Estimating Heterogeneous Treatment Effects Using Hierarchical Bayesian Models: A Simulation-Based Study

George Tzimas

November 2024

Abstract

Randomized controlled trials typically assess treatment effectiveness based on average treatment effects, which do not account for heterogeneity of responses in subgroups of patients. In this study, we create a hierarchical Bayesian model to estimate individual treatment effect estimates, taking into consideration the heterogeneity of treatment effect as well as baseline risk directly from simulated clinical trial data. We generated synthetic data to simulate a real world randomized controlled trial with important clinical covariates, like age and gender status, as well as bleeding history. Experiencing a major bleeding event was the main outcome of interest and specifically modeled as a binary probabilistic outcome (1, 0) determined by risk factors at baseline and treatment effects.

The results showed strong correlations (r = 0.78) between predicted and true conditional average treatment effects, with a root mean squared error of 0.53, confirming that the model does a good job capturing treatment response variability. Subgroup analysis based on baseline risks revealed that treatment effects were most pronounced among high-risk individuals, with a mean treatment effect of 0.547 in the highest risk quintile (Q5), compared to -0.359 in the lowest quintile (Q1), where treatment may even be harmful. These findings highlight the importance of stratifying patients by baseline risk to optimize treatment decisions and advance personalized medicine. However, the study is limited by the use of simulated data, and further validation on real-world clinical datasets is necessary to confirm the model's applicability.

Keywords: Bayesian hierarchical model, heterogeneous treatment effects, personalized medicine, statistical modeling

1 Introduction

1.1 Background

In order to determine the effectiveness of any given drug, randomized controlled trials (RCTs) tend to focus on the average response that that treatment will produce on a representative sample

of a population. However, that average treatment effect (ATE), whether significant or not, does not always tell us the full story. Treatment responses can often vary substantially among different patient subgroups at varying magnitudes and directions. Understanding this heterogeneity in treatment effects (HTE) is crucial for the implementation of personalized medicine based on a given patient's baseline risk factors, as well as empowering healthcare professionals to make more informed, data-driven decisions.

1.2 Objectives

The goal of this study is to utilize Bayesian methods to estimate the average and subgroup-specific treatment effects with the use of simulated clinical trial data. By incorporating both fixed and random effects, we will measure the effectiveness of these Bayesian methods as an alternative to traditional frequentist methods.

2 Methods

2.1 Data Generation

The synthetic data were emulated so that they resemble a typical randomized control trial for the purposes of testing the effectiveness of a drug. Patient-level clinical covariates such as age, gender, systolic and diastolic blood pressure, etc. were generated to reflect a realistic clinical setting (Table 2). Treatment assignment was randomized with equal probability across subjects to ensure the criteria of randomization and exchangeability are met and to minimize unconfoundedness (Table 1). To quantitatively confirm the absence of confounding, propensity score estimation using a logistic regression formula was performed to verify that randomization effectively balanced the covariates (Figure 1).

Table 1: Baseline Characteristics by Treatment Group

Covariate	level	Control	Treatment	p
n		1016	984	
age (mean (SD))		55.54 (12.03)	55.12 (11.76)	0.431
gender (%)	Male	526 (51.8)	492 (50.0)	0.455
	Female	490 (48.2)	492 (50.0)	
bmi (mean (SD))		27.97 (5.11)	28.19 (5.04)	0.316
sbp (mean (SD))		129.94 (14.16)	130.65 (13.91)	0.263
dbp (mean (SD))		79.89 (10.01)	80.47 (10.07)	0.201
hypertension (%)	No	735 (72.3)	690 (70.1)	0.295
	Yes	281 (27.7)	294 (29.9)	
diabetes (%)	No	775 (76.3)	724 (73.6)	0.179
	Yes	241 (23.7)	260 (26.4)	

(continued)

	level	Control	Treatment	p
smoking (%)	No	774 (76.2)	750 (76.2)	1.000
	Yes	242 (23.8)	234 (23.8)	
anticoagulant (%)	No	693 (68.2)	645 (65.5)	0.224
	Yes	323 (31.8)	339 (34.5)	
prior_bleeding (%)	No	901 (88.7)	896 (91.1)	0.092
	Yes	115 (11.3)	88 (8.9)	
event (%)	0	966 (95.1)	921 (93.6)	0.181
	1	50 (4.9)	63 (6.4)	
true_te (mean (SD))		-0.15 (0.66)	-0.18 (0.65)	0.376

