A Simulation Study to Validate Joint Models Fitted for Longitudinal and Survival Data

Nam-Anh Tran

1 Introduction

Longitudinal data offer a unique lens for tracking biological processes over time, capturing both within-subject dynamics and between-subject variability. Linear mixed models are well-suited for such analyses, disentangling population-level trends (fixed effects) from individual deviations (random effects) to enable nuanced interpretation. Time-to-event outcomes, such as survival or disease progression, are critical for identifying the timing of key clinical events; however, analyzing them in isolation can overlook valuable information from longitudinal trajectories. Joint modelling addresses this by integrating longitudinal and survival data, capturing the evolving nature of biomarkers and their influence on event risk. To fully understand the strengths and limitations of this approach across different scenarios, a thorough investigation of its performance is essential.[1,2]

Simulation studies are integral to validating the performance of joint models, particularly when integrating longitudinal and time-to-event data. By constructing controlled datasets with known parameter values, we can rigorously evaluate how accurately the joint model reconstructs individual longitudinal trajectories and associated event risks. We can also illuminate the interactions between the longitudinal and survival submodels, thereby guiding enhancements in model formulation and inference. Such comprehensive validation is essential prior to the application of the joint model to complex, real-world biomedical data, ensuring reliable, personalized risk assessments and informed clinical decision-making.

The primary objective of this study is to evaluate the performance and robustness of a Bayesian joint model through the simulation of datasets that encompass both longitudinal and time-to-event outcomes. By generating controlled and synthetic data, we assess the model's capability to accurately capture the evolution of longitudinal trajectories alongside the associated event risks. The model is fitted using a Bayesian framework, which offers distinct advantages.

While Bayesian linear mixed models are widely used for analysing longitudinal data, survival models are still predominantly estimated using Frequentist methods. In contrast, Bayesian approaches, particularly those adapting counting process formulations, are rarely applied to survival modelling. However, the Bayesian framework offers several distinct advantages: it enables full probabilistic inference through posterior distributions, facilitates uncertainty quantification, allows incorporation of prior knowledge, and provides greater flexibility in modelling complex time-to-event structures. Despite these strengths, the application of Bayesian methods in survival analysis remains limited, and comprehensive validation—especially in the context of joint modelling with longitudinal data—has not been fully explored. This study addresses that gap by evaluating both the survival component and its integration with longitudinal data under a Bayesian joint modelling framework.[2,3]

The remainder of this report is organized as follows: we first present the joint model, outlining its theoretical foundation and motivation for integrating longitudinal and time-to-event data. We then describe the simulation study, including the generation of realistic synthetic data and the Bayesian framework used for model fitting. Results are presented and discussed in relation to existing literature, highlighting key findings and limitations. The report concludes with a summary of insights and directions for future research.

2 Joint longitudinal and survival models

Joint models include two sub-models: longitudinal and event time components. The longitudinal sub-model is expressed as follows:

$$y_i(t) = \boldsymbol{x}_i'(t)\boldsymbol{\beta} + \boldsymbol{z}_i'(t)\boldsymbol{b}_i + \epsilon_i(t),$$

where $\{y_i(t)\}_{i=1}^{n}$ denotes the longitudinal outcome for ith subject at time t. The vector $\boldsymbol{x}_i(t)$ is a p-dimensional covariate vector associated with ith subject at time t. We frequently set the first element of the vector to 1 to account for the intercept. This vector includes both time-invariant and time-varying covariates. We can define $x_i(t) := x_i, \forall t$ to indicate the baseline covariate. The vector $\boldsymbol{\beta}$ represents the corresponding fixed effects, including the fixed intercept and regression coefficients. Similarly, $\boldsymbol{z}_i(t)$ is a covariate vector used for modelling random effects. $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_i(t)$ are not necessarily identical. The associated vector $\boldsymbol{b}_i \sim N(0, \Sigma_b)$ captures the subject-specific random effects. Finally, $\epsilon_i(t) \sim N(0, \sigma^2)$ denotes the residual error after accounting for both fixed and random effects.

To simplifiy the process, we set $\boldsymbol{x}_i(t) = (1, \text{Treatment})$ and $\boldsymbol{z}_i(t) = (1, \text{Time})$, where `Treatment` is a binary variable representing two arms of the study; `Time` is the discrete time point at which the longitudinal outcome is measured. The above model formula is rewritten as

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 \text{Treatment}_i + b_{0i} + b_{1i}\text{Time}_{ij} + \epsilon_{ij} \\ &\equiv m_i(t) + \epsilon_{ij}, \end{aligned} \tag{1}$$

where $m_i(t)$ is known as the trajectory function. The event time submodel is formulated as follows:

$$h_i(t) = h_0(t) \exp\Big\{\alpha_2 m_i(t) + \boldsymbol{w}_i' \boldsymbol{\alpha}_1\Big\},$$

where $h_i(t)$ denotes the hazard of an event for the ith subject at time t, and $h_0(t)$ represents the baseline hazard. The trajectory function serves as the link between the longitudinal and event-time components of the joint model. The parameter α_2 quantifies the effect of the longitudinal trajectory on the hazard function. Vector $\boldsymbol{w}_i$ contains covariates specific to the ith subject, with $\boldsymbol{\alpha}_1$ denoting the corresponding coefficients. In this study, we assume that the covariates $\boldsymbol{w}_i = \text{Treatment}_i$ consist exclusively of baseline treatment, and that $m(t)$, the conditional mean of the longitudinal outcome given $\boldsymbol{b}$, is the only time-varying covariate. While α_2 represents the indirect treatment effect on $h(t)$, as measured through longitudinal outcome, α_1 represents the direct treatment effect on $h(t)$. The event time submodel is then redefined as follows.

