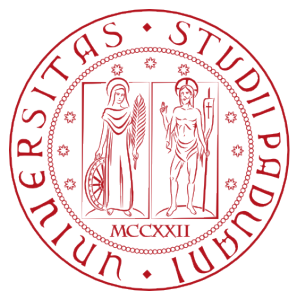
Introduction to Bayesian Computation and Application to Regression Models and Survival Analysis

IBIG 2018

Daniele Bottigliengo[1]

Padova, Italy, November 22, 2018



UNIVERSITÀ DEGLI STUDI DI PADOVA

[1] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padua, Italy

Survival Analysis Case Study

Survival Ovarian Cancer

- Randomized trial comparing treatment of patients with advanced ovarian carcinoma (stages *IIIB* and *IV*) (Edmonson et al. 1979)
- Two groups of patients:
  - Cyclophosphamide alone ($1\ g/m^2$)
  - Cyclophosphamide ($500\ \mu g/m^2$) plus Adriamycin ($40\ \mu g/m2$)
- Intravenous (IV) injection every 3 weeks

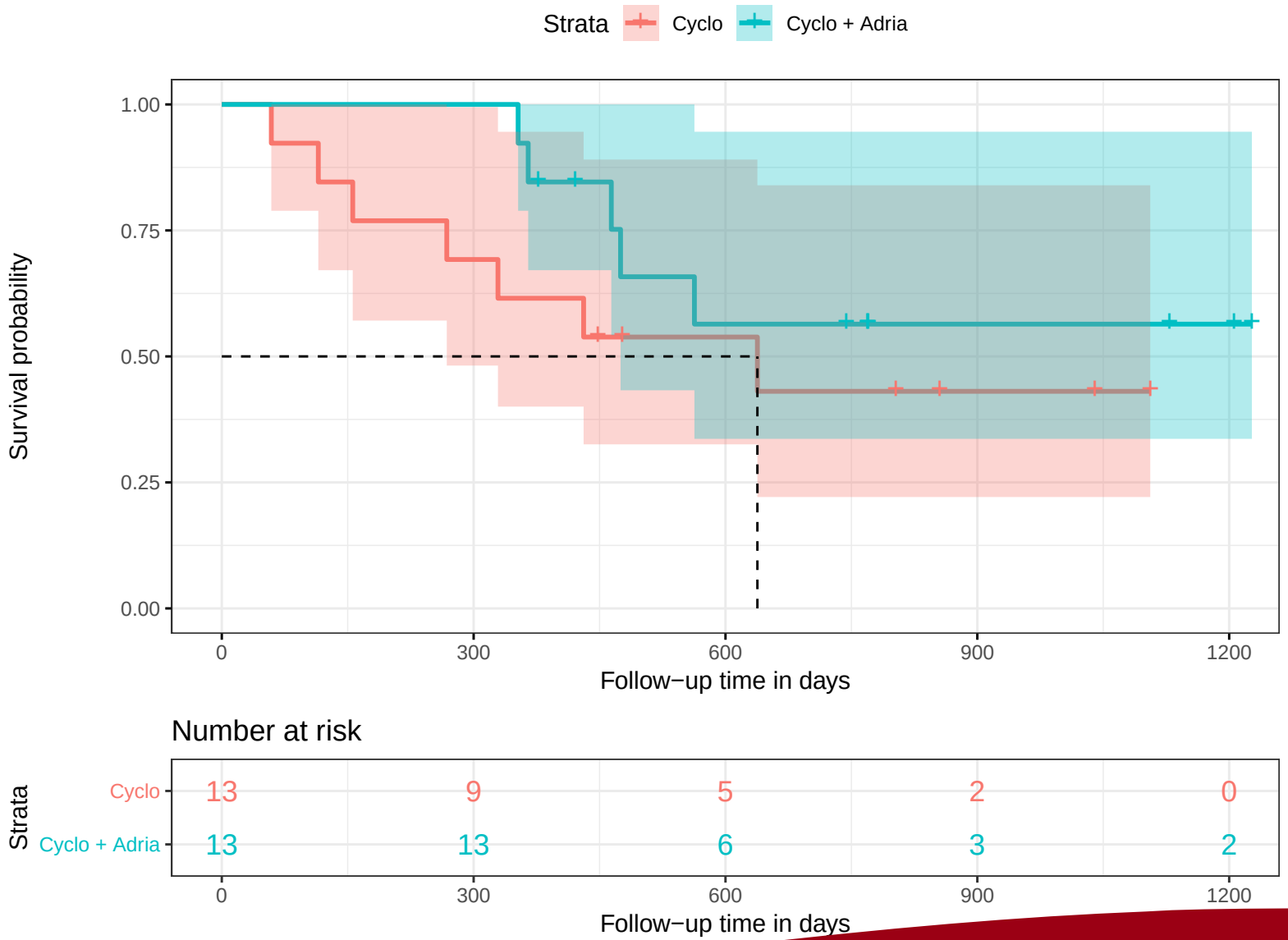
The dataset (1)

- 26 women enrolled
- The following information were retrieved:
  - Age
  - Presence of residual disease
  - ECOG performance
- Median follow-up time in the Cyclophosphamide group: 448 days
- Median follow-up time in the Cyclophosphamide plus Adriamycin 563 days
- 12 patients died during the study and 14 were right-censored

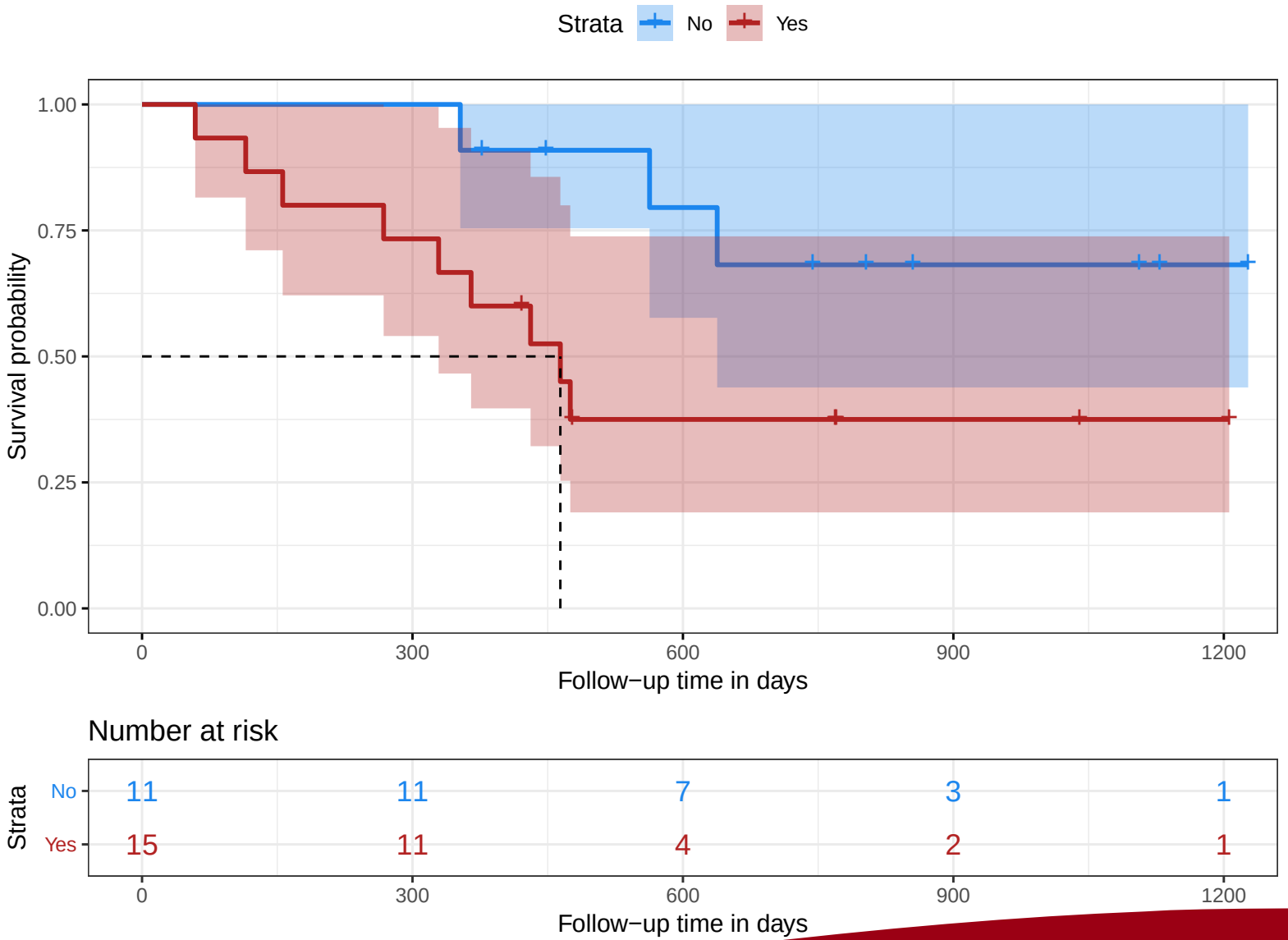
The dataset (2)

