune diseases marked by hardening of the skin and internal organs. There is currently no universally accepted treatment for the disease's skin thickening due to the paucity of studies demonstrating a significant effect and the associated adverse event profiles. Diffuse scleroderma is a type of the disease in which skin thickening occurs over large areas of the body and is associated with significant organ damage. In Scleroderma Lung Study II[34], MMF resulted in improvements in the modified Rodnan skin score (mRSS) among diffuse patients at the end of 24 months. We compare the efficacy of MMF-containing versus MMF-free treatment regimens for skin and lung measurements in patients who have demonstrated tolerance to MMF, whether diffuse or nondiffuse. In this observational study, there are multiple practical challenges: treatment assignment is not randomized based on measured factors, biomarkers are measured irregularly, missingness patterns may be informative about biomarker values, and natural heterogeneity among subjects exists beyond what the observables can explain. In order to tackle these issues, we use a Bayesian approach under the potential outcomes framework[29], which defines causal effect as a comparison of potential outcomes for the same set of subjects under different treatment regimes. The approach has the advantages of being able to handle structural missingness, incorporating Bayesian models with the flexibility to address complex data, and naturally quantifying uncertainty, all of which are important for decision-making in precision medicine.

The primary factor in evaluating treatment efficacy, both in this and many other scenarios of comparing treatment regimes for precision medicine, is subject heterogeneity or unmeasured factors in treatment assignment and biomarker dynamics. Individual treatment decisions are intuitively sensitive to unmeasured variables that may confound disease progression. Often, the practitioner deciding on whether or not and when to treat a patient will have access to private signals about the patient's potential outcomes, such as frailty, willingness to be treated, and potential risk of adverse effect, etc. It is not always possible to assemble a set of observed variables that serve as a proxy for the available information from all of the signals. Unmeasured variables influence not only time-varying decisions but also biomarker progression. Heckman and Willis [12] reasoned that when unobserved permanent components exist, subjects with similar observables may have heterogenous distribution of responses, i.e. an individual's sequential responses differ systematically from the group's average behavior.

The majority of existing causal inference methods for comparing time-varying treatment assume unconfoundedness, also known as the no unmeasured confounders assumption or sequential exchangeability, i.e. the treatment assignment is independent of the potential outcomes conditional on some observed variables. The potential existence of unmeasured factors that may confound the treatment assignment and biomarker dynamics violates this fundamental assumption and thus undermines these methods, including g-estimation[26;41], structural nested models [10], history-restricted marginal structural models [21], and longitudinal targeted maximum likelihood estimation [35]. Econometric literature, on the other hand, uses unobserved effects models (UEM) or unit fixed-effects models [9;15] to eliminate time-invariant unmeasured confounding by including subject-specific intercepts and having each subject act as their own control. Imai and Kim[14] used UEM in matching to estimate contemporaneous treatment effect, i.e. comparing the outcome right before and immediately after a change in the treatment status over a short time period. The main drawback of using an UEM is that due to its assumption of strict exogeneity, it is difficult to simultaneously address biases from reverse causation and time-dependent confounding[3], which are common in the causal comparison of dynamic treatment regimes.

From a modeling perspective, we account for the unmeasured patient heterogeneity in both treatment assignment and biomarker dynamics via multivariate generalized linear mixed-effects models (MGLMM)[39;1], which allows partial identification of unobserved permanent components through repeated measurements for each individual in a larger population. Behavioral and social science researchers have long used mixed-effects model[2;4;20;17;24] in research involving longitudinal data . The ability of mixed-effects models to estimate subject-specific random effects allows for quantitative characterization of between-subject heterogeneity due to unobserved factors[31;5]. Furthermore, these models describe the within-subject dependence in the time-varying outcome, which improves parameter estimation efficiency. However, due to the nonlinear link functions in MGLMM, estimated parameters in the generalized model often only have causal interpretations conditional on the random effects, that is, fixed-effects coefficients no longer lead to marginal causal effect based on potential outcomes[8] even when all covariates are exogenous[40;11].

To address this issue and enable the estimation of marginal causal effect for comparing treatment regimes with MGLMM on both population and subgroup levels, we use the g-computation algorithm, which underpins the majority of Bayesian causal inference methods. This approach directly simulates potential outcomes under a treatment path based on the joint distribution of time-varying confounders and outcomes conditional on patient history, consistently estimating potential outcomes and thus causal effects if all the conditional distributions are correctly specified. Standard g-estimation methods lead to biased effect estimates when unmeasured confounders are present, as the unobserved potential outcomes are not missing at random. From a sensitivity analysis perspective, Yang and Lok[37] assumes a nonidentifiable bias function quantifying the impact of unmeasured confounding on the average potential outcome under structural nested mean models. Sitlani et al.[33] and Qian et al.[23] compared treatment paths that differ only at a single point in time and discussed likelihood decomposition, which supports the causal interpretation of the fixed-effects coefficients estimated from a linear mixed model, i.e. as a "blip" of a structural nested model. Shardell and Ferrucci[32] incorporated joint mixed-effects models in the g-computation algorithm to estimate the population average effect of treatment regimes over time.

In this paper, we relax the unconfoundedness assumption and provide a framework for causal comparison of treatment paths using MGLMM, which accounts for the presence of unmeasured time-invariant factors as latent subject heterogeneity in treatment assignments, longitudinal outcomes, and time-varying confounders. We aim to synthesize evidence from the population pertinent to clinical decisions of an individual and to account for the dynamic progression of the individual's trajectories, all while addressing the unobserved permanent factors in selection bias and adhering to the generic causal inference ideology of only using the past to infer on the current status. Existing works on causal inference with longitudinal data using mixed-effect models often marginalize over the latent components and identify causal estimand as a function of the treatment path, covariates, and fix-effects coefficients. While the unobserved stable trait factors influencing disease progression remain constant over time, our proposal dynamically updates the information relevant to these factors by sequentially estimating the subject-specific latent variables in the longitudinal outcome and time-varying confounder models based on subject's accumulating observed or counterfactual history over time. In addition, we note that the distribution of treatment assignment heterogeneity is not fully identifiable when treatment paths are binary and monotonic because it is not a recurring process. Our discussion focuses on binary monotonic treatment process and the variance in the population distribution of treatment assignment heterogeneity is introduced as a built-in sensitivity parameter for treatment regime comparison.

Our proposal engages the treatment assignment model as part of a larger picture to bridge the gap between the confoundedness in selection bias and the heterogeneity of patients' dynamic disease progression. The work has several advantages. First, existing ways of incorporating propensity score (PS) in Bayesian causal inference[18] include specifying outcome distribution conditional on PS[41], having shared priors between propensity and outcome models, or using an inverse probability weighting or doubly robust estimator[30]; our method provides a new way to connect the propensity with the outcomes and time-varying confounders via the dependence structure on the subject-specific unobserved heterogeneity of the model components. Second, our method naturally incorporates unmeasured time-invariant factors via the random effects in MGLMM, for which the estimated covariances partially inform possible existence of unmeasured confounders. Third, we provide a new perspective to investigating the impact of potential time-invariant unmeasured confounding by using the distribution of treatment assignment heterogeneity as a sensitivity parameter involved in causal estimation, rather than quantifying unmeasured confounders in post hoc sensitivity analyses[27;37]. While random effects in the model components for outcomes and confounders reflect unobserved stable traits such as physiological factors of disease progression, treatment assignment heterogeneity is usually contextual and may be tractable based on knowledge about data collection and practice routine. As a result, the sensitivity parameters can be tailored to practitioners' needs as a controllable component to test the sensitivity of the causal estimates. Finally, under certain conditions, such as when treatment assignment heterogeneity is assumed to be absent and thus no unmeasured confounders exist, our approach identifies marginal subgroup treatment effect without making additional assumptions about the sensitivity parameter.

2 Notation and Model

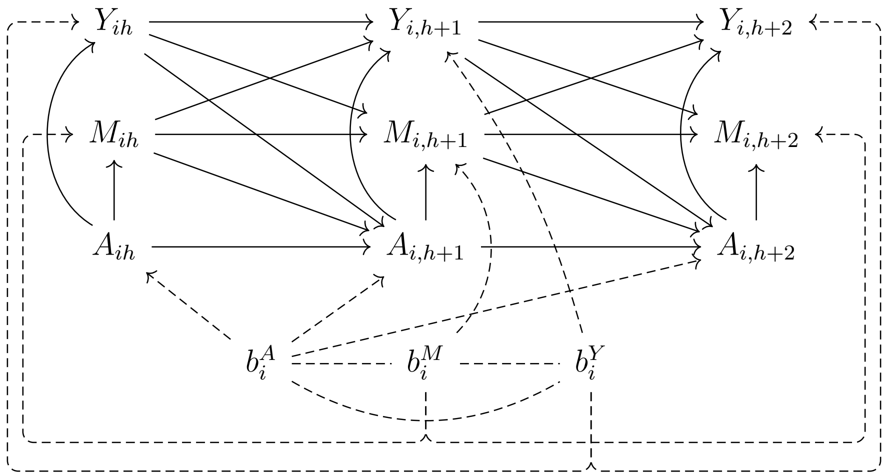

Figure 1: Directed acyclic graph (DAG) for the generalized linear mixed model displaying temporal order of the observed variables and time-invariant unmeasured heterogeneity in both treatment assignment and biomarker dynamics. Baseline characteristics V_i is excluded from the figure for simplicity.

We consider a longitudinal study that involves sequentially assigned treatment paths, and assume that time-invariant unobserved heterogeneity exists in both biomarker dynamics and treatment assignment. This paper demonstrates the method under the assumed temporal relationship of the variables as described in Figure 1, where an arrow suggests the potential of causal relationship (single arrow) or covariance (no arrow), whereas a missing arrow implies zero influence or zero covariance. There are multivariate stochastic processes, $\{(Y_t, M_t, A_t) : t \geq 0\}$, where Y_t, M_t, and A_t represent the outcome process, time-dependent confounders, and sequential treatment, respectively, where $A_t \in \{0, 1\}$. The confounders are affected by previous exposure and influence future outcomes and treamtent assignment. Let $\overline{Y}_{i,t_1:t_2}$, $\overline{M}_{i,t_1:t_2}$, and $\overline{A}_{i,t_1:t_2}$ denote the longitudinal paths observed for biomarkers, confounders, and interventions, respectively, during times $t = t_1, \dots, t_2$ for subject i, $i = 1, \dots, N$. At any time t, practitioners decide on $A_{i,(t+1)}$ based on clinical history recorded up to time t, i.e. past treatment path $\overline{A}_{i,0:t}$ and measurement history $\mathcal{H}_{i,t+1} = (V_i, \overline{Y}_{i,0:t}, \overline{M}_{i,0:t})$, where V_i is the vector of baseline information. The updated clinical history, $(\mathcal{H}_{it}, \overline{A}_{i,0:(t+1)})$, which includes the most recent treatment decision, is then the observable information for explaining the dynamics of $(Y_{i,t+1}, M_{i,t+1})$.