$$h_i(t) = h_0(t) \exp \Big\{ \alpha_2 m_i(t) + \alpha_1 \text{Treatment}_i \Big\}. \tag{2}$$

It is essential to note that the trajectory function $m(t)$ can take various forms, often specified based on the underlying research objectives.[2,4] In this study, we adopt one of the earliest and most interpretable forms of the trajectory function, which facilitates a straightforward understanding of the overall treatment effect.[5] Moreover, the joint modelling framework is not restricted to linear mixed models (as in Equation Equation 1,) it can be extended to generalized linear mixed models to accommodate discrete longitudinal outcomes. Such extensions have been previously explored by Faucett et al.[6] and Li et al.[7]

The joint model is estimated within a Bayesian framework. While the Bayesian formulation of the longitudinal submodel has been extensively studied in the literature, the survival submodel in this work is modelled using a Gamma process, a less commonly applied approach in Bayesian survival analysis. We briefly summarise the framework of Gamma process, which we adopt for fitting Cox model. Details can be found in Ibrahim et al.[8] Under the Cox model, the joint probability of survival of n subjects is

$$P(\boldsymbol{Y}^{(survival)} > \boldsymbol{y}^{(survival)} | \boldsymbol{\alpha}, X, H_0) = \exp\Bigg\{ - \sum_{j=1}^{n} \exp(\boldsymbol{x}_j' \boldsymbol{\alpha}) H_0(y_j^{(survival)}) \Bigg\}$$
$$H_0 \sim \text{GammaProcess}(c_0 H^*, c_0).$$

Under assumption of exponential distribution, we have $H^*(y^{(survival)}) = \gamma_0 y^{(survival)}$. We define

$$h_j \equiv H_0(s_j) - H_0(s_{j-1}) \sim Gam(\kappa_{0j} - \kappa_{0,j-1}, c_0),$$

where $\kappa_{0j} = c_0 H^*(s_j)$. Thus,

$$h_j \sim Ga(c_0 \gamma_0 (s_j - s_{j-1}), c_0), \tag{3}$$

where c_0 and γ_0 are hyper-parameters, and $0 < s_1 < s_2 < \cdots < s_J$ with $s_J > y_i^{(survival)} \forall i$. Hence, the likelihood function is

$$L(\boldsymbol{\alpha}, \boldsymbol{h} | D) \propto \prod_{j=1}^{J} G_j,$$

where $\boldsymbol{h} = (h_1, \ldots, h_J)'$ and

$$G_j = \exp\left\{ -h_j \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp \boldsymbol{x}'\boldsymbol{\alpha} \right\} \prod_{l \in \mathcal{D}_j} \left[1 - \exp\{-h_j \exp(\boldsymbol{x}'\boldsymbol{\alpha})\} \right],$$

where $\mathcal{R}_j$ and $\mathcal{D}_i$ are the risk and event set at time j. Since H_0 enters the likelihood only through the h_j's, parameters in the likelihood are $(\boldsymbol{\alpha}, \boldsymbol{h})$. As both the longitudinal and survival time submodels are conditional independent given random effect, the complete data log-likelihood is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \left\{ \phi(y_i^{(longitudinal)} | \boldsymbol{b_i}, \boldsymbol{\beta}, \sigma^2) \phi(\boldsymbol{b}_i | \Sigma_b) \phi(y_i^{(survival)} | \boldsymbol{b}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) \right\},$$

where $\phi(.|.)$ denotes the appropriate normal probability density function.

The Bayesian models are completed after defining the prior distribution. We define the priors as follows.

$$\begin{aligned} &\{\beta_k\}_{k=1}^{2} \sim N(0, 10), \quad b_{0i} \sim N(0, 10), \quad b_{1i} \sim U(0, 4) \\ &\epsilon_{ij} \sim N(0, \sigma^2), \quad \sigma \sim N^{+}(0, 5) \end{aligned} \tag{4}$$

and the hyperparameters in Equation 3 are

$$\begin{aligned} &\alpha_1 \sim N(0, 10), \quad \alpha_2 \sim N(0, 3) \\ &\gamma_0 = 0.1, \quad c_0 = 0.01, \end{aligned} \tag{5}$$

suggested by Ibrahim et al.[8]

All models were implemented in the R environment using `stan`, which performs full Bayesian inference via Hamiltonian Monte Carlo. For each model, 3,000 posterior samples were drawn following 4,000 warm-up iterations. Convergence was assessed using the potential scale reduction statistic $\hat{R}$, with values below 1.1 indicating satisfactory convergence.

3 Simulations

Data are simulated based on four steps:

1. We simulate time for each subject;
2. The longitudinal observations are generated for each subject;
3. We remove all longitudinal observations with time beyond survival time;
4. The censors are defined as the event times beyond the 60% quantile.