follow_up_days	status	age	residual_disease	treatment	ecog_performance
59	dead	72.3315	yes	Cyclo	1
115	dead	74.4932	yes	Cyclo	1
156	dead	66.4658	yes	Cyclo	2
421	alive	53.3644	yes	Cyclo + Adria	1
431	dead	50.3397	yes	Cyclo	1
448	alive	56.4301	no	Cyclo	2
464	dead	56.9370	yes	Cyclo + Adria	2
475	dead	59.8548	yes	Cyclo + Adria	2
477	alive	64.1753	yes	Cyclo	1
563	dead	55.1781	no	Cyclo + Adria	2

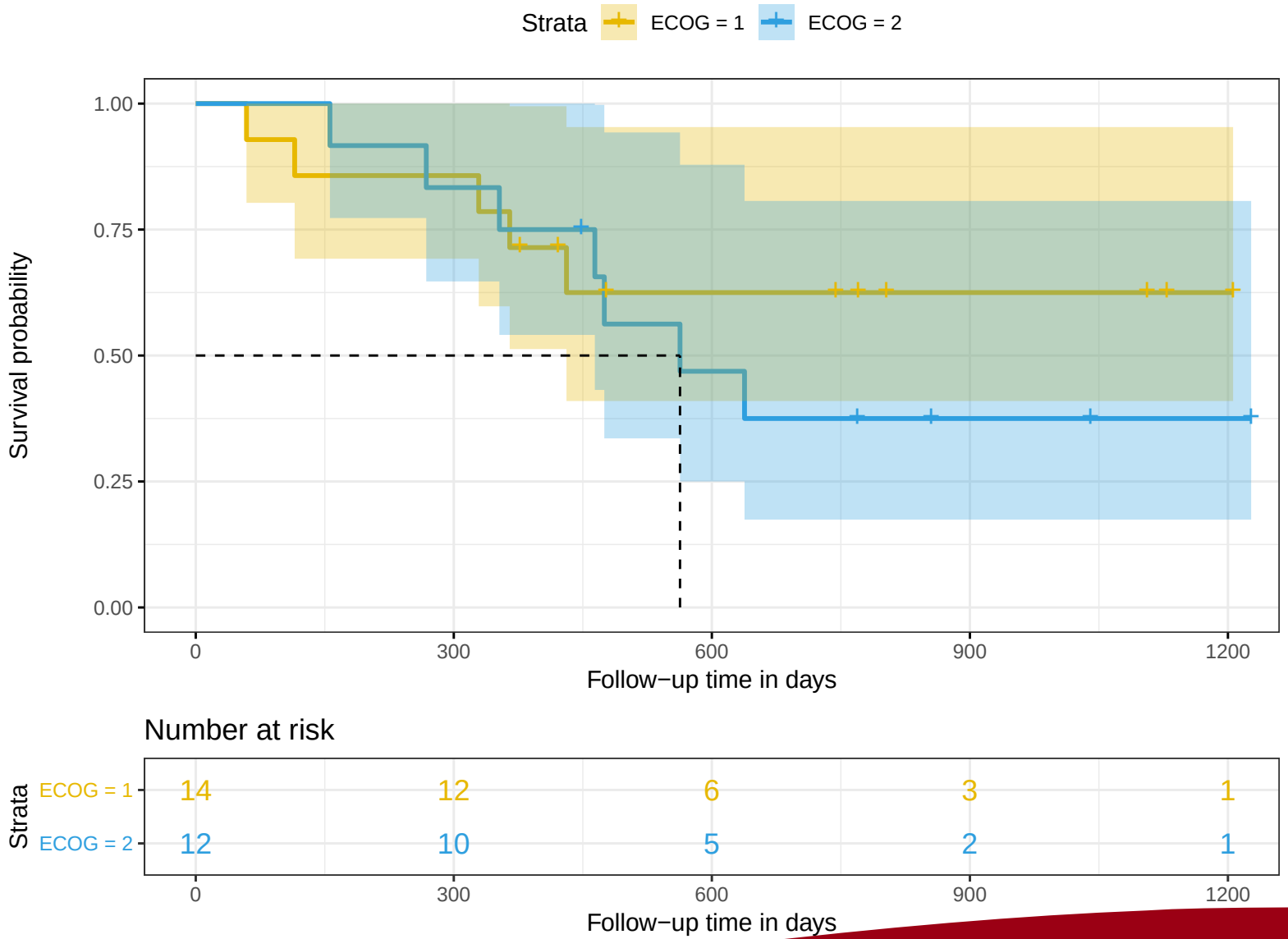
Exploratory data analysis (1)



Exploratory data analysis (2)



Exploratory data analysis (3)



Survival Model

Weibull parametric proportional hazard model:

$$f\left(t|\alpha,\sigma\right) = \frac{\alpha}{\sigma}\left(\frac{t}{\sigma}\right)^{\alpha-1} e^{-\left(\frac{t}{\sigma}\right)^{\alpha}}$$

where:

- α is the shape parameter
- σ is the scale parameter defined as $\sigma = e^{-\left(\frac{\eta}{\alpha}\right)}$.
- η is the linear predictor and it can be expressed as function of some covariates

Fake data simulations

- Starting point of model fitting
- Check if the model makes sense

How to proceed:

1. Simulate fake data from the prior predictive distributions
2. Fit the model to the simulated data
3. Are true parameters values included in the posterior distributions?

The model: data block


```
"
data {

  int<lower = 0> n_obs;              // Number of deaths
  int<lower = 0> n_cens;             // Number of censored
  vector[n_obs] y_obs;               // Death vector
  vector[n_cens] y_cens;             // Censored vector
  int<lower = 0> k;                  // Number of covariates
  matrix[n_obs, k] x_obs;            // Design matrix for deaths
  matrix[n_cens, k] x_cens;          // Design matrix for censoring

}

transformed data {

  real<lower = 0> tau_beta_0;        // Sd of intercept
  real<lower = 0> tau_alpha;         // Sd alpha

  tau_beta_0 = 10;
  tau_alpha = 10;

}
"
```