Table 2: Distributions of Covariates in the Synthetic Dataset

Covariate	**Formula**
Age	$\text{Age} \sim \mathcal{N}(\mu = 55, \sigma = 12)$, truncated between 25 and 85
Gender	$\text{Gender} \sim \text{Bernoulli}(p = 0.52)$, (1 = Female, 0 = Male)
SBP and DBP	$\begin{bmatrix}\text{SBP}\\ \text{DBP}\end{bmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix}130\\ 80\end{bmatrix}, \Sigma = \begin{bmatrix}200 & 100\\ 100 & 100\end{bmatrix}\right)$, truncated at realistic values
BMI	$\text{BMI} \sim \mathcal{N}(\mu = 28, \sigma = 5)$, truncated between 16 and 50
Hypertension	$\text{Hypertension} \sim \begin{cases}1, & \text{if SBP} \geq 140 \text{ or DBP} \geq 90\\ 0, & \text{otherwise}\end{cases}$
Diabetes	$\text{Diabetes} \sim \text{Bernoulli}(p = 0.2 + 0.2 \cdot (\text{BMI} > 30))$
Smoking	$\text{Smoking} \sim \text{Bernoulli}(p = 0.25 - 0.1 \cdot (\text{Age} > 65))$
Anticoagulant Use	$\text{Anticoagulant} \sim \text{Bernoulli}(p = 0.3)$
Prior Bleeding	$\text{Prior Bleeding} \sim \text{Bernoulli}(p = 0.1)$

The primary outcome of interest in this simulation is the occurrence of a bleeding event. This binary outcome is modeled as a probabilistic event, determined by both the subject's baseline risk and the potential treatment effect. It was calculated using a logistic regression model that incorporates all the key covariates that are known clinically to contribute to the risk of bleeding:

$$\begin{aligned}\text{Logit}(P_{\text{baseline}}) = \beta_0 &+ \beta_1(\text{Age} - 55) + \beta_2(\text{Anticoagulant Use})\\ &+ \beta_3(\text{Prior Bleeding}) + \beta_4(\text{Diabetes}) + \beta_5(\text{SBP} - 130)\end{aligned} \tag{1}$$

This baseline risk is converted into a probability using the logistic function:

$$P_{\text{baseline}} = \frac{1}{1 + \exp(-\text{Logit}(P_{\text{baseline}}))} \tag{2}$$

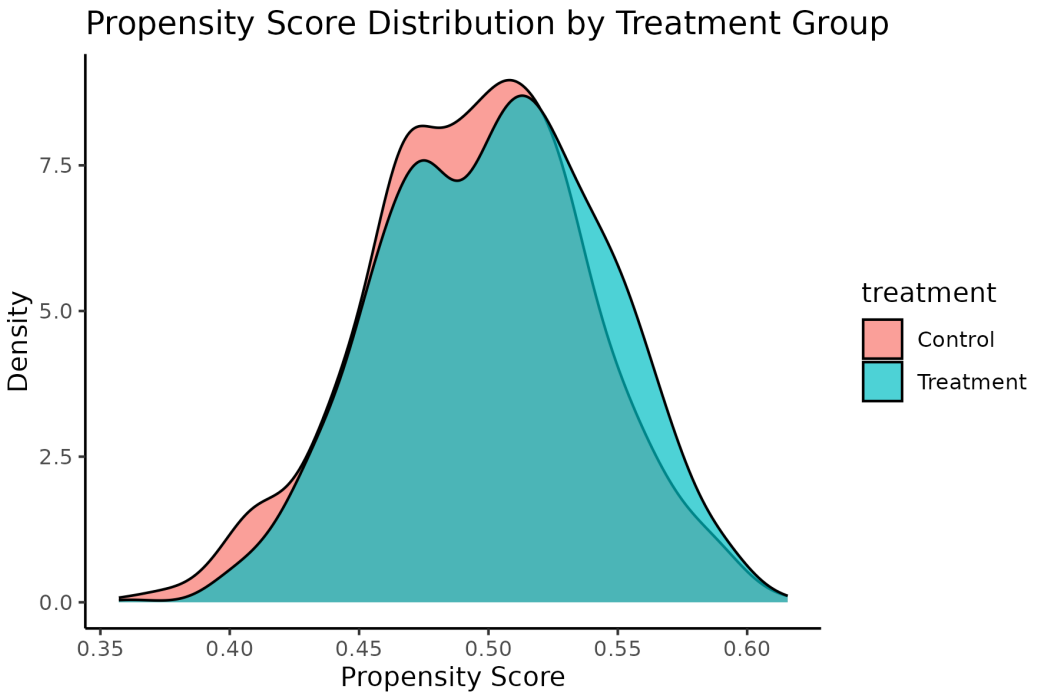

Figure 1: Distribution of propensity scores by treatment arm.

In order to account for heterogeneity in treatment effects, interaction terms were included to reflect how the treatment's impact varies across different subgroups:

$$TE_i = \gamma_0 + \gamma_1(\text{Age} > 65) + \gamma_2(\text{Anticoagulant Use}) + \gamma_3(\text{BMI} > 30), \text{ where} \tag{3}$$

- γ_0: The average increase in bleeding risk caused by treatment.
- γ_1: Older individuals (> 65) experience a greater increase in risk.
- γ_2: Anticoagulant use increases the treatment effect.
- γ_3: Individuals with high BMI (> 30) are more susceptible to bleeding events under treatment.