First, we assume that the event time follows an exponential distribution and employ the method proposed by Austin et al.[9], which enables the event time to be generated using a closed-form expression. The authors define the hazard function as $h_0(t)\exp\left\{\beta_t z(t) + \beta' x\right\}$, where $z(t)$ is a time-varying covariate. They assume that $z(t)$ is proportional to time, specifically $z(t) = \nu t$, with $\nu > 0$. Under this assumption, the authors derived the closed-form expression for event time as

$$T = \frac{1}{\beta_t \nu} \ln\left[1 + \frac{\beta_t \nu(-\ln u)}{\lambda \exp(\beta x)}\right],$$

where λ is the scale parameter of the exponential distribution and u is a random draw from the uniform distribution on (0,1). We note that the closed-form is obtained based on the integral with respect to the time t, which is a linear predictor. This suggests that the event time in our model (Equation 1 and Equation 2) can also be generated using a closed-form expression. We adopt this idea to show that the closed-form, used to simulate our survival time of our event time submodel, has the following form.

$$T = \ln\left[\frac{\alpha_2 \beta_2(-\ln u)}{\lambda \exp \alpha_2 k + \alpha_1 \text{treatment}} + 1\right] \cdot \frac{1}{\alpha_2 \beta_2}, \tag{6}$$

the derivation of this closed-form expression is provided in Appendix Section A.1.

It is important to note that this approach performs well only when the time-varying covariate is a well-defined function of time. In our setting, this condition holds, as the longitudinal model includes only individual-level covariates. However, if within-individual (time-dependent) covariates were incorporated, a more sophisticated method would be required. Hendry[10] proposed an approach for generating event times under time-dependent covariates using truncated piecewise exponential models. While this method aligns more closely with our event-time submodel—fitted using a piecewise constant hazard function—the simpler method proposed by Austin remains sufficient for our current framework.

Furthermore, while Hendry's method offers a natural simulation flow—first generating longitudinal data and then simulating survival times based on those trajectories—Austin's approach follows the reverse order. This reversal arises because the closed-form expression in Equation 6 is independent of time, as it is derived from integrating the exponential distribution over time t. Consequently, our simulation begins with generating event times, followed by the simulation of longitudinal measurements. As the longitudinal outcomes are accurately estimated, the generated event times more closely reflect their true underlying distribution (i.e., the exponential distribution), thereby improving the accuracy of coefficient estimates in the survival submodel.

We simulate data using the following setting for fixed and random effects.

$$\begin{aligned} &\beta_0 = 0.5, \quad \beta_1 = -1.2, \quad \lambda = 0.05 \\ &b_0 \sim N(0,1), \quad b_1 \sim U(1,3), \quad \epsilon \sim N(0,1) \\ &\alpha_2 = 0.8, \quad \alpha_1 = -0.5. \end{aligned} \tag{7}$$

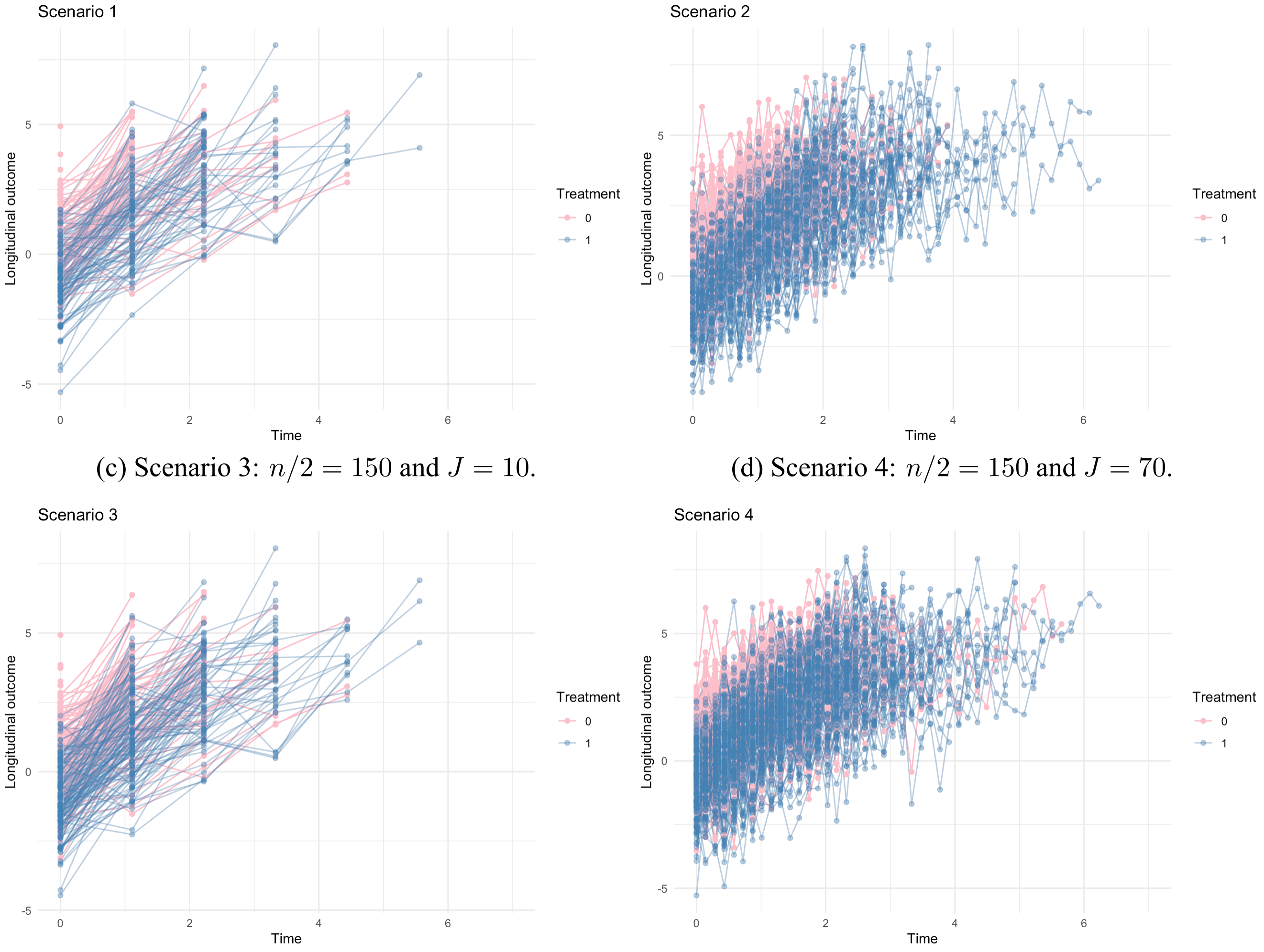
This setting implies that treatment decreases the longitudinal outcome (β_1 is negative); the considered disease is worse over time (b_1 is non-negative); thus, as longitudinal outcome decreases, the hazard must decrease (α_2 is positive); finally, treatment also decreases hazard (α_1 is negative.)