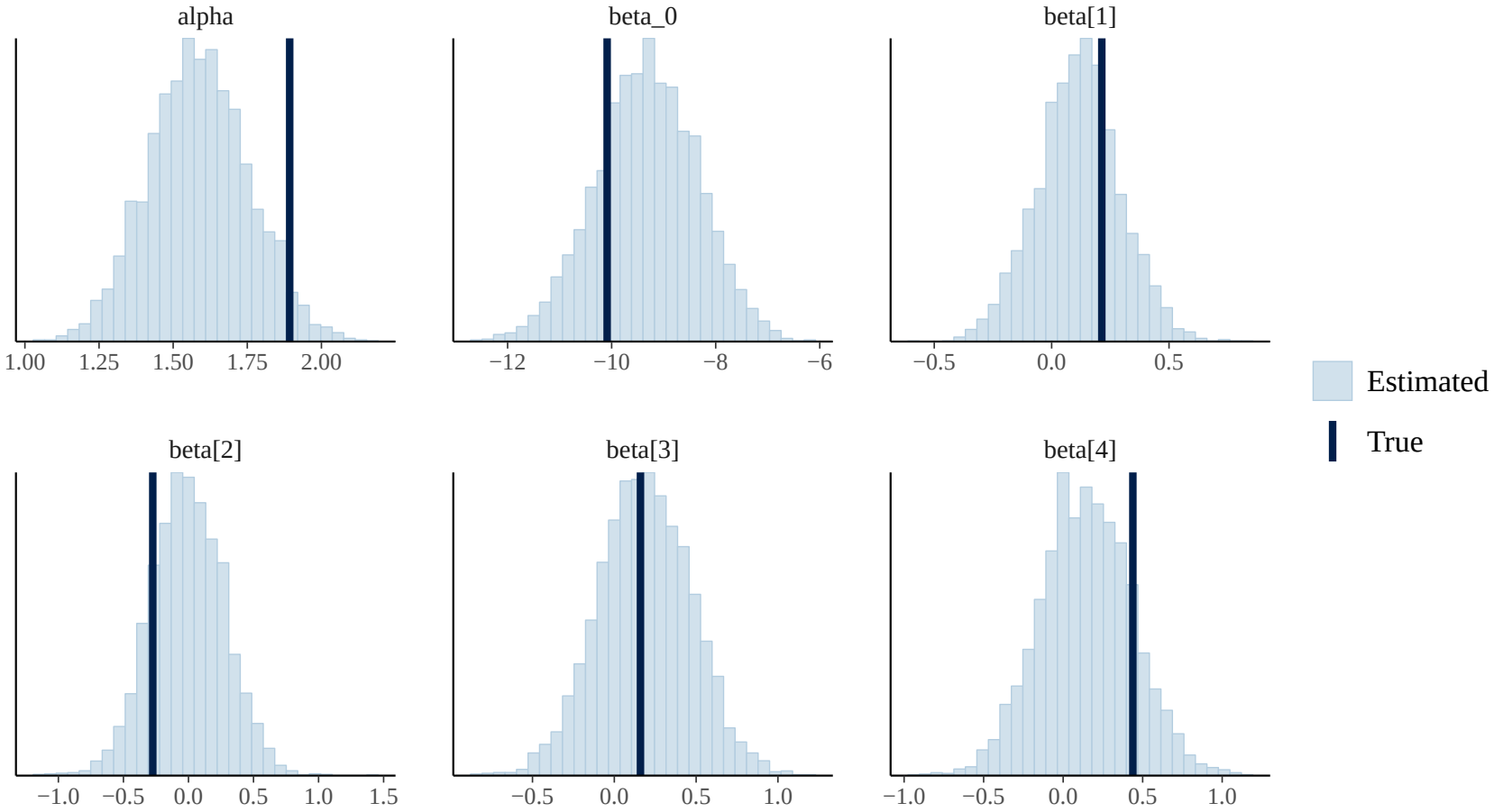
The model: parameters block


```
"
parameters {

  real<lower = 0> alpha;           // Alpha parameter on the log scale
  real beta_0;                     // Intercept
  vector[k] beta;                  // Coefficients of covariates

}
"
```

The model: model block


```
"
model {

  // Linear predictors
  vector[n_obs] eta_obs = beta_0 + x_obs * beta;
  vector[n_cens] eta_cens = beta_0 + x_cens * beta;

  // Define the priors
  target += normal_lpdf(alpha | 0, tau_alpha) +
            normal_lpdf(beta_0 | 0, tau_beta_0) +
            normal_lpdf(beta | 0, 1);

  // Define the likelihood
  target += weibull_lpdf(y_obs | alpha, exp(-eta_obs/alpha)) +
            weibull_lccdf(y_cens | alpha, exp(-eta_cens/alpha));

}
"
```

Recover the parameters values



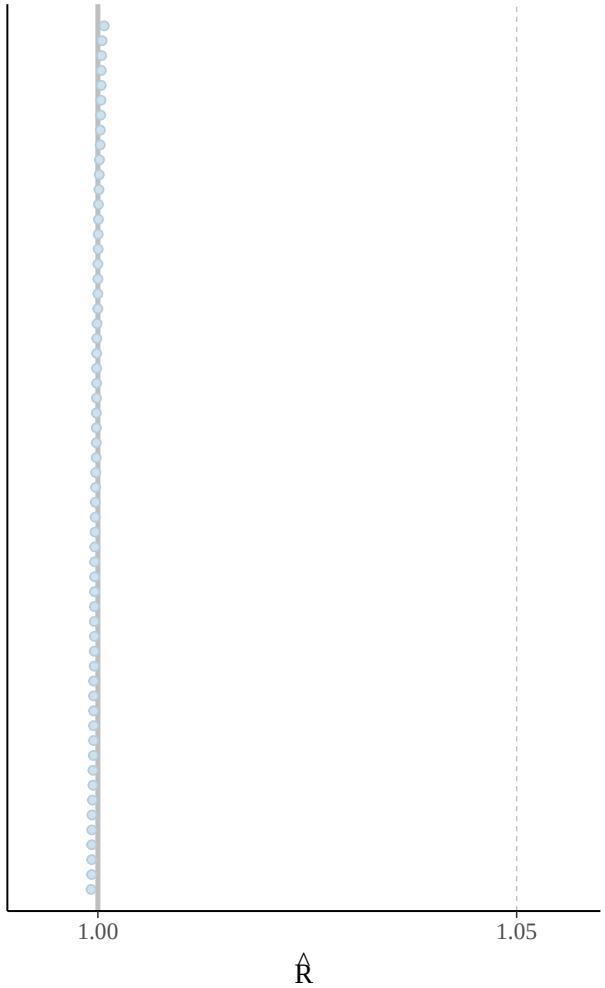
Fit the model to the real data

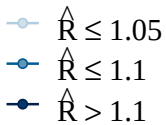
- If the fitted model is able to recover the true parameters values it is possible to proceed by fitting the model to real data
- Prior Predictive checks can be very useful to question the the correctness of the model
- Before fitting the model to the real data, centering and scale the covariates is useful to ease the sampling process

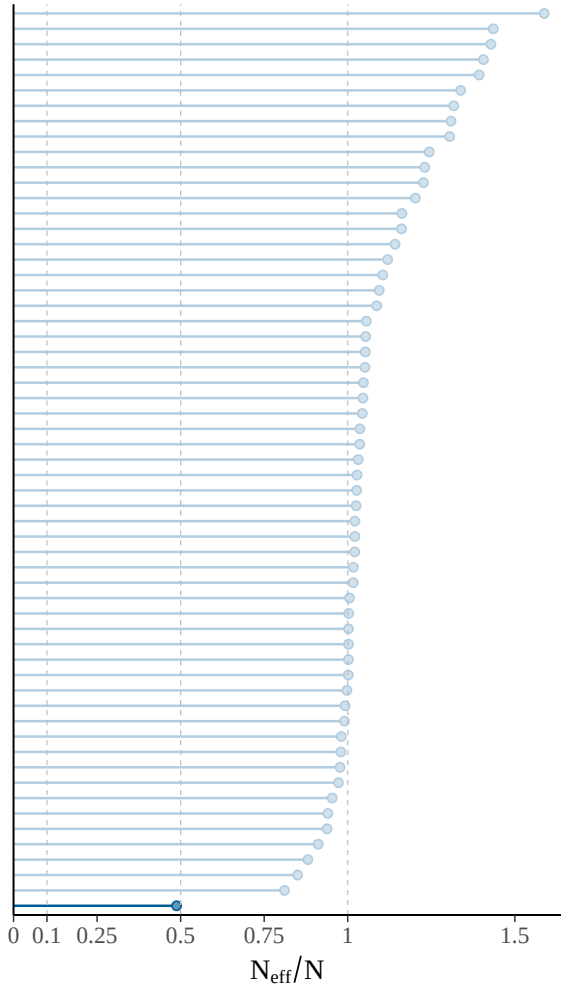
Two steps are important to evaluate the robustness of the analysis:

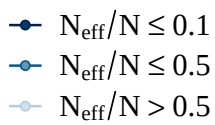
- MCMC diagnostics
- Posterior Predictive Checks