In this paper, we restrict the discussion to studying treatment initiations such that an initiation occurs at a single time and we assume subjects to remain treated after the initiation. Without loss of generality, we consider the outcomes to be continuous and the time-dependent confounders to be the pattern of subject visits. We model the confounders as binary variables based on the missing structure of the longitudinal outcomes. We propose using the longitudinal multivariate generalized linear mixed model (MGLMM) described below to characterize individual-level time-specific progression of biomarkers and treatment assignments. For $t = 1, \dots, T$, the continuous outcomes have a linear mixed-effects model specification,

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

where λ_Y is the link function, $\phi_A(\overline{A}_{i,0:t})$ may be a function of dosage information for person i at time t with maximum dose K, e.g. $(\mathbb{1}\{\sum_{s=1}^t A_{is} = 1\}, \dots, \mathbb{1}\{\sum_{s=1}^t A_{is} = K\})$, $b_i^Y = (b_{i0}^Y, b_{i1}^Y)$ is the vector of random effects, ψ_i^Y is the stochastic randomness following a mean zero distribution, e.g. $N(0, 1)$, and $\phi_2^Y(\mathcal{H}_{it}) \subseteq \phi_1^Y(\mathcal{H}_{it})$, $\phi_3^Y(\mathcal{H}_{it}) \subseteq \phi_1^Y(\mathcal{H}_{it})$, and $\phi_4^Y(\mathcal{H}_{it}) \subseteq \phi_2^Y(\mathcal{H}_{it})$. Outcome model parameters may take the form of $\theta^Y = (\beta_1^Y, \beta_2^Y, \sigma)$, where σ is the standard deviation of outcome distribution.

Treatment initiation is modeled as

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

where λ_A is the logit function, $\theta^A = (\beta_1^A, \beta_2^A)$, $b_i^A = b_{i0}^A$ is the random effect , and $\phi_2^A(\mathcal{H}_{it}) \subseteq \phi_1^A(\mathcal{H}_{it})$. With a binary dependent variable, the randomness satisfies $\psi_{it}^A \sim U(0, 1)$ and indicates a realization of A_{it} via $\mathbb{1}\{\psi_{it}^A \leq \lambda_A^{-1}(\eta_{it}^A)\}$ under $A_{i,t-1} = 0$. We recognize that the distribution of the heterogeneity in treatment assignment, b_i^A, is not fully identifiable from the observed data when the assignment is not a recurrent process, i.e. happens at most once for each subject. For identifiability in model estimation, it is necessary in this instance to posit values on the variance of b_i^A.

For time-dependent confounders M_{it}, the model specification is similar to equation (1),

$$M_{it} \sim f_M(\mathcal{H}_{it}, \overline{A}_{i,0:t}, b_i^M; \theta^M, \psi_{it}^M), \quad \mathbb{E}(M_{it} | \mathcal{H}_{it}, \overline{A}_{i,0:t}, b_i^M; \theta^M) = \lambda_M^{-1}(\eta_{it}^M),$$

$$\eta_{it}^M = \phi_1^M(\mathcal{H}_{it})\beta_1^M + \phi_2^M(\mathcal{H}_{it})\phi_A(\overline{A}_{i,0:t})^T\beta_2^M + \phi_3^M(\mathcal{H}_{it})b_{i0}^M + \phi_4^M(\mathcal{H}_{it})\phi_A(\overline{A}_{i,0:t})^T b_{i1}^M, \quad (3)$$

where $\theta^M = (\beta_1^M, \beta_2^M)$ if confounders are categorical, $b_i^M = (b_{i0}^M, b_{i1}^M)$ is the vector of random effects, and $\phi_2^M(\mathcal{H}_{it}) \subseteq \phi_1^M(\mathcal{H}_{it})$, $\phi_3^M(\mathcal{H}_{it}) \subseteq \phi_1^M(\mathcal{H}_{it})$, $\phi_4^M(\mathcal{H}_{it}) \subseteq \phi_2^M(\mathcal{H}_{it})$. In the motivating application, M_{it} represents missing indicators of the outcomes so we set λ_M as the logit function. In order for binary confounders to be identifiable, η_{it}^M has to have a parametric specification and we assume an additive model. The vector of randomness $(\psi_{it}^Y, \psi_{it}^M, \psi_{it}^A)$ is i.i.d and independent of b_i; $(\psi_{it}^Y, \psi_{it}^M, \psi_{it}^A)$ characterizes the stochasticity of counterfactual realizations. To control for stochasticity, we use the same set of randomness for causal estimation across different treatment regimes, ensuring that projected potential outcomes are comparable and reproducible.

The three model components, (1), (2), and (3), are connected through a ccovariance structure between the random effects,

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

In our application, we assume G_i to be the same across subjects, i.e. $b_i \sim MNV(0, G)$. Subject-specific covariance G_i can be realized by further parameterizing under assumed structures with individual level parameters. These random effects are interpreted as unobserved time-invariant subject-specific heterogeneity; they represent stable traits that influence the clinical trajectories and treatment assignment processes directly via random intercepts and indirectly through the effect of factors via random slopes. Specifically, b_i^A is the unmeasured static heterogeneity in treatment assignment, such as a patient's frailty observed but not recorded in clinic. Without loss of generality, we assume that $\psi_{it}^Y \sim N(0,1)$, identity link for λ_Y, and logit link λ_A and λ_M for the rest of the manuscript.

3 Bayesian G-Computation with MGLMM

3.1 Causal Quantities and Target Estimand

Until now, we have focused on using MGLMM to describe the data-generating mechanism as illustrated in Figure 1. When the MGLMM is correctly specified, the posterior predictive samples of the model parameters concentrates on the true data distribution. In most cases, model parameters in MGLMM do not have a causal interpretation due to the random effects, with an exception described in supplementary material.

A treatment regime dynamically defines a patient's present treatment status as a function $q(\cdot)$ of the observed or counterfactual clinical history, i.e. given a past treatment path and measurement history up to time t, $(\overline{A}_{i,0:(t-1)}, \mathcal{H}_{it})$, the treatment sequence under regime q is sequentially determined by $a_t(q) = q(\overline{A}_{i,0:(t-1)}, \mathcal{H}_{it})$. For any variable X, $X(q)$ represents the value of X had the individual received treatment under regime q. We define $\overline{Y}_{i,0:t}(q)$, $\overline{M}_{i,0:t}(q)$, and $\overline{a}_{0:t}(q) = (a_1(q), \ldots, a_t(q))$ as the counterfactual longitudinal trajectories of outcomes, confounders, and treatment path under regime q, and write the counterfactual measurement history under regime q up to before time t as $\mathcal{H}_{it}(q) = \{V_i, \overline{Y}_{i,t-1}(q), \overline{M}_{i,t-1}(q)\}$.

The heterogeneity in treatment assignment, b_i^A, represents clinician-observed private signals or stable trait factors that are not captured by data but are relevant to clinicians' judgment about the potential outcomes of patients. To account for potential time-invariant unmeasured confounding that may occur naturally in the process of treating patients in clinic, we stratify the causal estimation based on treatment assignment heterogeneity. Given a specific regime of interest, q, and the unobserved time-invariant heterogeneity, b_i^A, we aim at identifying the joint distribution of a future τ days of counterfactual trajectories conditional on observed history up to a present time h,

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

Based on (4) and g-computation[26], we can identify causal effects from the expectation of counterfactual outcomes that are functions of the fix effects parameters and the time-evolving estimations of (b_i^Y, b_i^M), by integrating over observed or counterfactual histories under the treatment path determined by the regime of interest.

Suppose we are interested in the conditional mixed average treatment effect[19] (CMATE) within a target subgroup T characterized by $\{V_i, \overline{A}_{i,0:h_i}, \overline{Y}_{i,0:h_i}, \overline{M}_{i,0:h_i}\}_{i\in T}$, in which person i contributes history information up to time h_i to the subgroup. For example, our application considers a subgroup T that contains follow-up information prior to MMF usage from scleroderma patients who were observed to be treated with MMF. Let $\widehat{P}_T$ be the empirical distribution from the observed values in subgroup T, then the target quantity CMATE is defined as

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

where N_T is the number of individuals in subgroup T. Replacing the empirical distribution with the corresponding population distribution yields the population version of CMATE, which may be viewed as the longitudinal extension of conditional ATE and realized by jointly modeling all the variables involved in the definition of subgroup T.

3.2 Assumptions and Method

In this section, we describe the assumptions and procedure for estimating CMATE, which jointly models multivariate time-varying components while accounting for the accumulation of individual information over time via a time-evolving update of time-invariant unobserved traits represented by (b_i^Y, b_i^M). The proposal enables us to assess the sensitivity of the longitudinal causal effect estimation to different distributions of unobserved treatment heterogeneity, while allowing potential existence of time-invariant unmeasured confounding. For simplicity, we leave out subscript i for the following discussion. In order to show that the conditional counterfactual joint distribution (4) can be identified without parametric form, we make the following assumptions:

Assumption. *For $t = 0, \ldots, T$,*

1. *Consistency: $\overline{Y}_{0:t} = \overline{Y}_{0:t}(q)$ and $\overline{M}_{0:t} = \overline{M}_{0:t}(q)$ if $\overline{A}_{0:t} = \overline{a}_{0:t}(q)$;*
2. *Positivity: $P(A_{t+1} = a_{t+1}(q)|V, \overline{A}_{0:t} = \overline{a}_{0:t}(q), \overline{Y}_{0:t}, \overline{M}_{0:t}, b_i^A) > 0$ with probability 1 for $t \geq 0$;*
3. *Sequential exchangeability given b_i^A: for $\tau > 0$,*

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

The consistency assumption states that when the observed treatment path follows the hypothesized regime of interest, the observed and counterfactual biomarker dynamics are equivalent. It is important to note that the equivalence does not imply the same value, but rather the same distribution. Positivity guarantees that there is no systematic exclusion of a plausible treatment pattern over time. The classic assumption of sequential exchangeability[7] is commonly adopted in the existing literature, assuming that the observed pretreatment history can sufficiently explain the dependence between a current treatment assignment and future counterfactuals. We extend this assumption to condition on the unobserved time-invariant heterogeneity in treatment assignment, which is quantified by the random effect b_i^A in model (2). In practice, the heterogeneity in treatment assignment may be attributable to patients' willingness to be treated, the potential risk of adverse effects from treatment, and the clinician's perception of treatment. Under these assumptions and that models (1), (2), and (3) are correctly specified, (4) can be nonparametrically identified as below (see Appendix A for details),

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

In Figure 2, we use a single-world intervention graph[13;25] (SWIG) to display the independencies that lead to (6) and show the counterfactual dependencies that would exist if we set the treatment path to that under regime q. The graph is constructed by splitting the treatment nodes $\overline{A}_{i,(h+1):(h+\tau)}$ of the causal diagram in Figure 1 and replacing all descendants of the assigned treatment with their potential outcomes, marking all counterfactuals in red. The conditional sequential exchangeability assumption is demonstrated in the SWIG by d-separation between the counterfactual trajectories $(\overline{Y}_{(h+1):(h+\tau)}(q), \overline{M}_{(h+1):(h+\tau)}(q))$ and $A_{i,h+1}$ conditional on $(\overline{A}_{0:h}, \overline{Y}_{0:h}, \overline{M}_{0:h}, b_i^A)$. If b_i^A is not controlled for, selection bias would be induced by paths $A_{i,h+1} \leftarrow b_i^A \leftrightarrow b_i^Y \rightarrow Y_{i,h+1}(q)$ and $A_{i,h+1} \leftarrow b_i^A \leftrightarrow b_i^M \rightarrow M_{i,h+1}(q)$, while stratifying on b_i^A blocks these paths. Variables inside rectangles of Figure 2 are quantities involved in (6) that are relevant to the time-evolving update of (b_i^Y, b_i^M). At each time point, the distribution of (b_i^Y, b_i^M) can be derived based on information backflow from observed or counterfactual biomarkers' history, resulting in a sequential update of these subject-specific unobserved permanent traits. Hypothesized treatment status $\overline{a}_{it}(q)$, $t \in [h+1, h+\tau]$, does not contribute to the sequential update of (b_i^Y, b_i^M) because it generates no additional information beyond the definition of the regime of interest.