We simulate 4 scenarios:

- 100 subjects per arm + 9 intervals ($J = 10$ time points).
- 100 subjects per arm + 69 intervals ($J = 70$ time points).
- 150 subjects per arm + 9 intervals ($J = 10$ time points).
- 150 subjects per arm + 69 intervals ($J = 70$ time points).

Based on these scenarios, we evaluate ability of parameter recovery of the joint model, characterized by different sample sizes, reflected by two levels: subjects and the longitudinal observations of each subject. Figure 1 shows trajectories of longitudinal outcomes of all subjects in the first generated dataset in four scenarios.

Figure 1: Trajectories of longitudinal outcomes in four scenarios.



(a) Scenario 1: $n/2 = 100$ and $J = 10$.

(b) Scenario 2: $n/2 = 100$ and $J = 70$.

(c) Scenario 3: $n/2 = 150$ and $J = 10$.

(d) Scenario 4: $n/2 = 150$ and $J = 70$.

4 Results

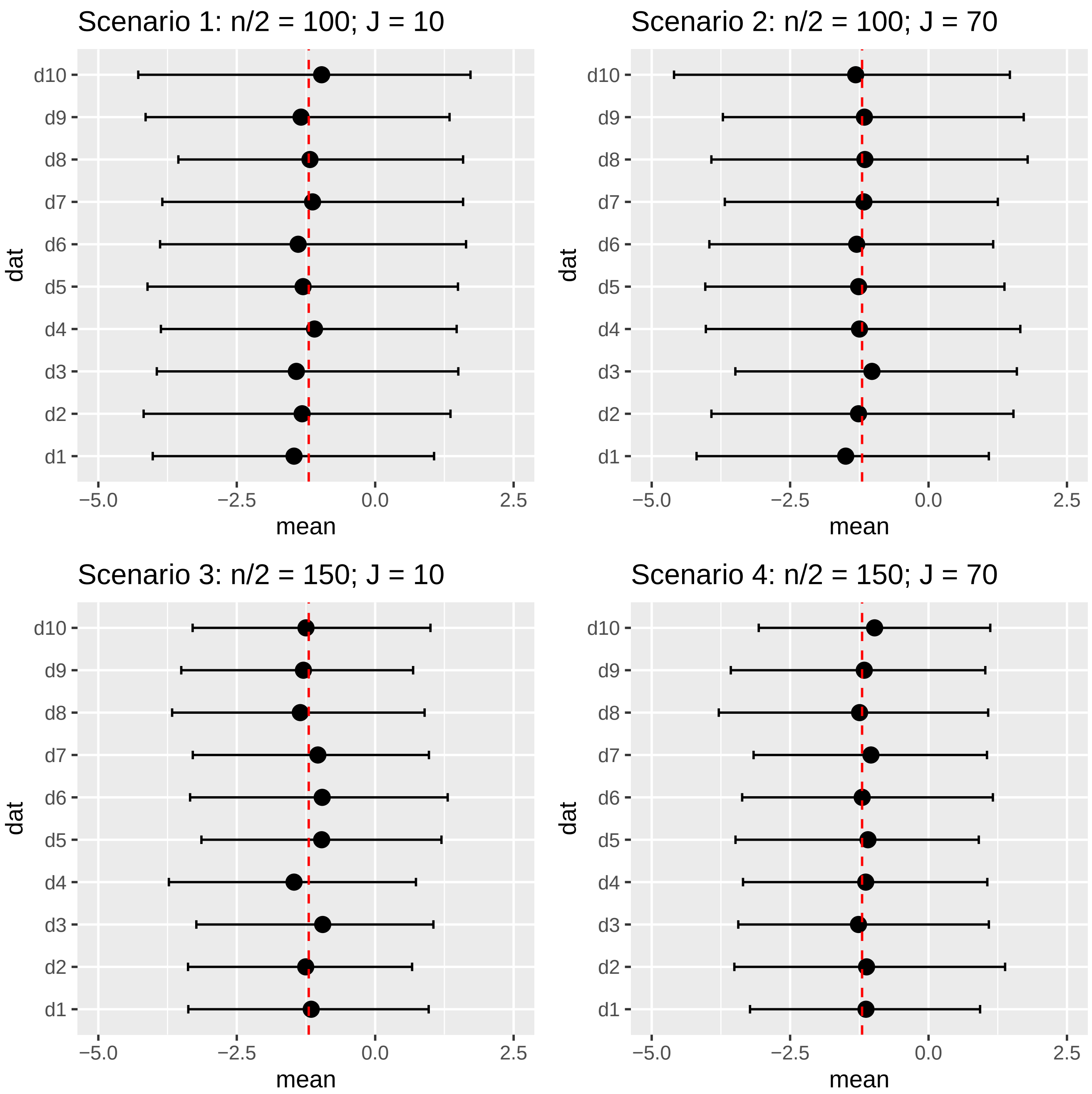
Figure 2 shows the means and 95% highest density intervals (HDIs) of β_1, the treatment effect on the longitudinal outcome derived from ten simulated datasets. The means of the treatment effect are distributed around the true value, indicated by the vertical red line, exhibiting minimal bias. This observation suggests that the true treatment effect is effectively recovered. Furthermore, as the number of subjects increases from 100 to 150 per arm, the posterior uncertainty diminishes (indicated by narrower intervals,) thereby enhancing the precision of the treatment effect estimation.