Finally, the overall probability of a bleeding event for each individual is a combination of the patient's baseline risk and the treatment effect:

$$\text{Logit}(P_{event}) = \text{Logit}(P_{\text{baseline}} + T_i \cdot TE_i) \tag{4}$$

Note that if the individual belongs in the placebo group ($T_i = 0$), their treatment effect is zero, and their probability of experiencing a bleeding event ($\text{Logit}(P_{\text{event}})$) depends only on their baseline risk of bleeding ($\text{Logit}(P_{\text{baseline}})$).

Table 3: First five records of the synthetic dataset

subject_id	age	gender	bmi	sbp	dbp	hypertension	diabetes	smoking	anticoagulant	prior_bleeding	treatment	event
1	48.3	Male	19.9	130	84	No	Yes	Yes	No	No	Control	0
2	52.2	Female	29.9	138	77	No	No	No	No	No	Control	0
3	73.7	Female	37.5	146	98	Yes	No	No	No	No	Control	0
4	55.8	Female	31	144	78	Yes	No	No	No	No	Treatment	0
5	56.6	Male	36.7	97	53	No	No	Yes	No	No	Treatment	0

2.2 Model

In order to estimate the treatment effects while accounting for individual variability and heterogeneity across subgroups, a hierarchical Bayesian model was implemented via JAGS in R. The JAGS model string can be found in Appendix B. The model consists of three levels:

- Population-Level Effects: These are the estimated fixed effects for the baseline covariates and the treatment effect. These aim to capture the average relationships across the entire patient cohort. The population-level treatment effect is defined as:

$$\mu_i = \beta_0 + \beta_1 T_i + \beta_2 X_{i1} + \beta_3 X_{i2} + ... + \beta_p X_{ip} \tag{5}$$

- Subgroup-Level Effects: These are random effects introduced to model subgroup-specific deviations from the population-level treatment effect (e.g. age ¿ 60 or anticoagulant use). This subgroup-level effect is defined as:

$$\tau_k \sim \mathcal{N}(\mu_\tau, \sigma_\tau^2), \tag{6}$$

  where k indexes the subgroup, μ_τ is the overall mean treatment effect, and σ_τ^2 captures the variability in treatment effects across subgroups.

- Individual-Level Effects: At the lowest level, individual-specific deviations are modeled using random effects:

$$\eta_i \sim \mathcal{N}(\tau_{k[i]}, \sigma_\eta^2), \tag{7}$$

  where $k[i]$ represents the subgroup to which individual i belongs, and σ_η^2 captures the within-subgroup variability in treatment effects.

For the likelihood function, the binary outcome of a bleeding event is modeled using a logistic regression framework:

$$\text{Logit}(P_i) = \beta_0 + \beta_1 T_i + \sum_{j=1}^{p} \beta_j X_{ij} + \tau_{k[i]} + \eta_i \tag{8}$$

For each individual i, the likelihood of the observed outcome Y_i is given by:

$$Y_i \sim \text{Bernoulli}(P_i), \tag{9}$$

where P_i is the probability of a bleeding event for individual i.

For our model priors:

- Fixed effects (β_j):

$$\beta_j \sim \mathcal{N}(0, 1) \tag{10}$$

- Random Effects ($\text{re}_k, \text{re\_trt}_k$):

$$\begin{aligned} \text{re}_k &\sim \mathcal{N}(0, \tau_{re}^{-1}) \\ \text{re\_trt} &\sim \mathcal{N}(0, \tau_{re\_trt}^{-1}), \end{aligned} \tag{11}$$

where hyperpriors $\tau_{re}, \tau_{re\_trt} \sim \text{Gamma}(2, 2)$

3 Results

The results from the Gelman-Rubin diagnostic test (Table 5) indicate that all of our parameters converged successfully, with individual parameter and multivariate psrf values very close to 1. Similarly, the Geweke diagnostics indicate that the z-scores for all parameters are within 2 in absolute value (Table 6).

Table 4: Summary statistics for each variable in the model.

	Mean	prob	SD	Naive.SE	Time.series.SE
b_age	0.201	0.550	0.145	0.003	0.003
b_anticoag	0.638	0.654	0.266	0.005	0.006
b_bmi	0.064	0.516	0.100	0.002	0.002
b_diabetes	-0.132	0.467	0.316	0.006	0.006
b_intercept	-1.635	0.163	0.968	0.018	0.091
b_prior	1.450	0.810	0.222	0.004	0.004
b_trt	-0.105	0.474	0.592	0.011	0.028
b_trt_age	0.337	0.583	0.202	0.004	0.004
b_trt_anticoag	0.635	0.654	0.361	0.007	0.008
b_trt_diabetes	0.255	0.563	0.402	0.007	0.008
te_gender[1]	-0.002	0.500	0.353	0.006	0.007
te_gender[2]	-0.393	0.403	0.355	0.006	0.008

Table 5: Gelman-Rubin diagnostic results for model convergence (Multivariate psrf=1.04).