At each time $s \in [h, h+\tau)$, Monte Carlo simulation of counterfactual outcomes and confounders $(Y_{s+1}(q), M_{s+1}(q))$ based on (6) involves integration over $P(b_i^Y, b_i^M | V, \overline{A}_{0:h}, \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^A)$, an updated conditional posterior distribution of (b_i^Y, b_i^M). The trajectories being conditioned on, $(\overline{Y}_{0:s}, \overline{M}_{0:s})$, is equivalent to $(\overline{Y}_{0:h}, \overline{Y}_{(h+1):s}(q), \overline{M}_{0:h}, \overline{M}_{(h+1):s}(q))$ in distribution, which is a mix of observed and counterfactual variables. Note that the counterfactual trajectories $(\overline{Y}_{(h+1):s}(q), \overline{M}_{(h+1):s}(q))$ have the following distribution

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

under regime q and treatment assignment heterogeneity b_i^A, as implied by the formulation of equation (6). The sampling of $(b_i^Y, b_i^M) \sim P(b_i^Y, b_i^M | V, \overline{A}_{0:h}, \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^A)$ may be complicated by nonlinear link functions in the MGLMM. We consider the following general strategy: first, calculate

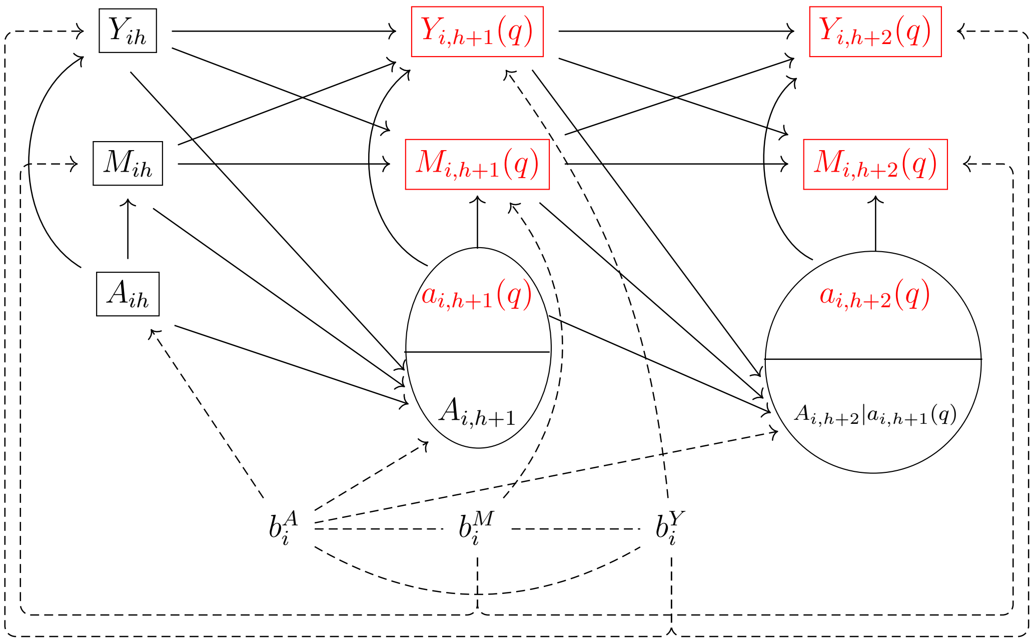

Figure 2: SWIG.

the Laplace approximation of the posterior distribution $(b_i^Y, b_i^M, b_i^A | V, \overline{A}_{0:h}, \overline{Y}_{0:s}, \overline{M}_{0:s})$, denoted by $MVN(\hat{b}_i, V)$, and then sample the heterogeneities via the corresponding conditional distribution, $(b_i^Y, b_i^M)|b_i^A$, with b_i^A set to a certain value. The procedure is illustrated in Appendix B. We provide in Appendix C the pseudocode for generating posterior samples of counterfactual trajectories from $P(\overline{Y}_{(h+1):(h+\tau)}(q), \overline{M}_{(h+1):(h+\tau)}(q)|V, \overline{A}_{0:h}, \overline{Y}_{0:h}, \overline{M}_{0:h}, b_i^A)$ based on (6).

We have been stratifying on b_i^A thus far in our discussion. The target estimand CMATE expressed in equation (5) is the marginal subgroup ATE, marginalizing over unobserved heterogeneity. Therefore, using the following formula, we integrate each component of CMATE over the distribution of b_i^A conditional on subgroup T,

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

As a result, CMATE is a functional of the counterfactual joint distribution (4) because the conditional expectation $\mathbb{E}(Y_{h+\tau}(q)|V, \overline{A}_{0:h}, \overline{Y}_{0:h}, \overline{M}_{0:h}, b_i^A)$ in (7) can be expressed as

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

When comparing regimes q_1 and q_2, we estimate the causal contrast by integrating

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

over the subgroup distribution of treatment assignment heterogeneity, $P(b_i^A|V, \overline{A}_{0:h}, \overline{Y}_{0:h}, \overline{M}_{0:h})$. Appendix C contains the computational details for calculating CMATE, and the motivating application illustrates subgroup causal effect estimation under the proposed method.

The subgroup distribution of b_i^A, which is controlled for and marginalized over, serves as the sensitivity parameter in the calculation of CMATE. Recall that MGLMM assumes that (b_i^Y, b_i^M, b_i^A) follows $MVN(0, G)$. Let v denote the variance of b_i^A, representing the assumed amount of variation in treatment assignment heterogeneity among subjects. When treatment assignment is a binary monotonic process as in the motivating application, v is unidentifiable and needs a posited value in model estimation because treatment assignment is not a recurring event. When subgroup T contains individual history of different lengths, subgroup distribution $P(b_i^A|V, \overline{A}_{0:h}, \overline{Y}_{0:h}, \overline{M}_{0:h})$ can be derived conditional on a given value of v for each individual. The evaluation of causal effectiveness may vary under different values of v . In this case, the variance of the unidentifiable time-invariant quantity b_i^A serves as a sensitivity parameter in the estimation of causal effects. For population ATE, the target quantity can be derived as

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

where $b_i^A \sim N(0, v)$ based on model assumption. Appendix A gives further details to the calculation of the mixed ATE of the target population[19], $\widehat{\mathbb{E}}(Y_{h+\tau}(q))$, which replaces the target population distribution with the corresponding empirical distribution.

When no treatment heterogeneity under MGLMM, i.e. setting $v = 0$, the proposal simplifies to the standard g-computation of utilizing only the model components for outcomes and confounders because the assignment mechanism is unconfounded[19]. Having $v = 0$ is a sufficient but unnecessary condition for having no unmeasured confounders. Under MGLMM, $\text{cov}(b_i^A, b_i^M) = \text{cov}(b_i^A, b_i^Y) = 0$ leads to no unmeasured confounders. When there are no unmeasured confounders, MGLMM still allows unobserved factors to influence treatment assignment, i.e. $v \neq 0$, as long as b_i^A is not correlated with the unobserved heterogeneity in biomarker dynamics (b_i^Y, b_i^M); examples of such non-confounding treatment assignment heterogeneity include preference for a treatment based on personal beliefs or social stigma. On the other hand, we note that the covariances $\text{cov}(b_i^A, b_i^M)$ and $\text{cov}(b_i^A, b_i^Y)$ are estimable given the variance of b_i^A, v. As a result, our method is able to provide insight into the potential existence of unmeasured confounders based on the MGLMM's estimated covariances.

4 Simulation

Assuming each person has two follow-up visits, $T_i = 2$, we simulate continuous biomarker Y_{it} and binary time-varying treatment A_{it} via

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

where $e_{it}^Y \sim N(0, 0.4^2)$, V_i is the baseline covariate, $V_i \sim$ Bernoulli(0.5), and $Y_{i0} \sim N(0, 1)$. Write $\rho = \text{Corr}(b_{i0}^A, b_{i0}^Y)$, $s_A = \sqrt{\text{Var}(b_{i0}^A)}$, $s_Y = \sqrt{\text{Var}(b_{i0}^Y)}$. We assume the random effects to follow $(b_{i0}^A, b_{i0}^Y) \sim N(0, G)$, where the covariance matrix G has elements $G_{11} = s_A^2$, $G_{12} = G_{21} = \rho s_A s_Y$, and $G_{22} = s_Y^2$. We set $s_Y = 0.8$ and randomly sample 100 replicates for each of the 101 settings defined by parameter combinations $(s_A, \rho) \in (0,0) \cup \{(s_A, \rho); s_A \in \{0.1, \ldots, 0.9, 1\}, \rho \in \{0, 0.1, \ldots, 0.9\}\}$. When $s_A > 0$ and $\rho < 1$, G is guaranteed to be positive definite because $\det(G) = (1-\rho)s_A^2 > 0$. When $\rho = 0$, matrix $G = \begin{pmatrix} s_A^2 & 0 \\ 0 & s_Y^2 \end{pmatrix}$ represents the case of no unmeasured confounding. We consider two scenarios, $\nu_k = k$ and $\nu_k = 0$ for $k = 1, 2$, where $\nu_k = k$ represents a stronger treatment effect over time and $\nu_k = 0$ indicates no treatment effect. For each scenario of ν_k and each setting of (s_A, ρ), we simulate 100 data replicates with sample size $n = 500$.

Define the regime of always treat as q_1 and not treated as q_0. Relative to each simulated dataset, we aim to compare the overall mixed ATE[19] at $t = 2$ under the regimes of being always on treatment, $\overline{a}_{1:2}(q_1) = (1, 1)$, versus not treated, $\overline{a}_{1:2}(q_0) = (0, 0)$. By consistency and conditional sequential exchangeability assumptions, for any treatment regime q of interest, i.e. $\overline{a}_{1:2}(q) = (a_1, a_2)$, target counterfactual quantity can be expressed as observed variables via g-formula as $\mathbb{E}[Y_{i2}(q)] = \mathbb{E}(Y_{i2}|A_{i1} = a_1, A_{i2} = a_2)$, derived in Appendix D. Population ATE, is the average difference between counterfactual outcomes $Y_{i2}(q_1)$ and $Y_{i2}(q_0)$,

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

We estimate the population ATE by mixed ATE, $\widehat{g}(q_0, q_1) = \widehat{\mathbb{E}}[Y_{i2}(q_1)] - \widehat{\mathbb{E}}[Y_{i2}(q_0)]$, where we use sample-specific empirical distribution $\widehat{P}(V = v)$ instead of $P(V = v)$, the population distribution of V involved in the derivation of population ATE.