Figure 3 presents the mean and 95% HDI's of the overall treatment effect on event time, formulated as $\beta_1\alpha_2 + \alpha_1$. As the number of subjects increases, the posterior uncertainty diminishes; however, the bias remains constant (compare scenarios 1 to 3). This constancy arises because we still regress on $m(t)$ observed with error (or biased toward its mean) due to lack of longitudinal outcomes, and thus the estimate of the slope α_2 is not improved, thereby remaining a biased overall treatment effect. Further, although the increase in the number of longitudinal outcomes results in a reduction of bias in the overall treatment effect but the posterior variance increases (scenarios 2 and 4 have wider intervals compared to that in scenarios 1 and 3). It is crucial to acknowledge that the low variance (i.e. narrow intervals) observed in scenarios 1 and 3, with fewer longitudinal observations does not indicate high precision but rather an underestimation of variance. Let us denote the overall treatment effect as $\kappa = \beta_1\alpha_2 + \alpha_1$, variance of κ is then

$$\mathrm{V}(\kappa) = \mathrm{V}(\alpha_1) + \beta_1^2\mathrm{V}(\alpha_2) + \ 2\,\beta_1\,\mathrm{Cov}(\alpha_1, \alpha_2). \tag{8}$$

When the number of longitudinal observations is small, the posterior distributions for each b_{0i} and b_{1i} are strongly pulled toward their prior means (zero), reflecting classical hierarchical shrinkage. As a result, the subject-to-subject variability is considerably reduced, indicating a small between-subject variance in $m(t)$. Consequently, when $m(t)$ is substituted into the event time model to estimate the indirect treatment effect α_2, the variance of α_2 approaches its prior variance. Moreover, the covariance term in Equation 8 shrinks toward zero. This occurs because $m_i(t)$ no longer differentiates between individuals, and thus the survival likelihood becomes insensitive to any relationship between treatment (reflected by α_1) and

Figure 2: The forest plots show the means and 95% HDIs of β_1 across 10 simulated datasets.



the longitudinal outcome (reflected by α_2). In effect, as α_2 becomes nearly constant, the covariance term vanishes, leading to a reduction in $\mathrm{V}(\kappa)$.

In contrast, when more observations are available for each subject, the variance of $m(t)$ is estimated more accurately, providing more information to estimate α_2, decreasing the uncertainty. Despite the decrease of $\mathrm{V}(\alpha_2)$, the covariance term in Equation 8 remains substantial and positive, contributing additional variability. As a result, the variance of the net effect κ is actually larger than in the "few-observation" case. The reduction in $\mathrm{V}(\alpha_2)$ in the "many-observation" scenario arises because this variance reflects two sources of uncertainty: one derived from the data and the other from the prior, i.e.,

$$\mathrm{V}(\alpha_2 \mid \text{data}) = \frac{1}{I(\alpha_2) + 1/\sigma^2_{\alpha_2}}, \tag{9}$$

where $I(\alpha_2)$ denotes the Fisher information of α_2 derived from the data, and $\sigma^2_{\alpha_2}$ represents the prior variance. Thus, as the number of observations per subject increases, I becomes large, resulting in a reduction of $\mathrm{V}(\alpha_2)$. This also explains why $\mathrm{V}(\alpha_2)$ tends to shrink toward the prior variance (i.e. $\sigma^2_{\alpha_2}$) when the number of observations per subject is small (i.e. $I \to 0$.)

Although the overall treatment effect, expressed as $\beta_1\alpha_2 + \alpha_1$, is sufficiently recovered, the individual estimates of the indirect effect (α_2) and the direct effect (α_1) on event time exhibit noticeable bias. As illustrated in Figure 4, the posterior means and 95% HDIs for α_2 reveal a consistent underestimation across four distinct scenarios. Increasing the number of longitudinal observations per subject (as seen in scenarios 2 and 4) reduces this bias; however, it does not eliminate it entirely.

Moreover, although the estimation of α_2 improves with additional longitudinal data, the variability of α_1 across simulated datasets increases. Figure 5 shows the means of α_1 distribute about the true value with higher dispersion in scenario 2 and 4. This pattern arises because, in the "low-observations" setting, α_2 is strongly pulled toward its prior mean of zero, resulting in α_1 being estimated largely independently of α_2. This independence leads to relatively low and stable sampling variability in α_1. In contrast, when sufficient data is available to inform the estimate of α_2, the model must allocate the treatment effect between the direct and indirect pathways. However, the event time model only identify $\alpha_2\beta_1 + \alpha_1$ as the treatment effect on survival outcome. Thus, the model might not well capture two treatment effects sep-

Figure 3: The forest plots show the means and 95% HDIs of the overall treatment effect $\beta_1\alpha_2 + \alpha_1$ across 10 simulated datasets.