MCMC diagnostics: R_{hat} and ESS

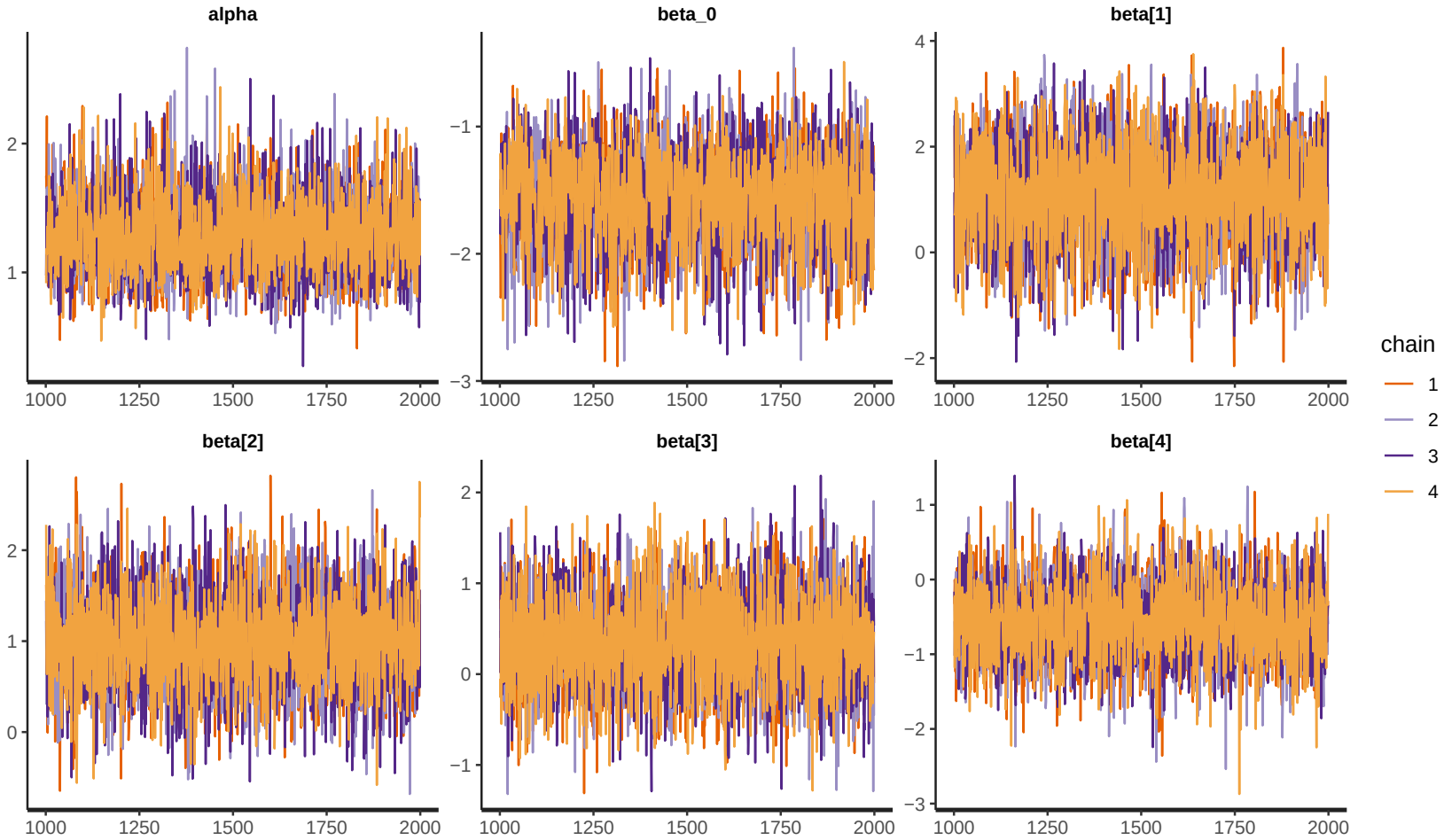




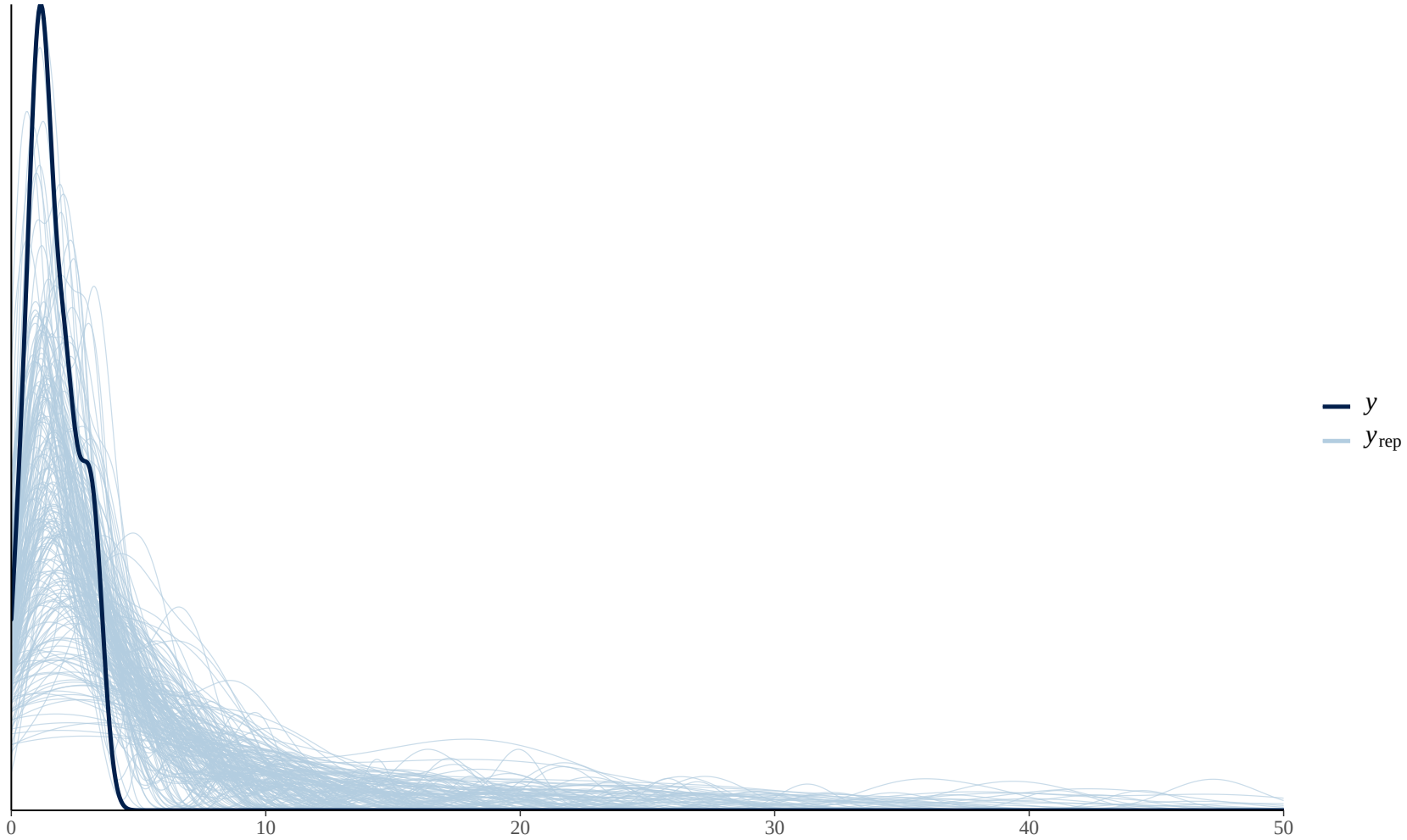




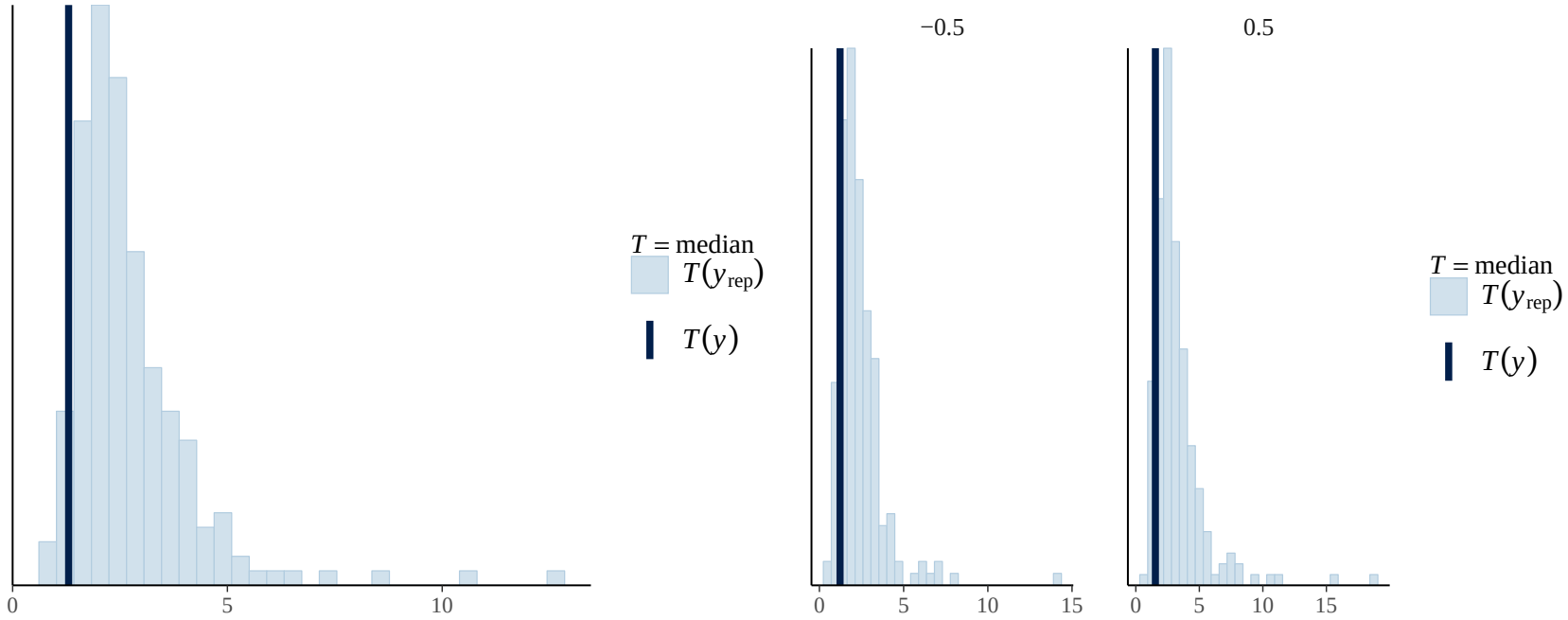
MCMC diagnostics: traceplot