By equation (8), the true population ATE is $g(q_0, q_1) = 1.2$ for scenario $\nu_k = k$ and 0 for scenario $\nu_k = 0$. For each data replicate, we sample the posterior predictive distribution of the mixed ATE by marginalizing over $b_i^A \sim N(0, \widehat{s}_A^2)$ under $\widehat{s}_A \in \{0, 0.3, 1\}$, where $\widehat{s}_A$ represents the assumed degree of variation in the unexplained treatment assignment heterogeneity. Different combinations of simulation truth (s_A, ρ) and assumed parameter $\widehat{s}_A$ explore the following three cases: (1) no unmeasured confounding ($\rho = 0$), (2) unmeasured confounding exists with correctly specified models ($\rho \neq 0, \widehat{s}_A = s_A$), and (3) unmeasured confounding exists with a mis-specified extent of treatment assignment heterogeneity ($\rho \neq 0, \widehat{s}_A \neq s_A$). Note that the statement about unmeasured confounding is based on the assumptions encoded in the causal DAG and model choices. Let N_{post} be the number of posterior draws. Given the rth data replicate under a simulation setting, we summarize the posterior samples of mixed ATE $(g_1^{(r)}, \ldots, g_{N_{post}}^{(r)})$ with its posterior mean $\overline{g}^{(r)} = \sum_{\ell=1}^{N_{post}} g_\ell^{(r)} / N_{post}$ and 95% credible interval $(L^{(r)}, U^{(r)})$. We aggregate across simulation replicates by mean squared error (MSE) $\frac{1}{100} \sum_{r=1}^{100} [\overline{g}^{(r)} - g(q_0, q_1)]^2$ and coverage $\frac{1}{100} \sum_{r=1}^{100} \mathbb{1}\{g(q_0, q_1) \in (L^{(r)}, U^{(r)})\}$.

For scenario $\nu_k = k$, Figure 3 displays the MSE and coverage of posterior mixed ATE in the first and second rows, respectively. The three columns from left to right correspond to estimations under $\widehat{s}_A$ being 0, 0.3, and 1, respectively. For each plot, the horizontal and vertical axes are the true parameters s_A and ρ under which data replicates were generated. Green indicates better estimation of the causal effect, i.e. lower MSE and higher posterior coverage. We observe that the causal effect is estimated relatively better when $\widehat{s}_A$ is no larger than the true value s_A. In addition, when there is no or close to no unmeasured confounding, i.e. ρ is close to zero, the estimated mixed ATE is robust to the posited value $\widehat{s}_A$. In other words, poor estimation and coverage occur when the assumed variation in treatment assignment heterogeneity differs significantly from its true value and there is substantial unmeasured confounding. For $\widehat{s}_A = 1$, estimation seems to be always better, so we also visualized in Figure 4 the ratio of MSE under $\widehat{s}_A$ being 0 versus 1 and 0.3 versus 1. We see that estimations are better when the $\widehat{s}_A$ is no larger than the truth and when no unmeasured confounding is true, smaller values of $\widehat{s}_A$ are favored regardless of the true value s_A. For the other scenario, $\nu_k = 0$, the treatment has no effect on the outcome at all times and $\mathbb{E}[Y_{it}(q)]$ does not depend on the treatment path. Figures 5 and 6 summarize the results and yield similar conclusions as under $\nu_k = k$. When $\widehat{s}_A$ is close to the truth ($\widehat{s}_A \approx s_A$) or when no unmeasured confounding is close to being true ($\rho \approx 0$), the method can find null-effect estimates well.

5 Application

The proposed method is applied to clinical data from scleroderma patients collected longitudinally through the Johns Hopkins PMAP Registry. The application aims to study the causal effectiveness of MMF initiation regimes among the subgroup who were treated with MMF. The inference is performed by sampling and comparing the posterior predictive distribution of counterfactual outcome trajectories across time intervals under different treatment regimes, where outcomes are continuous time-varying multivariate biomarkers. Disease onset is defined by the emergence of symptoms, which typically occurs prior to and is inquired about during the enrollment visit. Patients whose enrollment visits occurred within six years of disease onset and between 2010 and 2020 are included in the analysis data. The analysis utilizes individual clinical histories prior to February 28, 2022, and the observed maximum duration of follow-up in this data is ten years. Specifically, the data include all available follow-up visits when no MMF was ever taken, and up to two years after the first occurrence of continuous MMF consumption. The study includes 506 scleroderma patients who had not previously been treated with MMF at the time of enrollment, with 194 of them started MMF during follow-up. Among these individuals, 80% are females, 20% are African Americans, and 40% have diffuse scleroderma. Age at disease onset has first, second, and third quartiles as 38, 48, and 58 years old. Because patient visits are anticipated to be every six months, our analysis frames the progression of time-varying variables by six-month intervals. Over 90% of the observed MMF initiation happened during the first five post-enrollment time intervals.

To investigate the efficacy of MMF over the course of continuous use, assuming tolerance to the drug, we focus on the first two years of MMF usage in patients who were treated with MMF. Suppose person i was observed to start using MMF at time $\tilde{s}_i$, we evaluate the effectiveness of MMF by comparing the regimes of continuously taking MMF versus without MMF during the two-year period of time intervals $[\tilde{s}_i, \tilde{s}_i + 4)$. In this study, we consider outcomes Y_{it} to be measurements of the modified Rodnan skin score (mRSS) and lung scores evaluated by forced vital capacity (FVC) and diffusing capacity for carbon monoxide (DLCO). FVC and DLCO are continuous scores and are standardized for analysis. We quantilized mRSS to the standard normal distribution as mRSS has 51 levels. Missingness in biomarker measurements is common due to the nature of clinical data being observed only when patients take the initiative to comply with visit schedules. Among the included patients in this study, 82.6%, 83.6%, and 50% had at least one interval without measurement for FVC, DLCO, and mRSS, respectively. Visit patterns may potentially confound the effectiveness of MMF on the biomarkers. We define confounders M_{it} to be indicators of whether FVC, DLCO, and mRSS were updated for person i at time interval t, representing visit pattern over time.

We define V_i to be the vector of baseline demographic variables including gender, race, and age.

Let B_i be the baseline disease type, i.e. indicator of diffuse scleroderma. Based on domain knowledge, biomarkers are considered to progress upon time since disease onset, denoted as $S_{it} = t + O_i$, where O_i is the duration between disease onset and the enrollment visit. Write $\tilde{Y}_{it}$ as the carried forward measurement of biomarkers at time t. Because we limit the study to no more than two years of continuous MMF use, i.e. four intervals, dosage information at time t can be summarized by the vector $D(A_{i,0:t}) = \{\mathbb{1}\big(\sum_{\ell=1}^{t} A_{i\ell} = 1\big), \dots, \mathbb{1}\big(\sum_{\ell=1}^{t} A_{i\ell} = 4\big)\}$, where $A_{i,0:t} = (A_{i1}, \dots, A_{it})$ is the binary indicator vector of whether person i was observed to be on MMF over time. In addition, time between disease onset and MMF initiation is denoted by I_{it}, which equals zero when $A_{it} = 0$ and equals $S_{it'}$ when $A_{it} = 1$, where t' is the time of treatment initiation satisfying $A_{i,t'-1} = 0$, $A_{it'} = 1$, and $t' \leq t$. The following joint model is assumed for outcomes, time-varying confounders, and treatment assignment,

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

where

$$(b_{i0}^Y, b_{i0}^M, b_{i0}^A)^T \sim N(0, G),$$

$$\phi_1(\mathcal{H}_{it}) = \{1, \tilde{Y}_{i,t-1}, V_i, B_i, ns(S_{it}, \nu_s), B_i \times ns(S_{it}, \nu_s)\},$$

$$\phi_2(\mathcal{H}_{it})\phi_A(\overline{A}_{i,0:t})^T = \{D(\overline{A}_{i,0:t}), V_i \times D(\overline{A}_{i,0:t}), B_i \times D(\overline{A}_{i,0:t}), I_{it} \times D(\overline{A}_{i,0:t})\},$$

and we assume $\nu_s = 4$. Note that both the outcomes Y_{it} and confounders M_{it} are multivariate, i.e. Y_{it} and $M_{it} \in \mathbb{R}^3$. Specifically, $(b_{i0}^Y, b_{i0}^M, b_{i0}^A) \in \mathbb{R}^7$ and the covariance matrix $G \in \mathbb{R}^{7\times7}$. Model validation results are summarized in Figure 7. Black triangles represent the observed mean of time-varying variables at each time. The colored curves and areas represent the posterior mean and 95% posterior credible interval of the one-step forward prediction for each time-varying variable under various posited values of b_{i0}^A's standard deviation.

For each person i who was observed to have MMF during follow-up, we initiate the comparison of two regimes q_1 and q_2 at time $\tilde{s}_i$, conditional on the person's clinical history up to time $\tilde{s}_i - 1$, i.e. $(V_i, \overline{A}_{i,0:(\tilde{s}_i-1)}, \overline{Y}_{i,0:(\tilde{s}_i-1)}, \overline{M}_{i,0:(\tilde{s}_i-1)})$. Based on the algorithm outlined in Appendix C, we sample from the posterior predictive distribution of $(\overline{Y}_{i,\tilde{s}_i:(\tilde{s}_i+3)}(q_z), \overline{M}_{i,\tilde{s}_i:(\tilde{s}_i+3)}(q_z))$, which is the counterfactual trajectories under regime q_z, $z = 1, 2$, and obtain posterior samples of the counterfactual trajectories, denoted as $\{\overline{Y}^{(\ell)}_{i,\tilde{s}_i:(\tilde{s}_i+3)}(q_z), \overline{M}^{(\ell)}_{i,\tilde{s}_i:(\tilde{s}_i+3)}(q_z); \ell = 1, \dots, N_{post})\}$. Causal comparative effectiveness between the two regimes is quantified by the averaged differences in biomarkers over time, $\mathbb{D}_j = \{\sum_{i\in T} d_{ij}^{(\ell)}/N_T; \ell = 1, \dots, N_{post}\}$, where $j \in \{0, 1, 2, 3\}$ indexes time since regimen application, $d_{ij}^{(\ell)} = Y^{(\ell)}_{i,\tilde{s}_i+j}(q_1) - Y^{(\ell)}_{i,\tilde{s}_i+j}(q_2)$, and T is the subgroup of interest. Figure 8 displays the CMATE of MMF among the treated individuals and compares the two regimes, initiate MMF as observed versus no MMF. The figure depicts the posterior mean and 95% credible interval of $\mathbb{D}_j$ over the two years of regimen comparison, i.e. $j = 0, 1, 2, 3$, under posited values of $\widehat{s}_A \in \{0, 0.1, 0.25, 0.5, 0.75, 1\}$ indicated by the decrease in opacity as $\widehat{s}_A$ increases, stratified by diffuse scleroderma statuses.