	Point est.	Upper C.I.
b_age	1.002	1.008
b_anticoag	1.002	1.007
b_bmi	1.000	1.000
b_diabetes	1.002	1.008
b_intercept	1.049	1.158
b_prior	1.000	1.003
b_trt	1.001	1.002
b_trt_age	1.002	1.007
b_trt_anticoag	1.001	1.004
b_trt_diabetes	1.000	1.000
te_gender[1]	1.000	1.003
te_gender[2]	1.000	1.001

Table 6: Geweke diagnostic results for each parameter and chain.

	Parameter	Chain_1	Chain_2	Chain_3
1	b_age	-0.806	-0.790	2.970
2	b_anticoag	-1.764	-0.915	-0.812
3	b_bmi	-0.856	-1.215	1.694
4	b_diabetes	1.086	-0.217	-0.109
5	b_intercept	0.416	0.890	-0.353
6	b_prior	-0.700	-0.277	0.770
7	b_trt	-1.806	-0.111	0.385
8	b_trt_age	2.476	0.590	-1.557
9	b_trt_anticoag	3.168	0.557	0.504
10	b_trt_diabetes	-1.427	1.229	-0.006
11	te_gender[1]	-1.853	0.140	0.722
12	te_gender[2]	-1.409	-0.309	-0.628

With the final model, we estimate the conditional average treatment effect (CATE) for each individual patient. These estimated CATES will be compared to the ground-truth CATEs from our simulated data. This allows us to assess the model's ability to capture variability in treatment effects based on all patient covariates.When comparing actual vs. predicted CATEs, we see that there is a strong positive correlation (0.78) as well as a relatively low root mean squared error (0.53), suggesting that our model is able to effectively capture how the covariates influence the baseline risk of experiencing a bleeding event (Table 7).

Table 7: Evaluation metrics for actual vs. predicted CATE estimates.

	Metric	Value
1	Correlation	0.784
2	RMSE	0.529
3	MAE	0.413
4	95% CI Coverage	0.998

Further analysis involved stratifying the population into five equally sized risk groups based on their predicted baseline risk of experiencing the outcome of a bleeding event. When looking at the mean treatment effect (CATE) within each risk subgroup (Table 8), we can see the variation in efficacy of the treatment, where treatment efficacy increases with baseline risk. Low-risk individuals (Q1 and Q2) are unlikely to benefit from treatment and may even be harmed, whereas high-risk individuals (Q4 and Q5) derive significant benefit.

Table 8: Comparison of treatment effect stratified based on baseline risk of experiencing a bleeding event.

	risk_group	n	mean_risk	mean_te	te_sd	te_lower	te_upper
1	Q1	400	0.169	-0.359	0.249	-1.663	0.985
2	Q2	400	0.197	-0.052	0.209	-1.261	1.189
3	Q3	400	0.226	0.203	0.224	-1.052	1.478
4	Q4	400	0.281	0.505	0.229	-0.813	1.832
5	Q5	400	0.421	0.547	0.502	-0.762	1.845

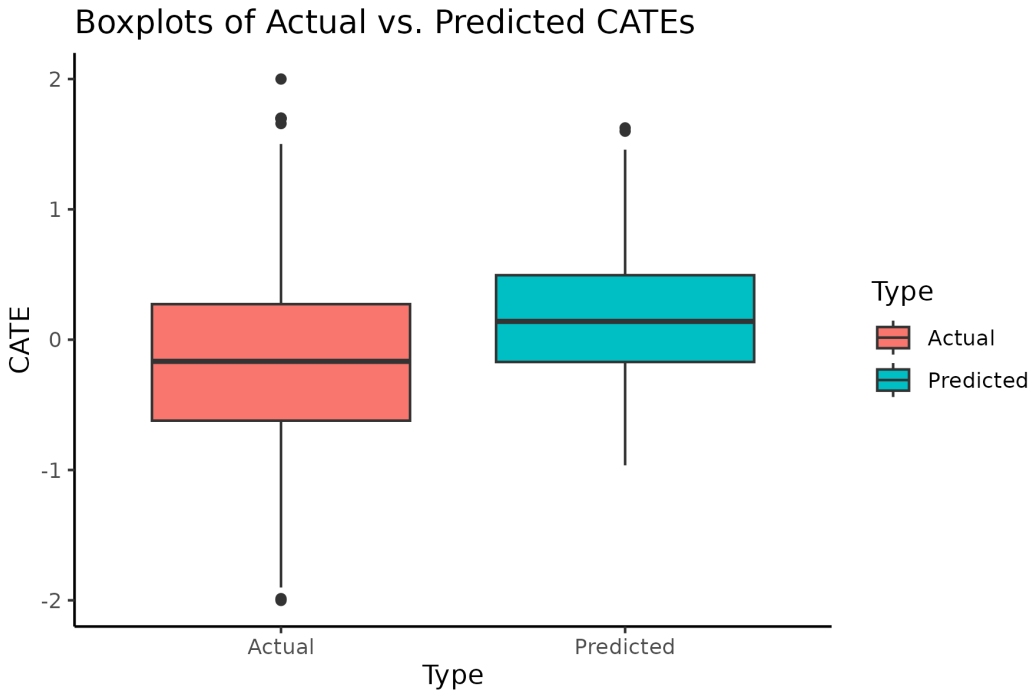

Figure 2: Distribution of actual and predicted CATE estimates.