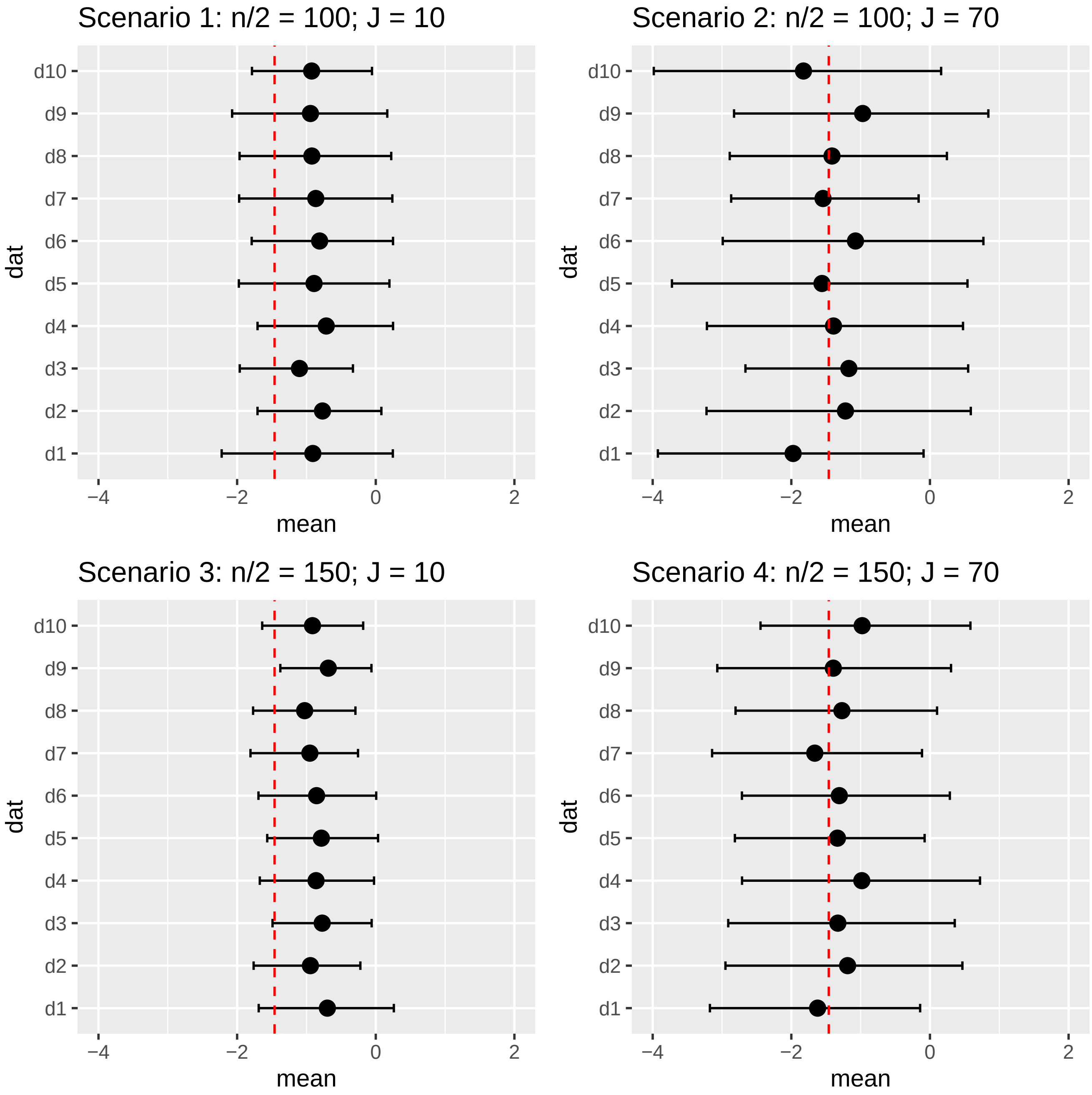
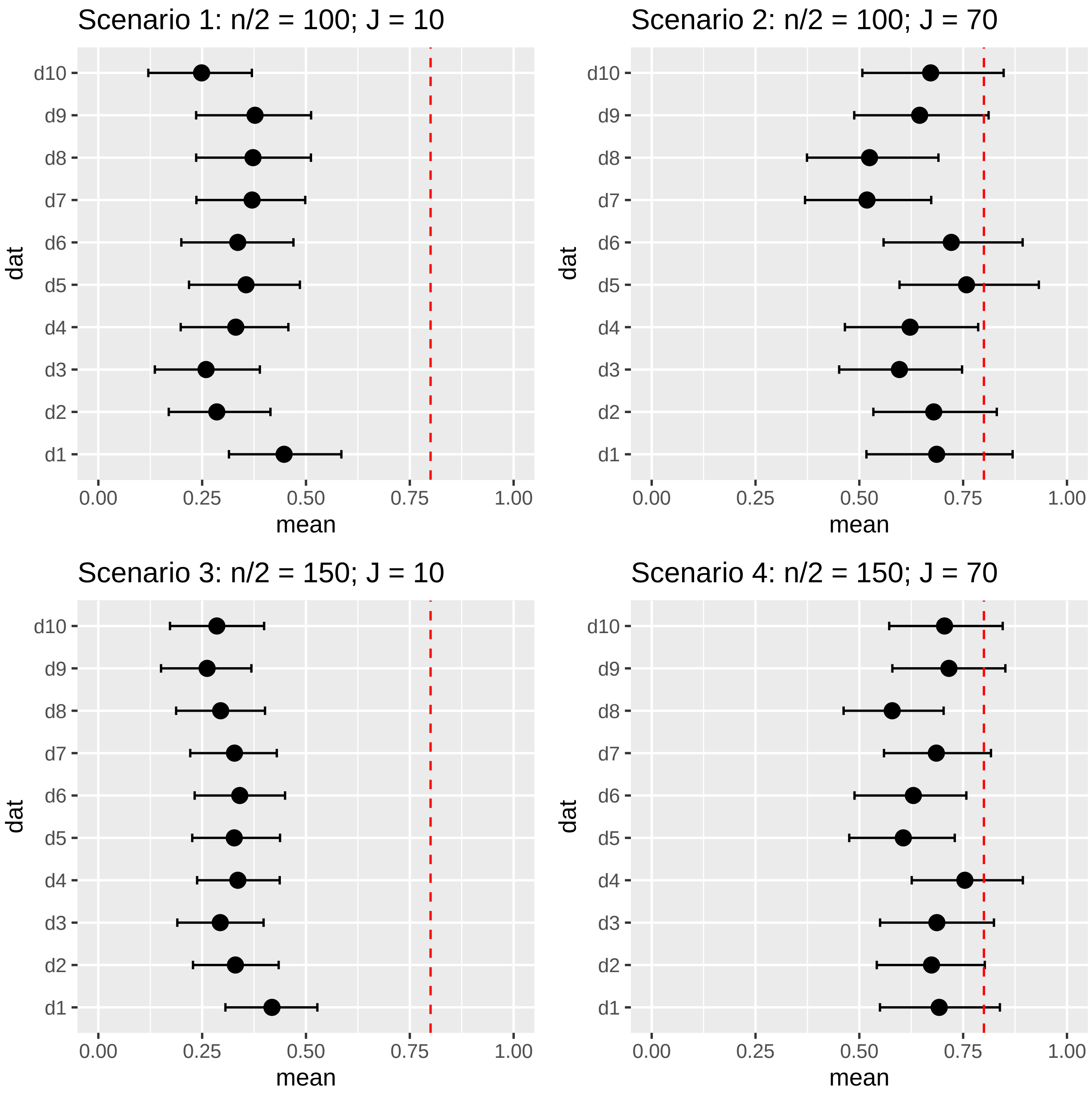