Figure 8 shows that incorporating MMF into the treatment of patients with diffuse scleroderma has a significant effect on skin score during the two years after drug initiation, assuming drug tolerance and continuous use of the drug. This is consistent with the Scleroderma Lung Study II, which found that MMF significantly improved mRSS in patients with diffuse scleroderma at the end of 24 months. We investigated the incorporation of MMF in the treatment of scleroderma patients, whereas the clinical trial focused on the effect of using MMF alone versus no treatment at all. Furthermore, our findings show that adding MMF to the treatment of nondiffuse patients has no long-term benefit; this has clinical implications because MMF is an immunosuppressive agent that may increase the risk of serious infection and blood pressure issues. Observe that drug combination and treatment practices for regimens containing or not containing MMF may differ in the real world; further research is needed to examine the impact of the difference in practices and treatment patterns resulting from the use or nonuse of MMF.

6 Discussion

Deciding which treatment regime is better for patients of a specific subgroup or history pattern is a basic question for treating patients in clinics. Causal inference is a natural tool for answering such questions, but characteristics of clinical data need to be accommodated for valid inference when evaluating the efficacy of treatment paths. Observational longitudinal clinical datasets often include treatment assignments that are not randomized based on observed patient history, as well as irregular measurements that may yield informative missingness from patients' visit patterns. A key factor in comparing treatment paths is the natural heterogeneity in treatment assignment and biomarker dynamics that goes beyond what observables can explain. These are typical features of longitudinal clinical datasets. Choosing the most effective treatment regimen for one type of patient requires summarizing evidence from a population of patients of a similar type while accounting for such person's specific biomarker trends. The statistical model used to answer this question is complex by nature. This paper describes the simplest possible model that accommodates these characteristics, such as nonrandom treatment assignment and patient heterogeneity, and provides a tool for the comparison of treatment paths.

The main contribution of this work is to develop a Bayesian framework for causal inference with observational longitudinal data, estimating the subgroup effectiveness of binary treatment paths on longitudinal outcomes while accounting for time-varying confounders and allowing the existence of time-invariant unmeasured confounding. We propose to simultaneously model biomarker dynamics and treatment assignment by MGLMM, which retains the capability to deal with unmeasured confounding to some degree when the model specification is reasonably close to being correct. Since mixed-effects models do not rely on the assumption of no unmeasured confounding, which is required by the majority of the existing g-estimation methods, our approach gives the possibility of consistently estimating the causal effects of treatment paths even when unmeasured confounding certainly exists or when an unconfounded variable or instrumental variable is not available. We note that the MGLMM introduced here does not deal with time-varying unmeasured confounding. When a random slope for a time-varying variable is specified in the model, it is considered an unobserved trait that is time-invariant but characterizes patient heterogeneity in dynamic progression. Furthermore, MGLMM has a specific representation of unmeasured confounding, the degree of which is governed by the unexplained variation in treatment assignment s_A and the correlation between the treatment assignment and biomarker dynamics that operates through the correlation parameter ρ. A small ρ and a large s_A, or a large ρ and a small s_A, may both lead to heavy unmeasured confounding. The method has the potential to be extended to guide the inclusion of other latent variable models in Bayesian causal inference.

Note that our proposal does rely on the parametric assumption about the joint distribution of outcomes, time-varying confounders, and treatment assignment. We cannot test the assumption of no unmeasured confounding unless the randomization of treatment assignment is guaranteed or controlled, i.e. through a cellphone application. Under the strong assumption of no unmeasured confounding, a strong parametric model assumption would not be required for valid causal inference. However, in the example of treating patients with chronic rare diseases in clinics, unmeasured confounders unavoidably exist. We recognize that there is no free lunch in causal inference and to relax the uncounfoundedness assumption, the tradeoff here is to make additional assumptions on model specifications, which is also largely untestable. A completely nonparametric causal effect in observational data cannot be identified and untestable assumptions are always needed for the identification of causal effects. The parametric model assumption facilitated the identification of treatment effects even when unmeasured confounders may exist. The structured unmeasured confounding represented in the MGLMM provides insight into how the very specific kind of unmeasured confounding impact the causal effect of treatment paths.

Method-wise, we adopt the Bayesian g-computation algorithm (GCA) by incorporating MGLMM as the time-evolving generative component, while accounting for the real-time update of subject-specific unobserved stable traits as patient history accumulates over time. Our proposal makes subgroup evaluation of treatment paths possible by involving time-varying estimation of latent variables in the GCA, instead of marginalizing them out as in population ATE. Furthermore, the method provides a way of incorporating propensity scores (PS) in Bayesian causal inference. Existing ways of combining PS and outcomes models include specifying outcomes distribution based on PS, having shared parameters or priors between PS and outcome models, or using posterior-based inverse probability weighting or doubly robust estimators[19]. Our method falls under the category of having shared parameters or priors between PS and outcome models, using a multivariate Gaussian latent structure to connect them through covariance between latent variables. Lastly, our method provides an alternative way to assess sensitivity analysis in causal inference. Instead of assuming no unmeasured confounding and conducting post hoc analysis to assess bias due to unobserved confounding, the proposal estimates causal effects conditional on different posited values of the sensitivity parameter, which is the variance of unobserved treatment assignment heterogeneity. Future extensions of the method should consider categorical and count outcomes, multiple or continuous treatments, and more flexible distributional assumptions on the patient heterogeneities.

References

[1] F. Achana, D. Gallacher, R. Oppong, S. Kim, S. Petrou, J. Mason, and M. Crowther. Multivariate generalized linear mixed-effects models for the analysis of clinical trial–based cost-effectiveness data. *Medical Decision Making*, 41(6):667–684, 2021.

[2] A. Agresti, J. G. Booth*, J. P. Hobert*, and B. Caffo*. Random-effects modeling of categorical response data. *Sociological Methodology*, 30(1):27–80, 2000.

[3] P. D. Allison, R. Williams, and E. Moral-Benito. Maximum likelihood for cross-lagged panel models with fixed effects. *Socius*, 3:2378023117710578, 2017.

[4] M. P. Berger and F. E. Tan. Robust designs for linear mixed effects models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(4):569–581, 2004.

[5] N. Bolger and J.-P. Laurenceau. *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford press, 2013.

[6] R. Y. Coley, A. J. Fisher, M. Mamawala, H. B. Carter, K. J. Pienta, and S. L. Zeger. A bayesian hierarchical model for prediction of latent health states from multiple data sources with application to active surveillance of prostate cancer. *Biometrics*, 73(2):625–634, 2017.

[7] S. Greenland and J. M. Robins. Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, 15(3):413–419, 1986.

[8] S. Greenland, J. Pearl, and J. M. Robins. Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46, 1999.

[9] F. I. Gunasekara, K. Richardson, K. Carter, and T. Blakely. Fixed effects analysis of repeated measures data. *International journal of epidemiology*, 43(1):264–269, 2014.

[10] J. He, A. Stephens-Shields, and M. Joffe. Structural nested mean models to estimate the effects of time-varying treatments on clustered outcomes. *The International Journal of Biostatistics*, 11 (2):203–222, 2015.

[11] P. J. Heagerty. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55(3):688–698, 1999.

[12] J. J. Heckman and R. J. Willis. A beta-logistic model for the analysis of sequential labor force participation by married women. *Journal of Political Economy*, 85(1):27–58, 1977.

[13] M. A. Hernán and J. M. Robins. Causal inference, 2010.

[14] K. Imai and I. S. Kim. When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, 63(2):467–490, 2019.

[15] J. S. Kaufman. Commentary: Why are we biased against bias? *International journal of epidemiology*, 37(3):624–626, 2008.

[16] M. R. Kosorok and E. B. Laber. Precision medicine. *Annual review of statistics and its application*, 6:263–286, 2019.

[17] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[18] D. Li, S. Iddi, W. K. Thompson, M. C. Donohue, and A. D. N. Initiative. Bayesian latent time joint mixed effect models for multicohort longitudinal data. *Statistical methods in medical research*, 28(3):835–845, 2019.

[19] F. Li, P. Ding, and F. Mealli. Bayesian causal inference: A critical review. *Philosophical Transactions of the Royal Society A, Mathematical, Physical and Engineering Sciences*, 2022.

[20] T. M. Luger, J. Suls, and M. W. Vander Weg. How robust is the association between smoking and depression in adults? a meta-analysis using linear mixed-effects models. *Addictive behaviors*, 39(10):1418–1429, 2014.

[21] R. Neugebauer, M. J. van der Laan, M. M. Joffe, and I. B. Tager. Causal inference in longitudinal studies with history-restricted marginal structural models. *Electronic journal of statistics*, 1:119, 2007.

[22] M. A. Omair, A. Alahmadi, and S. R. Johnson. Safety and effectiveness of mycophenolate in systemic sclerosis. a systematic review. *PLoS One*, 10(5):e0124205, 2015.

[23] T. Qian, P. Klasnja, and S. A. Murphy. Linear mixed models with endogenous covariates: modeling sequential treatment effects with application to a mobile health study. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 35(3):375, 2020.

[24] S. W. Raudenbush and A. S. Bryk. *Hierarchical linear models: Applications and data analysis methods*, volume 1. sage, 2002.

[25] T. S. Richardson and J. M. Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

[26] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

[27] J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

[28] A. Rosen, S. L. Zeger, et al. Precision medicine: discovering clinically relevant and mechanistically anchored disease subgroups at scale. *The Journal of clinical investigation*, 129(3):944–945, 2019.

[29] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[30] M. E. Schnitzer, J. Sango, S. Ferreira Guerra, and M. J. Van der Laan. Data-adaptive longitudinal model selection in causal inference with collaborative targeted minimum loss-based estimation. *Biometrics*, 76(1):145–157, 2020.

[31] J. E. Schwartz and A. A. Stone. The analysis of real-time momentary data: A practical guide. *The science of real-time data capture: Self-reports in health research*, pages 76–113, 2007.

[32] M. Shardell and L. Ferrucci. Joint mixed-effects models for causal inference with longitudinal data. *Statistics in medicine*, 37(5):829–846, 2018.

[33] C. M. Sitlani, P. J. Heagerty, E. A. Blood, and T. D. Tosteson. Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. *Statistics in medicine*, 31(16): 1738–1760, 2012.

[34] D. P. Tashkin, M. D. Roth, P. J. Clements, D. E. Furst, D. Khanna, E. C. Kleerup, J. Goldin, E. Arriola, E. R. Volkmann, S. Kafaja, et al. Mycophenolate mofetil versus oral cyclophosphamide in scleroderma-related interstitial lung disease (sls ii): a randomised controlled, double-blind, parallel group trial. *The lancet Respiratory medicine*, 4(9):708–719, 2016.

[35] M. J. Van der Laan, S. Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 10. Springer, 2011.