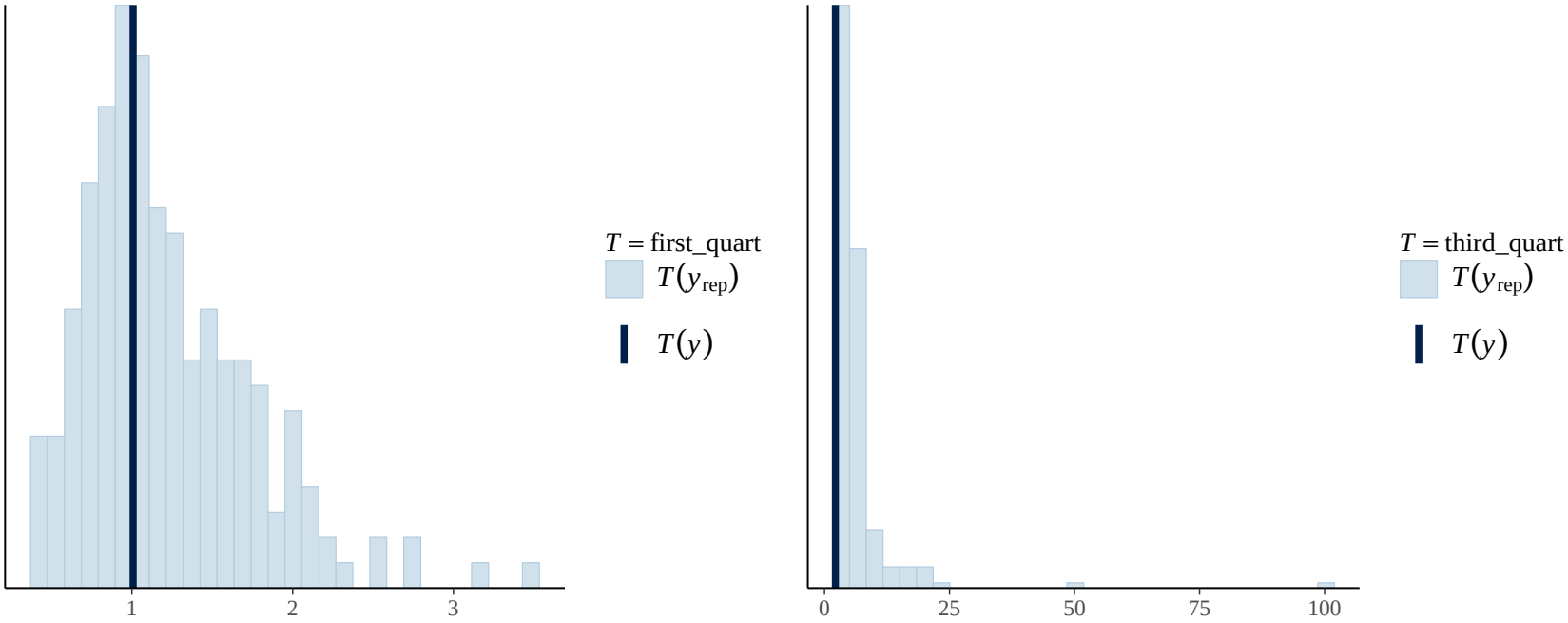
Posterior Predictive Checks (1)



Posterior Predictive Checks (2)



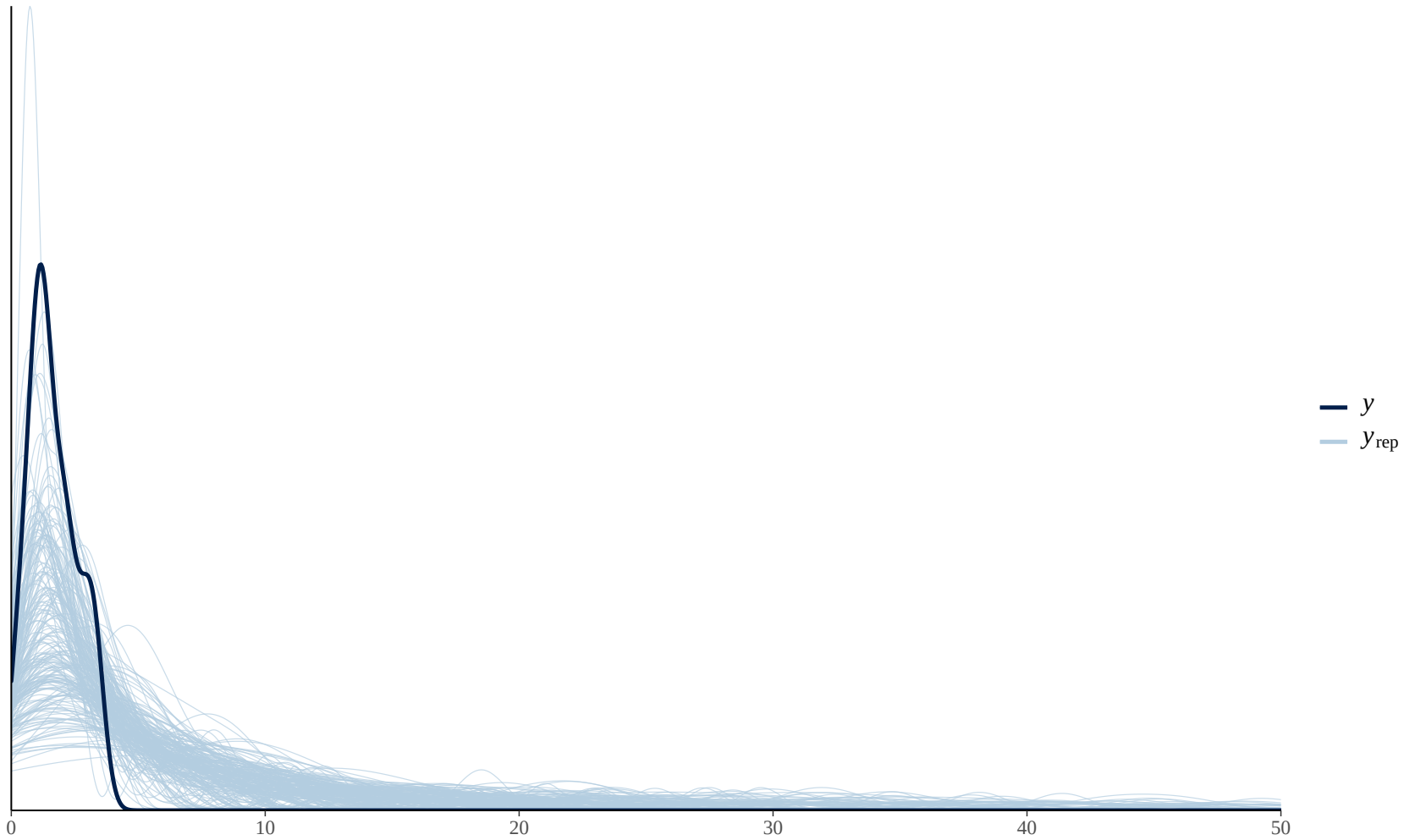
Posterior Predictive Checks (3)