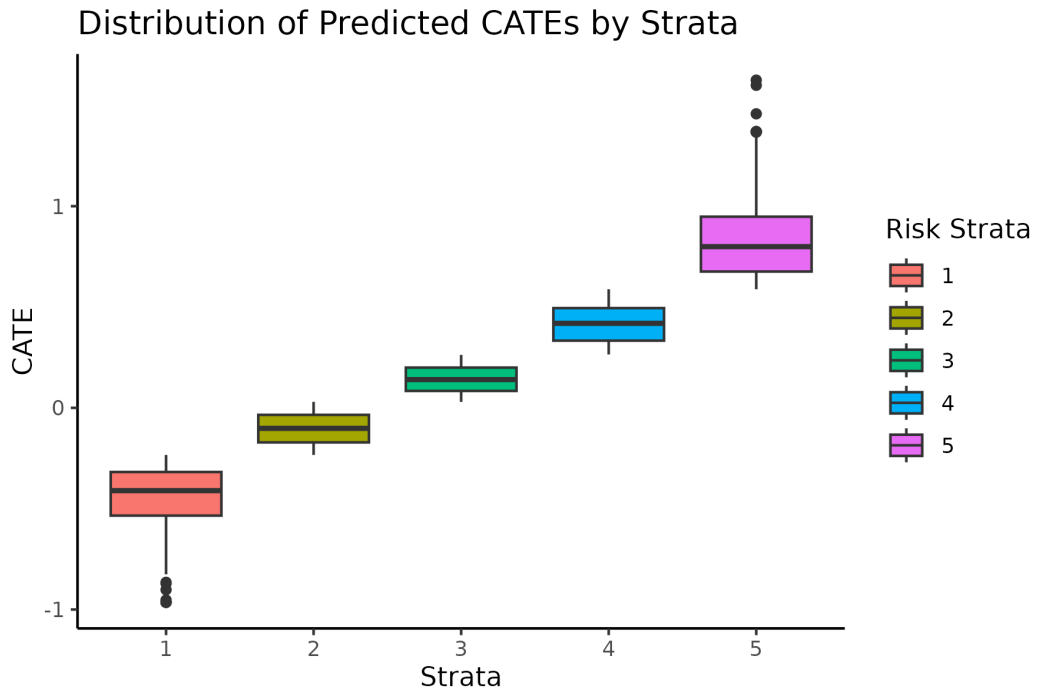

Figure 3: Distribution of CATE estimates by risk strata.