[36] Z. Wang, M. G. Bowring, A. Rosen, B. Garibaldi, S. Zeger, and A. Nishimura. Learning and predicting from dynamic models for covid-19 patient monitoring. *Statistical Science*, 37(2): 251–265, 2022.

[37] S. Yang and J. J. Lok. Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statistica Sinica*, 28(4):1703, 2018.

[38] A. C. Zamora, P. J. Wolters, H. R. Collard, M. K. Connolly, B. M. Elicker, W. R. Webb, T. E. King Jr, and J. A. Golden. Use of mycophenolate mofetil to treat scleroderma-associated interstitial lung disease. *Respiratory medicine*, 102(1):150–155, 2008.

[39] S. L. Zeger and M. R. Karim. Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American statistical association*, 86(413):79–86, 1991.

[40] S. L. Zeger, K.-Y. Liang, and P. S. Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.

[41] T. Zhou, M. R. Elliott, and R. J. Little. Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association*, 114(525):1–19, 2019.

Figures

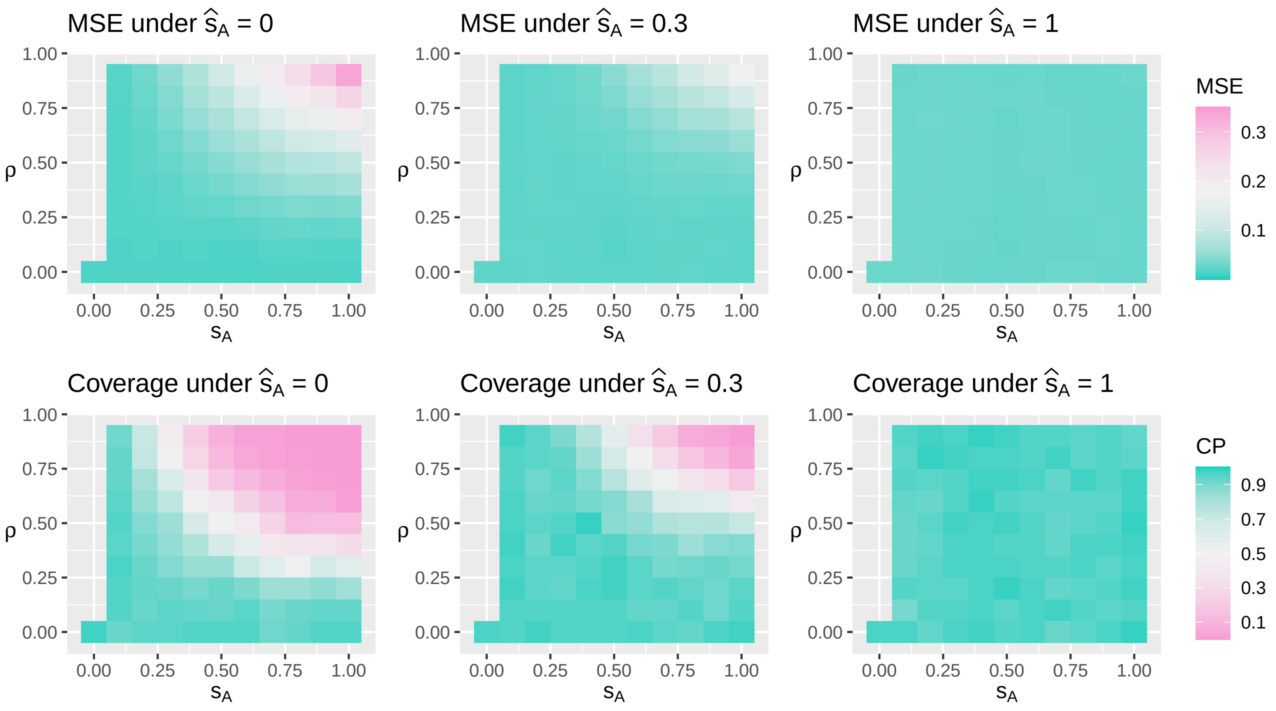

Figure 3: Under true treatment effect being 1.2 at the second time point, the figure displays mean squared error (MSE) and posterior coverage for mixed ATE under different simulation truth (s_A, ρ) and assumed model parameter $\widehat{s}_A$. Color green refers to better estimation, e.g. lower MSE and higher coverage probability.

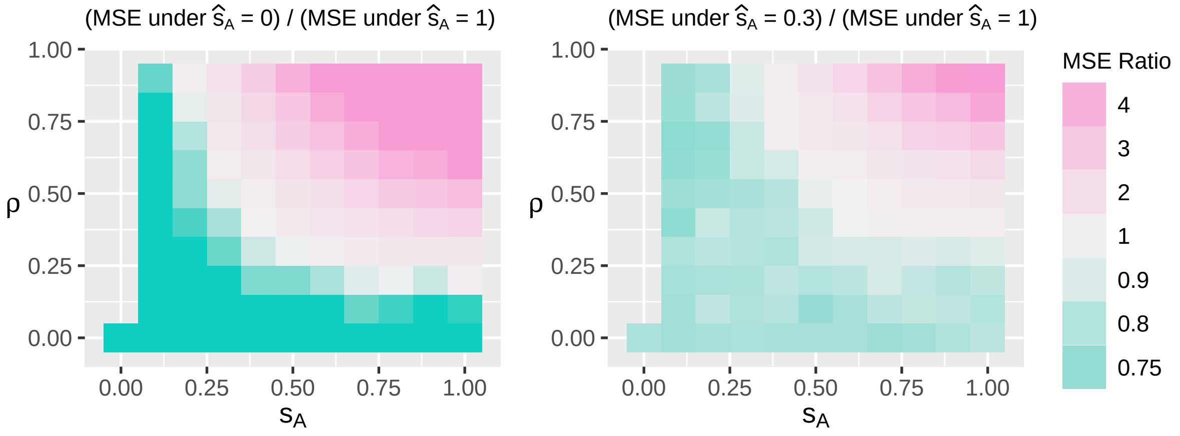

Figure 4: MSE ratio under true treatment effect being 1.2 at the second time point.

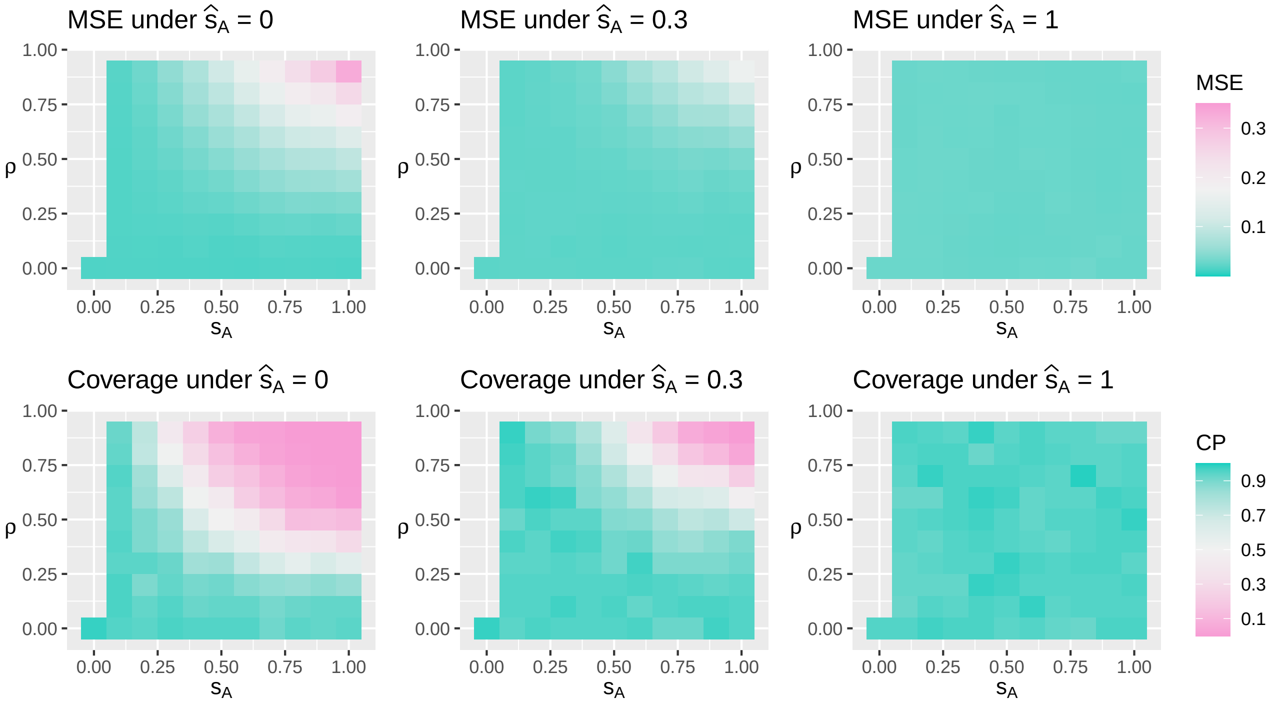

Figure 5: Under true treatment effect being 0 at the second time point, the figure displays mean squared error (MSE) and posterior coverage for mixed ATE under different simulation truth (s_A, ρ) and assumed model parameter $\widehat{s}_A$. Color green refers to better estimation, e.g. lower MSE and higher coverage probability.

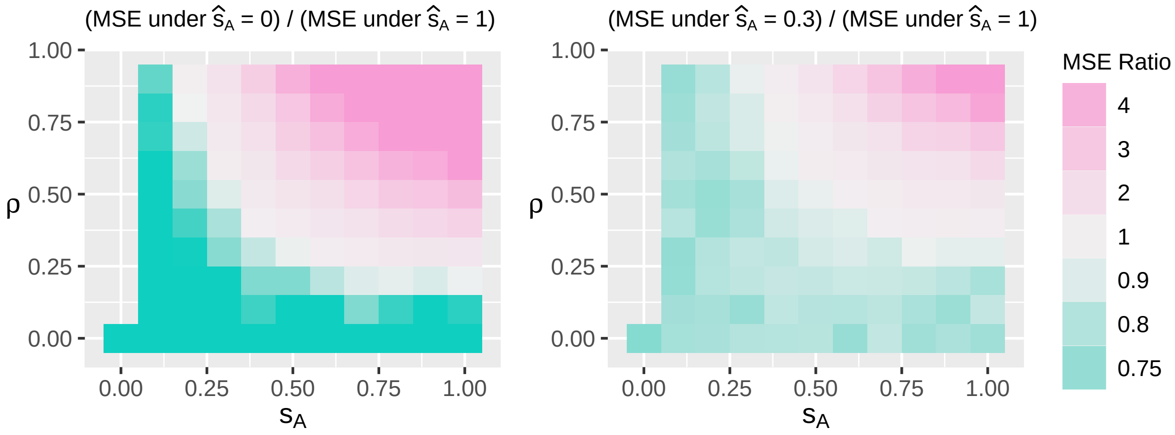

Figure 6: MSE ratio under no treatment effect.