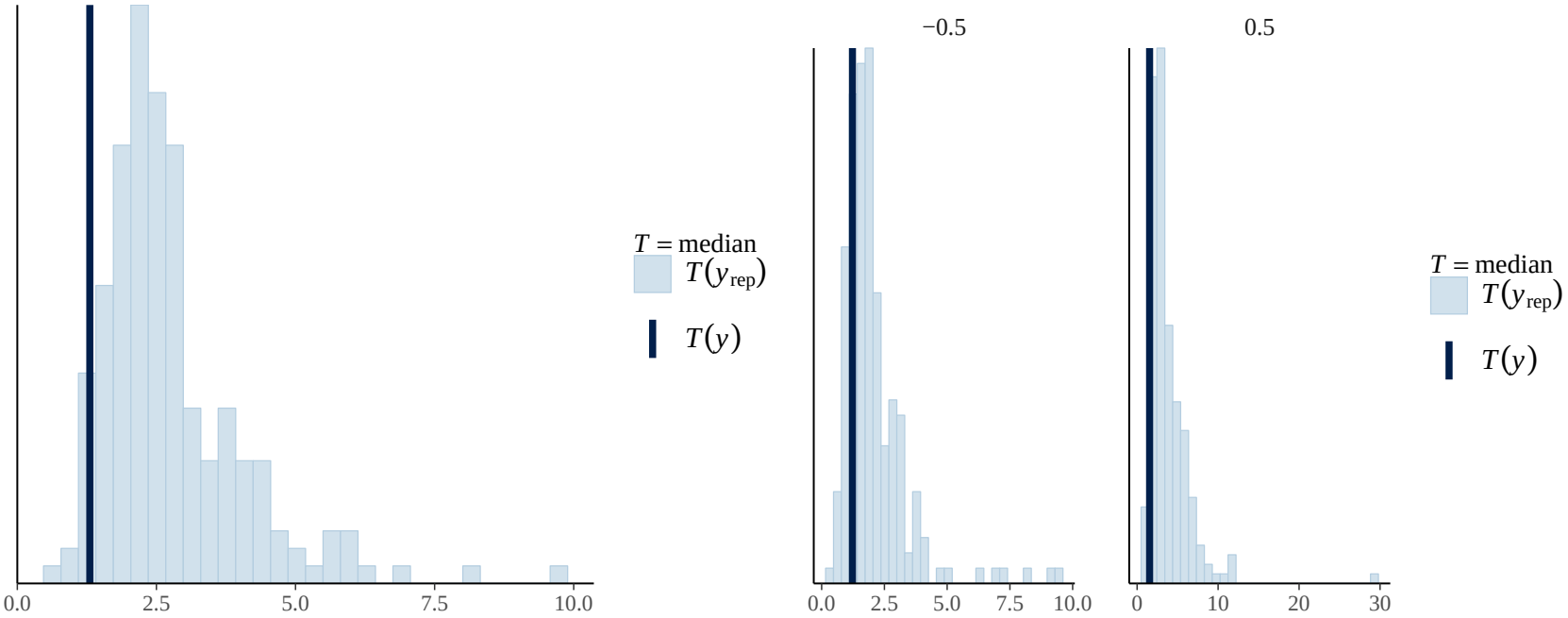
Revise the model

- The model predicts greater follow-up times than those observed in the ovarian cancer data
- Weibull distribution may not be the best one to model time-to-deaths of subjects with ovarian cancer
- Different family distributions can be considered, e.g. log-normal, gamma, ...

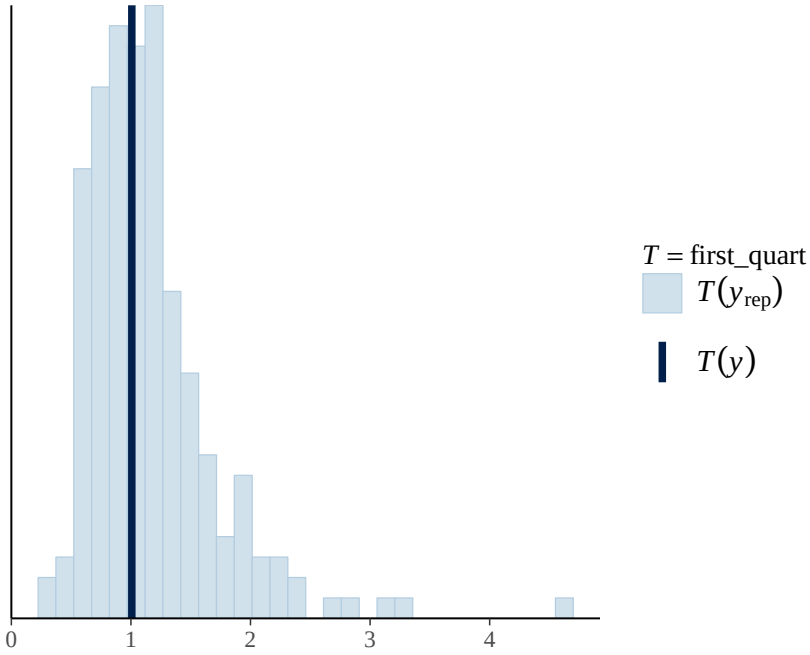
Log-normal (1)

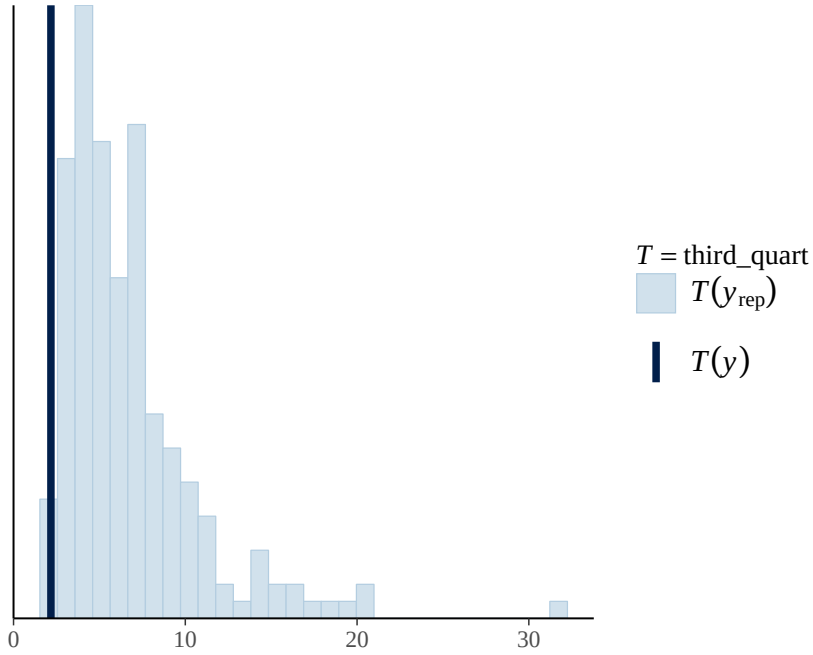


Log-normal (2)