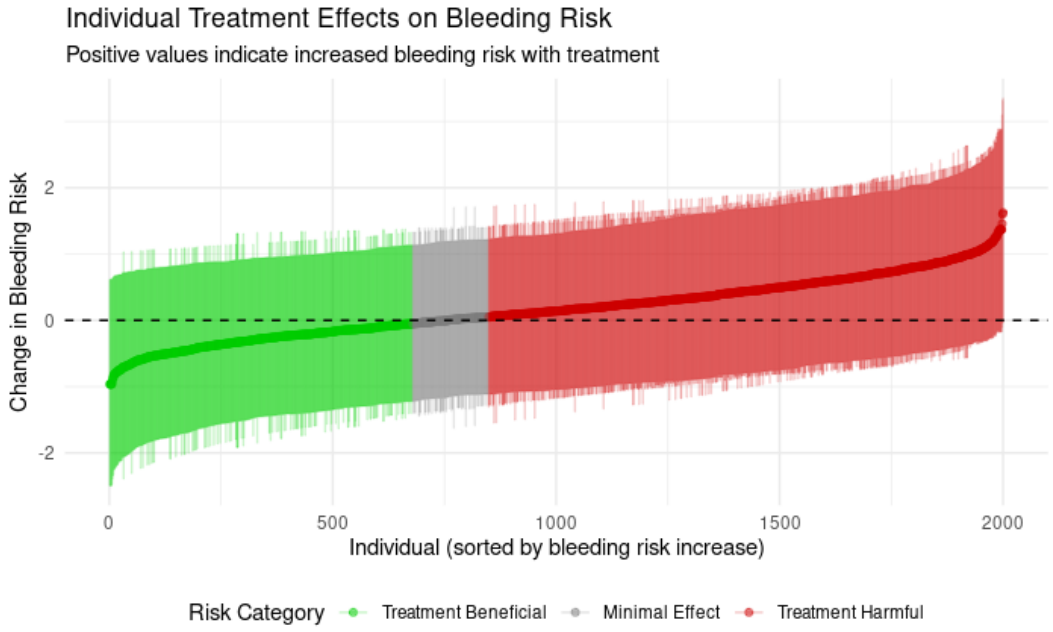

Figure 4: Caption

4 Discussion

4.1 Key Results

This study demonstrated the importance of taking treatment effect heterogeneity into account when evaluating the results of a clinical intervention. Using a hierarchical Bayesian model, we estimated the individual CATEs and compared them to the true effects derived from our simulated data. Our results indicate that treatment efficacy is not always uniform within a patient cohort, but can vary significantly based on baseline covariate measurements.

Further subgroup analysis highlighted the importance of identifying patient subgroups that are most likely to benefit from treatment interventions. Doing so can enable practitioners to provide personalized medicine decisions for their patients and optimize patient outcomes, while at the same time minimizing harm. In addition, the substantial heterogeneity observed within higher-risk groups highlights the need for further investigation into individual-level predictors of treatment response, which could refine patient selection criteria even further.

4.2 Limitations

This study relied solely on simulated data to evaluate model performance. While this approach allows for controlled evaluation and comparison to a ground truth, it may not fully capture the complexity and noise present in real-world clinical datasets. Doing further validation on clinical trial data is necessary to confirm the model's applicability and effectiveness. In addition, the number of covariates used in the simulated dataset is relatively limited when compared to real-world clinical data, where you would have access to a more complex set of covariates, including time-to-event measurements.

A Data Generation Code

```r
generate_synthetic_rct <- function(n_subjects = 1000, max_followup = 365
    * 3) {
  set.seed(123)

  # basic demographic covariates
  age <- rnorm(n_subjects, mean = 55, sd = 12)
  age <- pmax(pmin(age, 85), 25)  # Truncate age between 25 and 85

  gender <- rbinom(n_subjects, 1, 0.5)

  # blood pressure variables
  bp_matrix <- MASS::mvrnorm(n_subjects,
                             mu = c(130, 80),
                             Sigma = matrix(c(200, 100,
                                              100, 100), 2, 2))
```

```r

  sbp <- bp_matrix[,1]
  dbp <- bp_matrix[,2]

  # BMI
  bmi <- rnorm(n_subjects, mean = 28, sd = 5)
  bmi <- pmax(pmin(bmi, 50), 16)  # Truncate BMI between 16 and 50

  # binary comorbidities with some correlation
  hypertension <- ifelse(sbp >= 140 | dbp >= 90, 1, 0)
  diabetes <- rbinom(n_subjects, 1, 0.2 + 0.2 * (bmi > 30))
  smoking <- rbinom(n_subjects, 1, 0.25 - 0.1 * (age > 65))

  # Additional risk factors
  anticoagulant <- rbinom(n_subjects, 1, 0.3)
  prior_bleeding <- rbinom(n_subjects, 1, 0.1)

  # Treatment assignment
  treatment <- rbinom(n_subjects, 1, 0.5)

  # Generate baseline risk
  baseline_risk <- plogis(-3.5 +
                          0.02 * (age - 55) +
                          0.8 * anticoagulant +
                          1.2 * prior_bleeding +
                          0.3 * diabetes +
                          0.01 * (sbp - 130))

  # Generate treatment effects with possibility of harm
  treatment_effect <- -0.2 +                    # Negative base effect
                      0.5 * (age - 55)/10 +     # Age effect centered
                        at 55
                      0.4 * anticoagulant +     # Positive effect for
                        anticoagulated
                      -0.3 * diabetes +         # Negative effect for
                        diabetics
                      -0.2 * (bmi > 30) +       # Negative effect for
                        obese
                      0.3 * prior_bleeding      # Positive effect for
                        prior bleeders

  # Treatment assignment
  treatment <- rbinom(n_subjects, 1, 0.5)

  # Generate baseline risk
  baseline_risk <- plogis(-3.5 +
                          0.02 * (age - 55) +
                          0.8 * anticoagulant +
                          1.2 * prior_bleeding +
                          0.3 * diabetes +
```

```
                        0.01 * (sbp - 130))

  # Generate outcome with treatment interaction
  event_prob <- plogis(logit(baseline_risk) + treatment * treatment_
     effect)
  event <- rbinom(n_subjects, 1, event_prob)

  # Create data frame
  data <- data.frame(
    subject_id = 1:n_subjects,
    age = round(age, 1),
    gender = factor(gender, labels = c("Male", "Female")),
    bmi = round(bmi, 1),
    sbp = round(sbp),
    dbp = round(dbp),
    hypertension = factor(hypertension, labels = c("No", "Yes")),
    diabetes = factor(diabetes, labels = c("No", "Yes")),
    smoking = factor(smoking, labels = c("No", "Yes")),
    anticoagulant = factor(anticoagulant, labels = c("No", "Yes")),
    prior_bleeding = factor(prior_bleeding, labels = c("No", "Yes")),
    treatment = factor(treatment, labels = c("Control", "Treatment")),
    event = event,
    true_te = treatment_effect  # Store true treatment effect for
       validation
  )

  return(data)
}
```

Listing 1: Code used to simulate the RCT dataset.