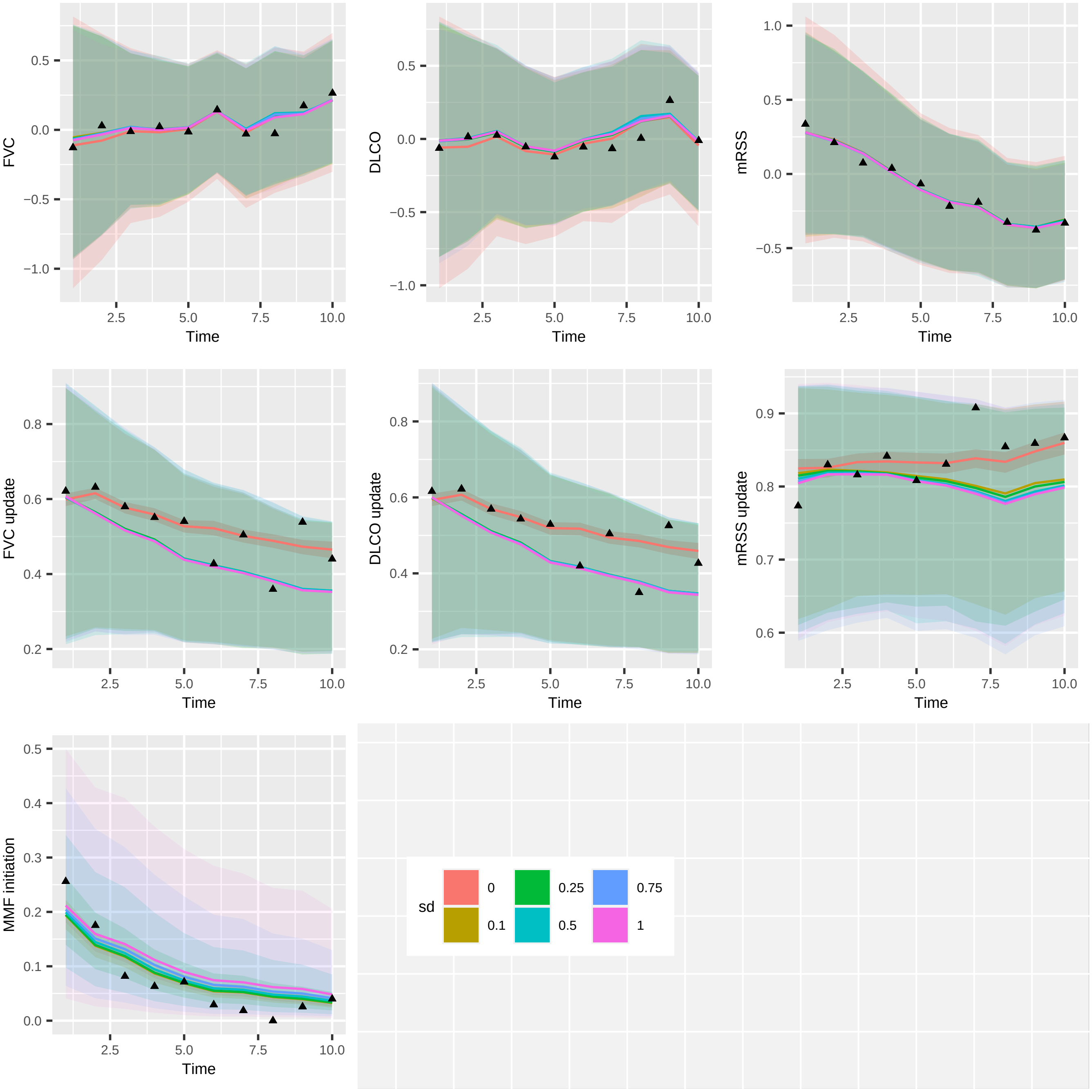

Figure 7: Application accuracy plot.

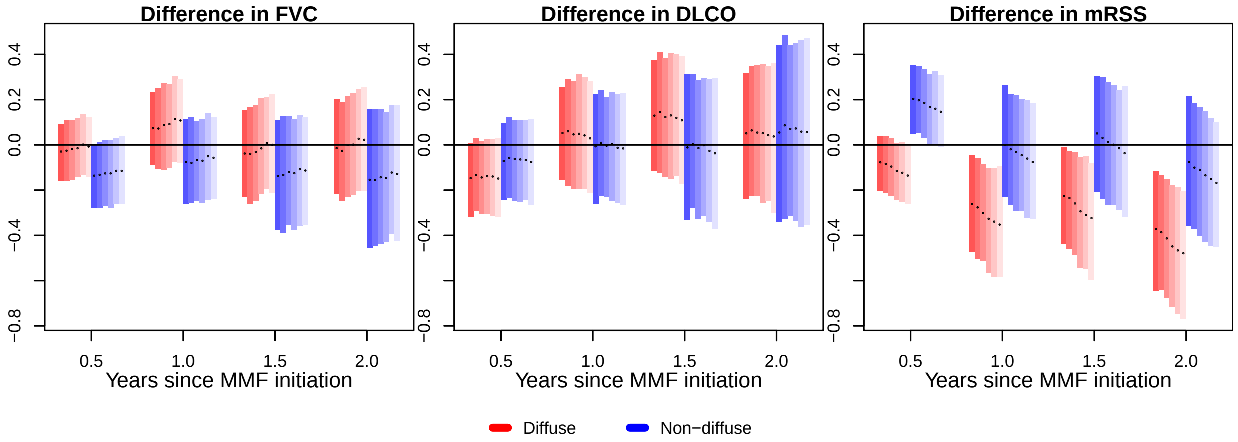

Figure 8: Application causal estimation by subgroup diffuse versus nondiffuse.

Appendices

A Identification of the G-formula

For simplicity, we ignore the subscript i for indexing subjects. Assuming $\overline{A}_{0:h} = \overline{a}_{0:h}(q)$ and time-invariant latent treatment heterogeneity $b_i^A = b_i^A$, the distribution of counterfactual trajectories for the future τ time intervals conditional on observed information up to time h can be processed as follows.

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

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,

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

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,

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

because counterfactual treatment path does not inform heterogeneity estimation,

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

by distributional assumptions illustrated in Figure 1,

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

parameterizing MGLMM as linear models, we get

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

Given b_i^A, the g-formula for a conditional subgroup ATE is defined as a conditional mean of the potential outcome at the end of follow-up at time $h+\tau$ under a user-specified regime q. It can then be derived as below,

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

The population ATE conditional on b_i^A can be obtained by further integrating over the distribution of observed clinical history in the target population,

$$
\mathbb{E}(Y_{h+\tau}(q)|b_i^A) = \int_v \int_{y_h}\int_{m_h} \cdots \int_{y_0}\int_{m_0} \mathbb{E}(Y_{h+\tau}(q)|V = v, \overline{A}_{0:h} = \overline{a}_{0:h}, \overline{Y}_{0:h} = \overline{y}_{0:h}, \overline{M}_{0:h} = \overline{m}_{0:h}, b_i^A)
$$

$$P(V = v, \overline{A}_{0:h} = \overline{a}_{0:h}, \overline{Y}_{0:h} = \overline{y}_{0:h}, \overline{M}_{0:h} = \overline{m}_{0:h}) dm_0 dy_0 \dots dm_h dy_h dv$$

The CMATE is computed as follows by substituting the target population distribution of the observable with the corresponding empirical distribution, $\widehat{P}(V = v, \overline{A}_{0:h} = \overline{a}_{0:h}, \overline{Y}_{0:h} = \overline{y}_{0:h}, \overline{M}_{0:h} = \overline{m}_{0:h})$.

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

Heterogeneity in treatment assignment, b_i^A, is assumed to be marginally $N(0, v)$ in the target population. The marginal mixed population ATE can then be obtained by integrating b_i^A over its distribution $P(b_i^A = w)$ as $\widehat{\mathbb{E}}(Y_{h+\tau}(q)) = \int_w \widehat{\mathbb{E}}(Y_{h+\tau}(q)|b_i^A = w) P(b_i^A = w) dw$.

B Sequential Update of Random Effects

Without loss of generality, assuming Gaussian distribution and logit model for continuous and binary variables, respectively, the structural model can be written as follows ,

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

where $\psi^{Y} \sim N(0, 1)$, $\eta_{it}^{Y} = \eta^{Y}(\mathcal{F}_{it}, b_i^{Y}; \theta^{Y})$, $\eta_{it}^{M} = \eta^{M}(\mathcal{F}_{it}, b_i^{M}; \theta^{M})$, and $\eta_{it}^{A} = \eta^{A}(\mathcal{F}_{it}^{A}, b_i^{A}; \theta^{A})$.

Sequential update for random effects is implemented for each individual, conditional on biomarker dynamics up to time t and observed treatment sequence up to time h, where $h \leq t$. For the observed trajectories of subject i, the joint likelihood is

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

where s_i is the observed treatment initiation time for subject i, and the random effect b_i has prior

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

The log posterior of b_i can then be written as

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

Using algorithms for constructing sampling chains, such as MCMC, in sampling b_i would consume a significant amount of computational resources due to the complexity of calculating counterfactual individual trajectories. We consider a Laplace approximation of the posterior distribution of b_i for an easier posterior sampling. The mean of the approximated distribution is obtained by solving the following equation for a posterior mode $\hat{b}_i = (\hat{b}_i^{Y}, \hat{b}_i^{M}, \hat{b}_i^{A})$,

$$
\frac{\partial}{\partial b_i} \log P(b_i|Y_{i,0:t}, M_{i,0:t}, A_{i,0:h}, \beta, \sigma, G) \bigg|_{b_i = \hat{b}_i} = 0,
$$

where

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

The variance of the approximated distribution is the asymptotic variance of $\hat{b}_i$, which is the inverse of the observed Fisher information matrix defined as follows

$$V = \left[- \frac{\partial^2}{\partial b_i \partial b_i^T} \log P(b_i | Y_{i,0:t}, M_{i,0:t}, A_{i,0:h}, \beta, \sigma, G) \bigg|_{b_i = \hat{b}_i} \right]^{-1},$$

where

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

As a result, an approximation to the posterior distribution $P(b_i | Y_{i,0:t}, M_{i,0:t}, A_{i,0:h}, \beta, \sigma, G)$ is the multivariate Gaussian distribution $MNV(\hat{b}_i, V)$.