Figure 4: The forest plots show the means and 95% HDIs of the indirect treatment effect α_2 across 10 simulated datasets.



arately. This introduces greater sampling uncertainty (i.e., among simulated datasets), thereby increasing the variability of α_1 across simulations.

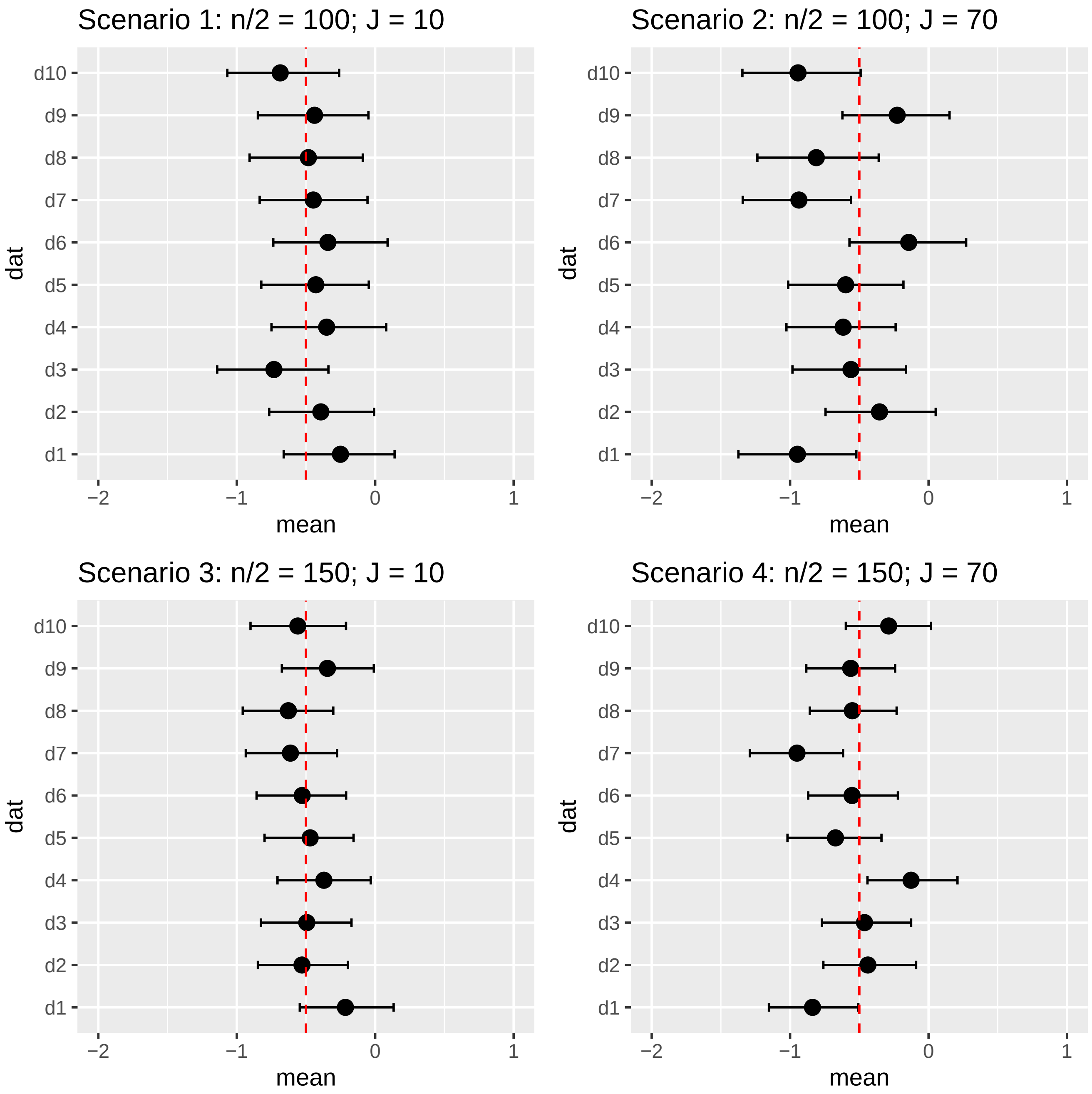
5 Discussion

This study evaluates the performance of a joint model for longitudinal and time-to-event data. The results show that both the treatment effect on the longitudinal outcome and the overall treatment effect on event time are accurately recovered. However, the model exhibits bias in estimating individual-level direct and indirect treatment effects on the time-to-event outcome. Specifically, although the longitudinal trajectories are well estimated, the survival submodel struggles to disentangle the direct and indirect pathways of treatment influence. Increasing the sample size may help reduce this bias and improve the precision of effect decomposition.

This study considered a simplified model setting that included only a baseline treatment covariate, with the time-varying covariate specified as a well-defined function of time. However, this assumption may be unrealistic, as models used in practice often involve greater complexity, for example, the functional form of time-varying covariates and time is typically unknown. A more applicable approach would involve generating data under a piecewise constant hazard framework, which explicitly captures the dynamic effects of time-varying covariates on event times across small time intervals.

Future research should consider extending the longitudinal submodel to include additional time-varying covariates, encompassing both categorical and continuous variables. A piecewise constant hazard framework can be employed to generate survival times, allowing the effect of covariates to vary over time. Additionally, generating survival times under different assumptions about the underlying event-time distributions will help evaluate the robustness of the joint modelling approach.

Figure 5: The forest plots show the means and 95% HDIs of the direct treatment effect α_1 across 10 simulated datasets.



References