B JAGS Model String

```
model {
  # Likelihood
  for (i in 1:N) {
    y[i] ~ dbern(p[i])
    logit(p[i]) <- mu[i]

    mu[i] <- b_intercept +
             b_trt * treatment[i] +
             b_age * age[i] +
             b_anticoag * anticoag[i] +
             b_diabetes * diabetes[i] +
             b_prior * prior_bleeding[i] +
             b_bmi * bmi[i] +
             # Treatment interactions
             b_trt_age * treatment[i] * age[i] +
```

```
            b_trt_anticoag * treatment[i] * anticoag[i] +
            b_trt_diabetes * treatment[i] * diabetes[i] +
            # Random effects
            re[gender[i] + 1] +
            re_trt[gender[i] + 1] * treatment[i]
  }

  # Priors for fixed effects
  b_intercept ~ dnorm(0, 1)
  b_trt ~ dnorm(0, 1)
  b_age ~ dnorm(0, 1)
  b_anticoag ~ dnorm(0, 1)
  b_diabetes ~ dnorm(0, 1)
  b_prior ~ dnorm(0, 1)
  b_bmi ~ dnorm(0, 1)

  # Priors for interaction effects
  b_trt_age ~ dnorm(0, 1)
  b_trt_anticoag ~ dnorm(0, 1)
  b_trt_diabetes ~ dnorm(0, 1)

  # Random effects for gender
  for (k in 1:n_gender) {
    re[k] ~ dnorm(0, tau_re)
    re_trt[k] ~ dnorm(0, tau_re_trt)
  }

  # Hyperpriors for random effects
  tau_re ~ dgamma(2, 2)
  tau_re_trt ~ dgamma(2, 2)

  # Calculate gender-specific treatment effects
  for (k in 1:n_gender) {
    te_gender[k] <- b_trt + re_trt[k]
  }
}
```

Listing 2: JAGS model statement.