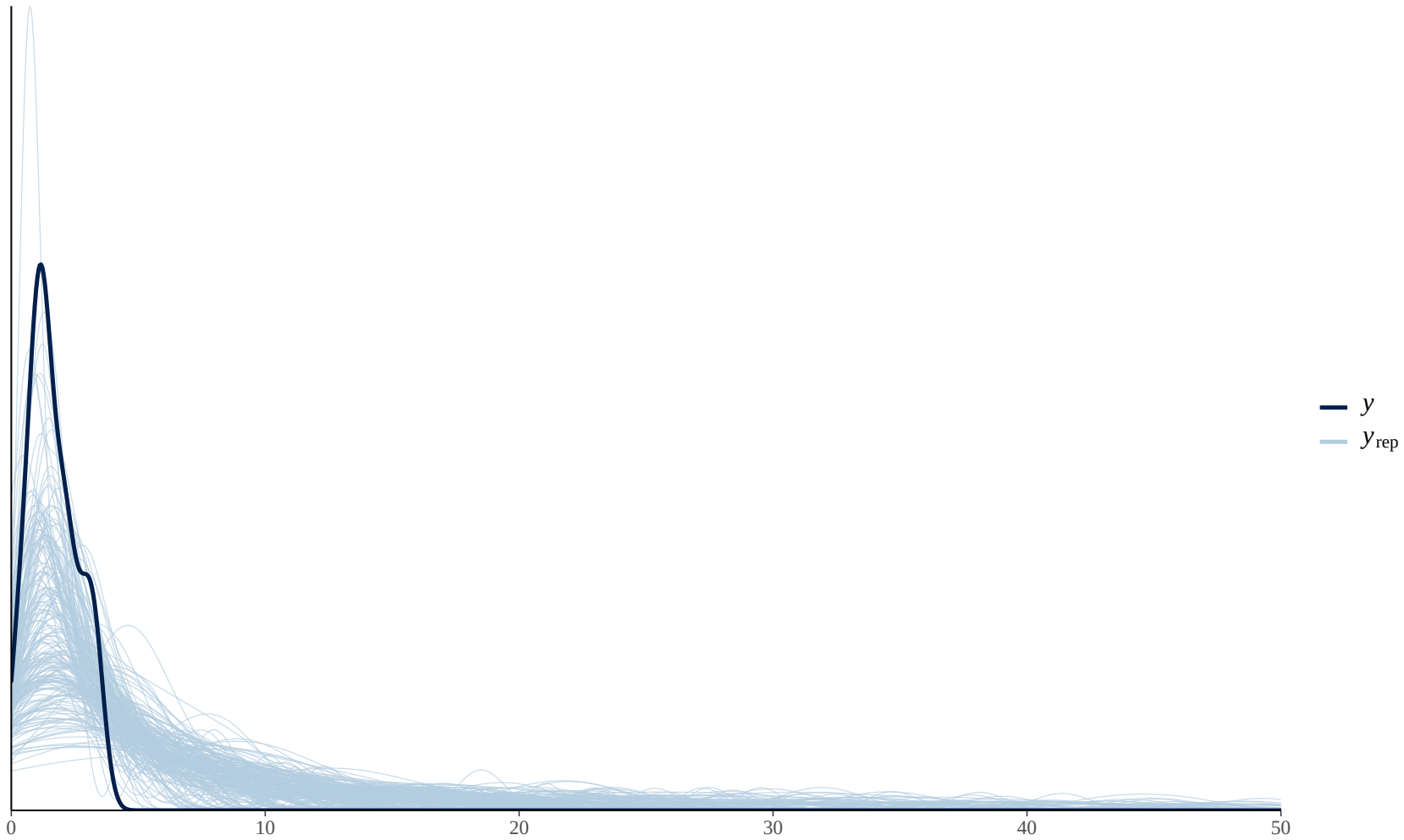

0.0
2.5
5.0
7.5
10.0
T = median
$T(y_{rep})$
$T(y)$
−0.5
0.5
0.0
2.5
5.0
7.5
10.0
0
10
20
30
T = median
$T(y_{rep})$
$T(y)$

Log-normal (3)