Sequential update of counterfactual trajectories is also conditinoal on b_i^A being a constant, i.e. $b_i^A = c$. We sequentially update the heterogeneity in biomarker dynamics conditional on history $(Y_{i,0:t}, M_{i,0:t}, A_{i,0:h})$, population level estimates (β, σ, G), and $b_i^A = c$ as follows

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

such that

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

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

Suppose we are simulating the counterfactual progression of patient's longitudinal measures with treatment sequence fixed as $\overline{a}_{0:t}^{q}$ under regime q, where the sequence up to time h is the observed treatment, i.e. $A_{i,0:h} = \overline{a}_{0:h}^{q}$. If we write the third row of G^{-1} as (C_1, C_2, C_3), then the derivative entry relative to b_i^A leads to

$$\sum_{j=1}^{min(h,s_i)} a_j - \frac{\exp(X_{ij}\beta^A + b_i^A{}_{i0})}{1 + \exp(X_{ij}\beta^A + b_{i0}^A)} = C_1 b_i^A + C_2 b_i^M + C_3 b_i^Y, \tag{11}$$

and we can see that the specification of the counterfactual treatment sequence $\overline{a}_{(h+1):t}$ does not affect the estimation of $\hat{b}_i$. Note that $\sum_{j=1}^{min(h,s_i)} a_j$ is either 0 or 1, because the summation stops at the time of initiation. In the application, we focus on studying the effect of treatment initiation among those who were not treated before a time h, i.e. $h < s_i$ and $\sum_{j=1}^{min(h,s_i)} a_j = 0$. Hence, for the estimation of $\hat{b}_i$, equation (11) imposes condition $-\frac{\exp(X_{ij}\beta^A + b_i^A{}_{i0})}{1+\exp(X_{ij}\beta^A + b_i^A{}_{i0})} = C_1 b_i^A + C_2 b_i^M + C_3 b_i^Y$, using only treatment information before an treatment initiation.

C Pseudocode for Generating Counterfactual Trajectories

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) $\mathcal{F}^{(\ell)}_{i,h+1}(q) = (V_i, \overline{Y}_{i,0:h}, \overline{M}_{i,0:h}, \overline{A}_{i,0:h}, a_{h+1}(q))$ for all $\ell$
    (c) for each $\ell$, 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 $(V_i, \overline{Y}^{(\ell)}_{i,0:(t-1)}(q), \overline{M}^{(\ell)}_{i,0:(t-1)}(q), \overline{A}_{i,0:h})$
            **Step 2**: Draw $(b_i^{Y(\ell)}(q), b_i^{M(\ell)}(q)) | b_i^{A(\ell)} \sim MVN(b^{(\ell)}_{t|A}(q), V^{(\ell)}_{t|A}(q))$, where
                $b^{(\ell)}_{t|A}(q)$ and $V^{(\ell)}_{t|A}(q)$ 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 $M_{it}^{(\ell)}(q) = \mathbb{1}\{\psi^M_{\ell,t-h} \leq p_{it}^{(\ell)}(q)\}$
            **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 $Y_{it}^{(\ell)}(q) = \eta_{it}^{Y(\ell)}(q) + \sigma^{(\ell)} \psi^Y_{\ell,t-h}$
            **Step 5**: Define

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

                where

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

                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**: $\{\overline{Y}^{(\ell)}_{i,(h+1):T}(q), \overline{M}^{(\ell)}_{i,(h+1):T}(q); \ell = 1, \ldots, N_{post})\}$ 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-formula in Simulation

$$
\begin{aligned}
\mathbb{E}[Y_{i2}(q)] =& \iint y_2 P(Y_{i1}(q) = y_1, Y_{i2}(q) = y_2 | V = v) P(V = v) dv dy_1 dy_2 \\
=& \iint y_2 P(Y_{i1}(q) = y_1, Y_{i2}(q) = y_2 | V = v, b_i^A = w) P(b_i^A = w) P(V = v) dw dv dy_1 dy_2 \\
=& \iint y_2 P(Y_{i1}(q) = y_1, Y_{i2}(q) = y_2 | A_{i1} = a_1, V = v, b_i^A = w) P(b_i^A = w) P(V = v) dw dv dy_1 dy_2 \\
=& \iint y_2 P(Y_{i1} = y_1, Y_{i2}(q) = y_2 | A_{i1} = a_1, V = v, b_i^A = w) P(b_i^A = w) P(V = v) dw dv dy_1 dy_2 \\
=& \iint y_2 P(Y_{i2}(q) = y_2 | A_{i1} = a_1, Y_{i1} = y_1, V = v, b_i^A = w) \\
& \quad P(Y_{i1} = y_1 | A_{i1} = a_1, V = v, b_i^A = w) P(b_i^A = w) P(V = v) dw dv dy_1 dy_2 \\
=& \iint y_2 P(Y_{i2}(q) = y_2 | A_{i1} = a_1, A_{i2} = a_2, Y_{i1} = y_1, V = v, b_i^A = w) \\
& \quad P(Y_{i1} = y_1 | A_{i1} = a_1, V = v, b_i^A = w) P(b_i^A = w) P(V = v) dw dv dy_1 dy_2 \\
=& \iint y_2 P(Y_{i2} = y_2 | A_{i1} = a_1, A_{i2} = a_2, Y_{i1} = y_1, V = v, b_i^A = w) \\
& \quad P(Y_{i1} = y_1 | A_{i1} = a_1, V = v, b_i^A = w) P(b_i^A = w) P(V = v) dw dv dy_1 dy_2 \\
=& \mathbb{E}(Y_{i2} | A_{i1} = a_1, A_{i2} = a_2).
\end{aligned}
$$

Supplementary Material

A Connection with Structural Nested Models

Sitlani et al.[33] and Qian et al.[23] studied instantaneous treatment effect as the "blip" of a structural nested model (SNM), using linear mixed models as the structural model and comparing treatment paths that only differ in the treatment status at a specific time m, i.e. comparing $A_m = 1$ versus $A_m = 0$ in the case of binary and monotone treatment. Our proposal, on the other hand, compares the effect of treatment paths under different regimes, i.e. $\overline{A}_{0:t}$ being $\overline{a}_{0:t}(q_1)$ versus $\overline{a}_{0:t}(q_2)$, where q_1 and q_2 are the regimes of interest. The motivating application investigates the treatment effect of taking a drug continuously over time, where the causal effect is cumulative over time and thus requires a fundamentally different characterization than a structural model approach. The instantaneous treatment effect, or the blip, can be characterized under our framework as the average causal effect comparing q_1 and q_2 where $a_t(q_1) = a_t(q_2)$ for $t \neq m$, $a_m(q_1) = 1$ and $a_m(q_2) = 0$. Specifically, assuming $\phi_A(\overline{A}_{i,0:t}) = A_{it}$, $\phi_4^Y(\mathcal{H}_{it}) = 0$, $\tau = 1$, and a linear mixed model for a continuous outcome leads to a special case in Qian et al.[23] , where we will have the instantaneous subgroup treatment effect at $h+1$ conditional on information up to time h being

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

Thus, the model parameter β_2^Y has a causal interpretation marginally over the subgroup defined by $(V_i, \overline{A}_{i,0:h}, \mathcal{H}_{ih})$ in this case and the MGLMM reduces to a linear structural mixed model. However, when $\phi_4^Y(\mathcal{H}_{it}) \neq 0$, equation (12) is no longer true because β_2^Y only remains with a causal interpretation conditional on b_i^Y, as showed in the conditional subgroup causal effect below,

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

and the conditional expectation $\mathbb{E}(b_i^Y|V_i, \overline{A}_{i,0:h}, \mathcal{H}_{ih})$ is not necessarily zero.

B Connection with Shardell and Ferrucci[32]

Shardell and Ferrucci[32] demonstrated longitudinal causal inference using joint mixed-effects models, assuming shared random effects between the model components for the outcome, confounders, and treatment assignment. Their model specification is similar to the MGLMM in Section 2, i.e. with $b_{i0}^A = b_{i0}^M = (b_{i0}^Y, b_{i1}^Y)$ and $\phi_4^M \equiv 0$, but distinctively different in that ϕ_2^A and ϕ_3^M are population-level coefficients instead of observed variables. Shardell and Ferrucci[32] assumed sequential exchangeability conditional on the unobserved heterogeneity in the outcome progression, i.e. (b_{i0}^Y, b_{i1}^Y), which is assumed to be proportionate to the heterogeneity in confounders and treatment assignment. Whereas we account for unobserved time-invariant traits in treatment assignment with the random effect b_i^A, assuming that it is correlated with (b_{i0}^Y, b_{i1}^Y) and having the sequential exchangeability conditional on b_i^A instead of (b_{i0}^Y, b_{i1}^Y).

The assumption of no unmeasured confounders in the model of Shardell and Ferrucci[32] implies no treatment assignment heterogeneity. While assuming no treatment heterogeneity under MGLMM is equivalent to setting $v = 0$, which is a sufficient but unnecessary condition for having no unmeasured confounders. In MGLMM, $\text{cov}(b_i^A, b_i^M) = \text{cov}(b_i^A, b_i^Y) = 0$ leads to no unmeasured confounders. That is, even when no unmeasured confounders is true, MGLMM still allows treatment assignment heterogeneity as long as it is not correlated with the unobserved heterogeneity in biomarker dynamics (b_i^Y, b_i^M); examples of such unconfounding treatment assignment heterogeneity include a patient's preference for a treatment based on personal beliefs or social stigma. On the other hand, we note that the covariances $\text{cov}(b_i^A, b_i^M)$ and $\text{cov}(b_i^A, b_i^Y)$ are estimable given v, the presumed variance of b_i^A. Henceforth, our method does partially inform the possible existence of unmeasured confounders based on the estimated covariances in the MGLMM.

When there is no treatment assignment heterogneiety, both Shardell and Ferrucci[32] and our method simplify to the standard g-computation of fitting only the outcome and confounders model using generalized linear mixed-effects model because the assignment mechanism becomes ignorable[19]. Let us consider a simplified scenario of looking at the subgroup ATE at time $h + 1$ conditional on history information up to time h, assuming no time-varying confounders and no treatment assignment heterogeneity. The subgroup ATE would not be identifiable under Shardell and Ferrucci[32]. The reason is as follows. Under their conditional sequential exchangeabiltiy assumption

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

we can directly identify the conditional counterfactual distribution as

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

However, the target quantity represented by $P(Y_{h+1}(q)|V, \overline{A}_{0:h}, \overline{Y}_{0:h})$ would not be calculable because

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

and the subgroup heterogeneity distribution $P(b_i^Y|V, \overline{A}_{0:h}, \overline{Y}_{0:h})$ is unknown. Whereas with our proposal, we assume a different conditional sequential exchangeabiltiy assumption

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

Given the assumption of no treatment assignment heterogeneity, we know var$(b_i^A) = 0$ and consequently cov$(b_i^A, b_i^Y) = 0$, leading to $P(b_i^Y|V, \overline{A}_{0:h}, \overline{Y}_{0:h}, b_i^A) = P(b_i^Y|V, \overline{A}_{0:h}, \overline{Y}_{0:h})$. As a result, the target quantity is identifiable via (6) as

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

Our proposal may be viewed as an extension of Shardell and Ferrucci's[32] work in the following aspects: (1) a softer assumption on the conditional sequential exchangeability, stratifying by b_i^A instead of the random effects shared across the outcome, confounders, and treatment model, (2) model specification as MGLMM, which is more generalized and has the potential to include their joint mixed-effects model as a special case, and (3) allows the identification of subgroup causal effects when assuming no treatment assignment heterogeneity. The merits of this extension come at a price of introducing the variance of b_i^A as a sensitivity parameter, and identifying subgroup causal effects under the existence of treatment assignment heterogeneity needs to be done under additional assumption on the subgroup distribution of b_i^A, which likely requires expert knowledge.