C Trace Plots and Density Plots

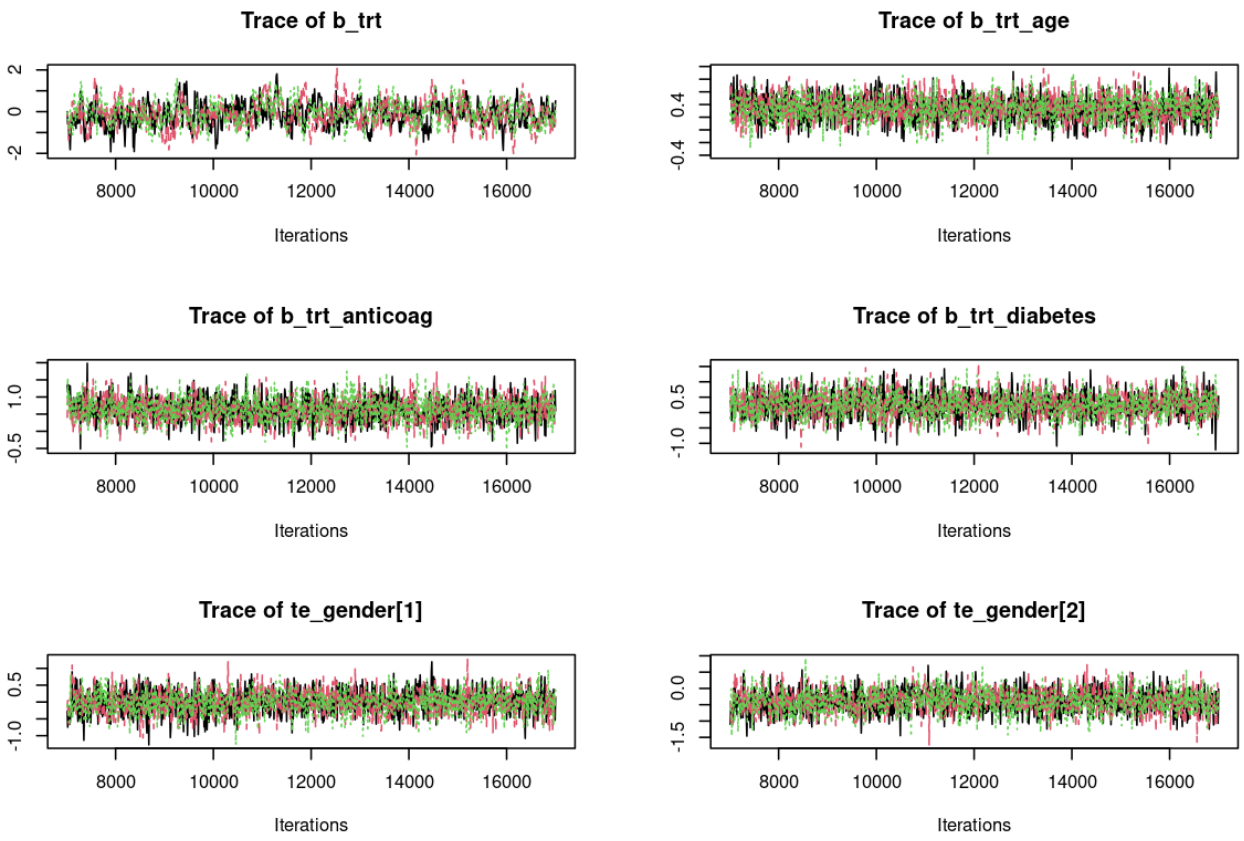

Figure 5: Traceplots of MCMC samples for key parameters.

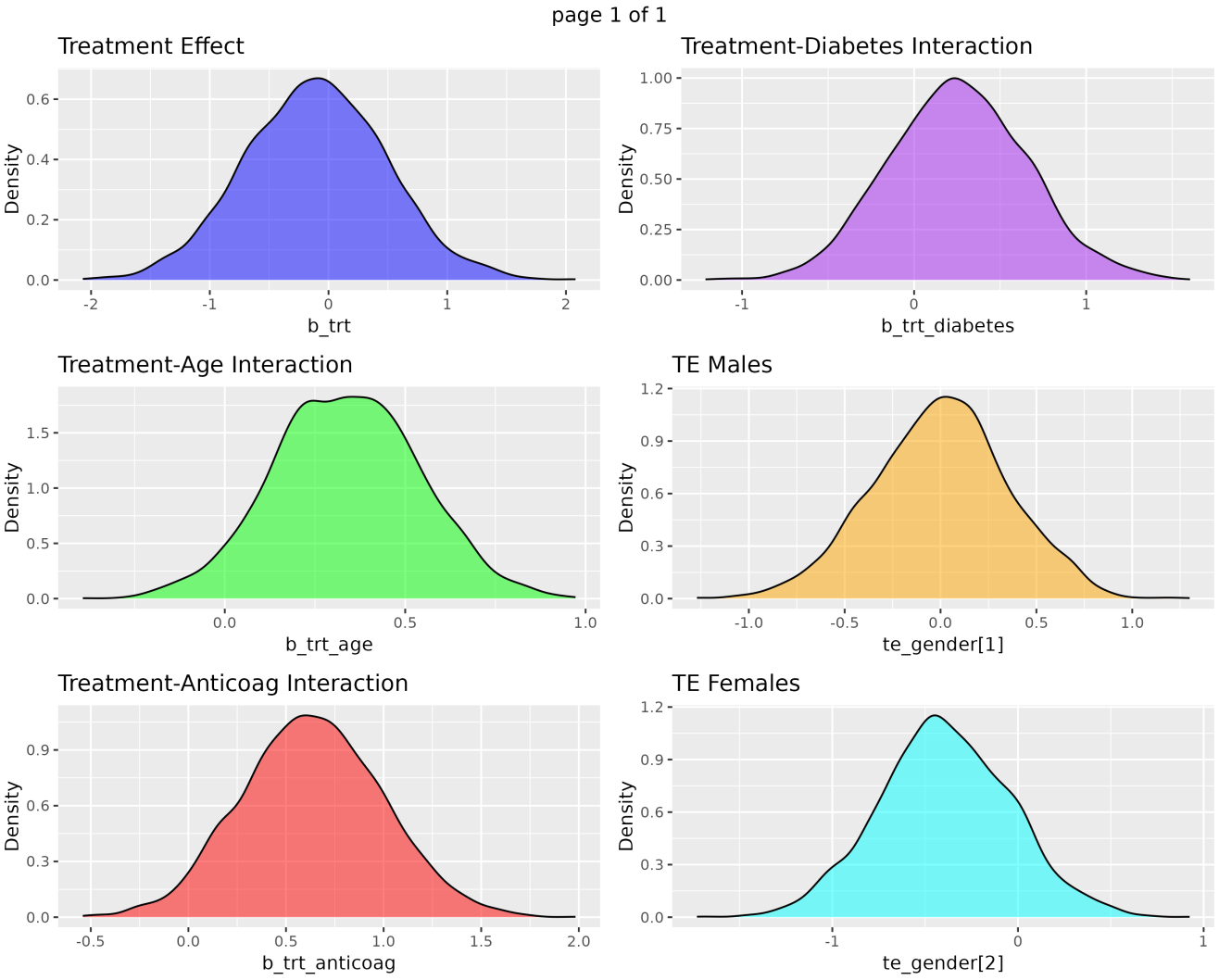

Figure 6: Posterior density estimates for key parameters.