[1] Verbeke G, Molenberghs G, Verbeke G. *Linear Mixed Models for Longitudinal Data*. Springer; 1997.

[2] Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in r*. CRC press; 2012.

[3] Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*. Published online 2004:809-834.

[4] Cremers J, Mortensen LH, Ekstrøm CT. A joint model for longitudinal and time-to-event data in social and life course research: Employment status and time to retirement. *Sociological Methods & Research*. 2024;53(2):603-638.

[5] Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of clinical oncology*. 2010;28(16):2796-2801.

[6] Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: A gibbs sampling approach. *Statistics in medicine*. 1996;15(15):1663-1685.

[7] Li N, Elashoff RM, Li G, Saver J. Joint modeling of longitudinal ordinal data and competing risks survival times and analysis of the NINDS rt-PA stroke trial. *Statistics in medicine*. 2010;29(5):546-557.

[8] Ibrahim JG, Chen MH, Sinha D. *Bayesian Survival Analysis*. Springer; 2001. doi:10.1007/978-1-4757-3447-8

[9] Austin PC. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in medicine*. 2012;31(29):3946-3958.

[10] Hendry DJ. Data generation for the cox proportional hazards model with time-dependent covariates: A method for medical researchers. *Statistics in medicine*. 2014;33(3):436-454.

A Appendix

A.1 Show the closed-form function to generate event time

Let us consider Equation 5 again. As event time follows an exponential distribution, we have that

$$\begin{aligned} H(t) &= \int_0^t \exp\Big\{\alpha_2 k + \alpha_1 \mathrm{treatment} + \alpha_2 \beta_2 u\Big\} \lambda du \\ &= \lambda \exp\Big\{\alpha_2 k + \alpha_1 \mathrm{treatment}\Big\} \int_0^t \exp[\alpha_2 \beta_2 u] du \\ &= \frac{\lambda \exp\{\alpha_2 k + \alpha_1 \mathrm{treatment}\}}{\alpha_2 \beta_2} (e^{\alpha_2 \beta_2 t} - 1) \end{aligned}$$

thus,

$$H^{-1}(u) = \ln\left[\frac{\alpha_2 \beta_2 u}{\lambda \exp\{\alpha_2 k + \alpha_1 \mathrm{treatment}\}} + 1\right] \frac{1}{\alpha_2 \beta_2}.$$

The event time can be generated using the closed-form function:

$$T = \ln\left[\frac{\alpha_2 \beta_2 (-\ln u)}{\lambda \exp\{\alpha_2 k + \alpha_1 \mathrm{treatment}\}} + 1\right] \frac{1}{\alpha_2 \beta_2},$$

where $u \sim \mathrm{Uniform}(0, 1)$.