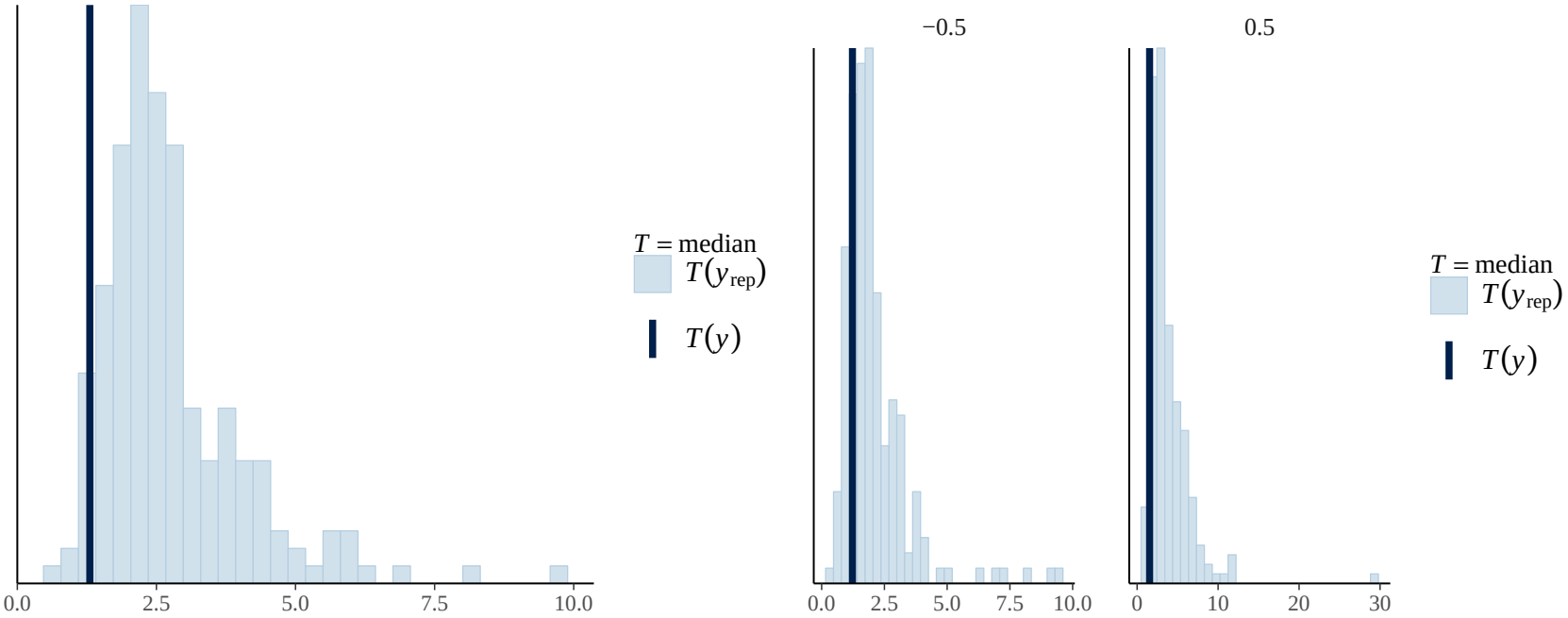

0
1
2
3
4
T = first_quart
$T(y_{\text{rep}})$
$T(y)$


0
10
20
30
T = third_quart
$T(y_{\text{rep}})$
$T(y)$

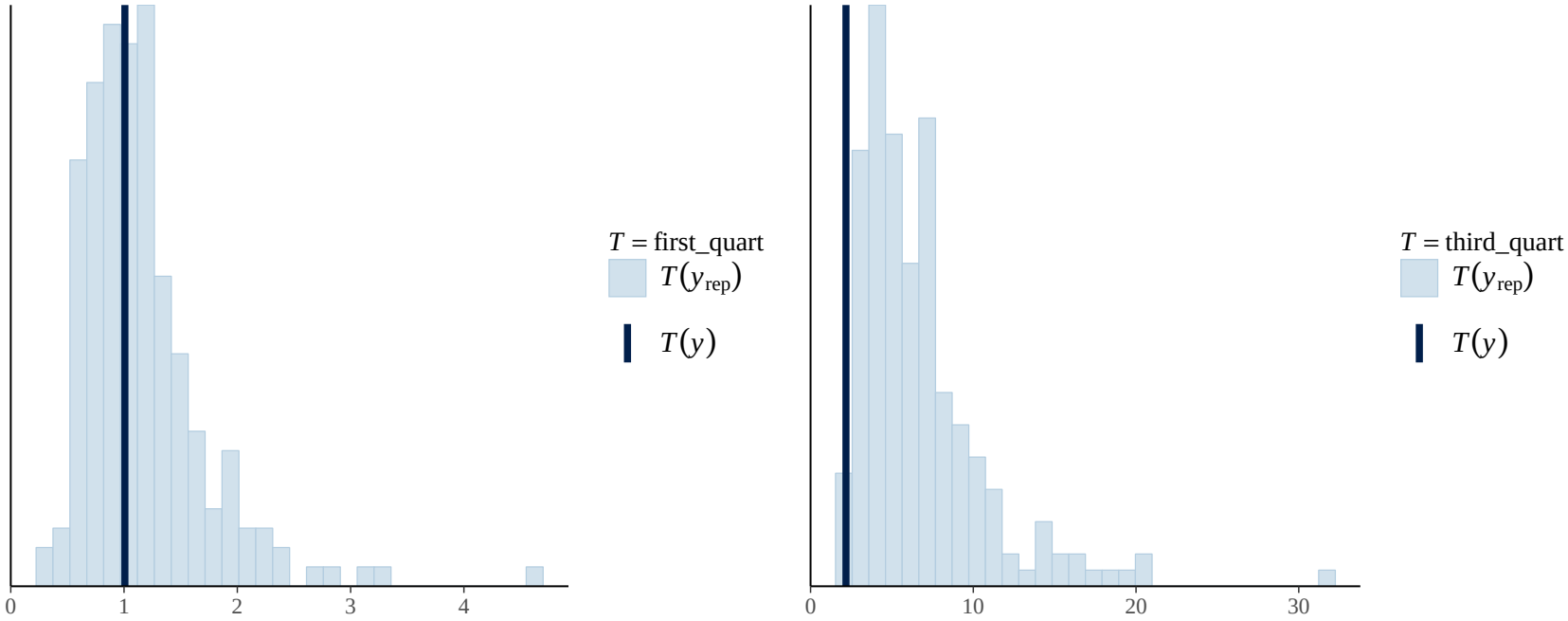
Gamma (1)



Gamma (2)


0.0
2.5
5.0
7.5
10.0
T = median
$T(y_{rep})$
$T(y)$
−0.5
0.5
0.0
2.5
5.0
7.5
10.0
0
10
20
30
T = median
$T(y_{rep})$
$T(y)$

Gamma (3)


0
1
2
3
4
T = first_quart
$T(y_{\text{rep}})$
$T(y)$
0
10
20
30
T = third_quart
$T(y_{\text{rep}})$
$T(y)$

Compare the models (1)

- Models can be compared by using leave-one-out cross-validation (LOO-CV)
- Expected log predictive density (ELPD) computed with LOO-CV can be used to evaluate which model has a better fit
- Predictive weights can be assigned to each model by using Stacking, Pseudo bayesian-model-averaging (Pseudo-BMA)
- Higher ELPD and predictive weights suggest better predictive performances

Compare the models (2)

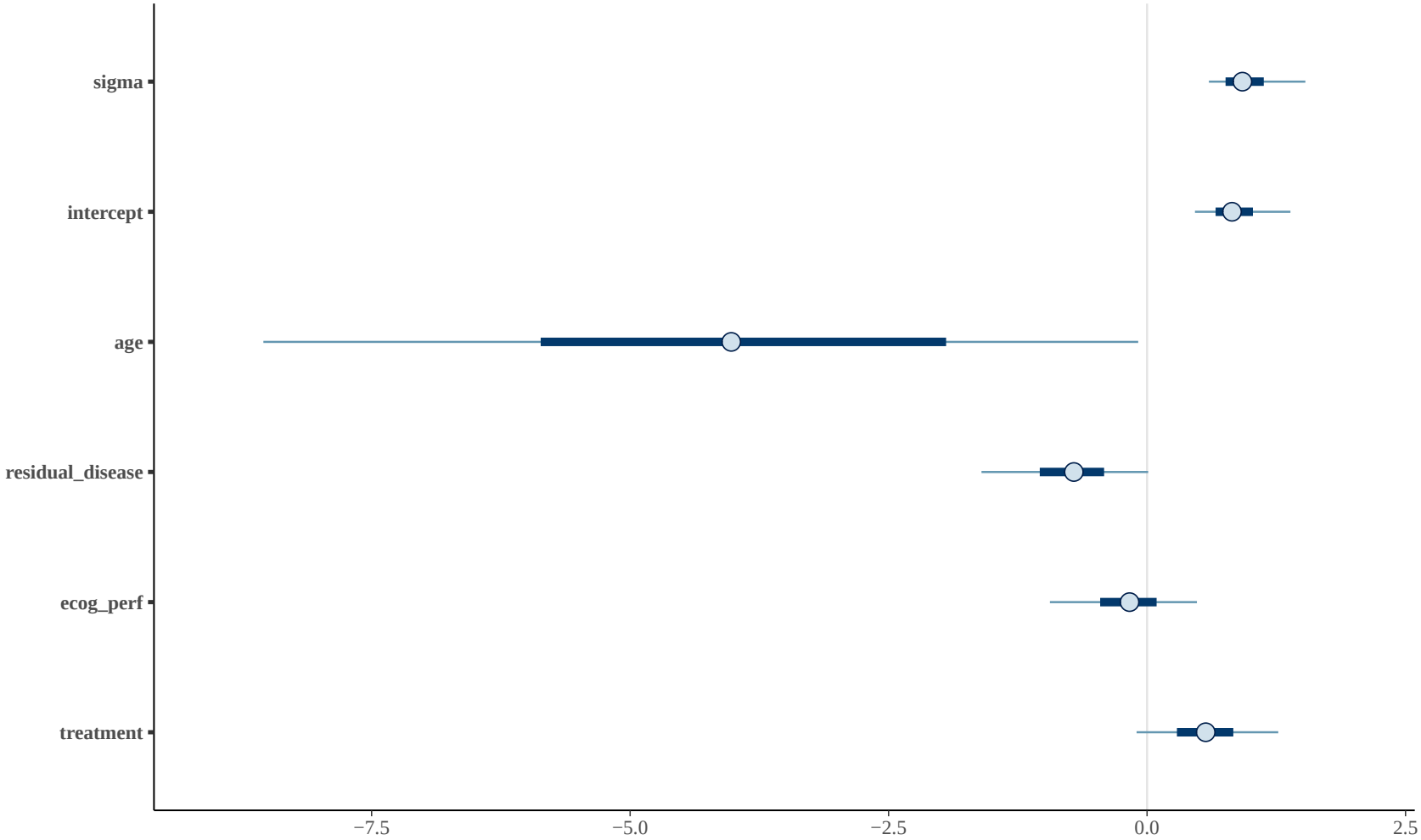
Table 2: Comparison of ELPD of the fitted models.

model	elpd_diff	elpd_loo	se_elpd_loo
lognormal	0.00	-23.95	3.13
gamma	-1.28	-25.23	3.27
weibull	-4.02	-27.97	3.40

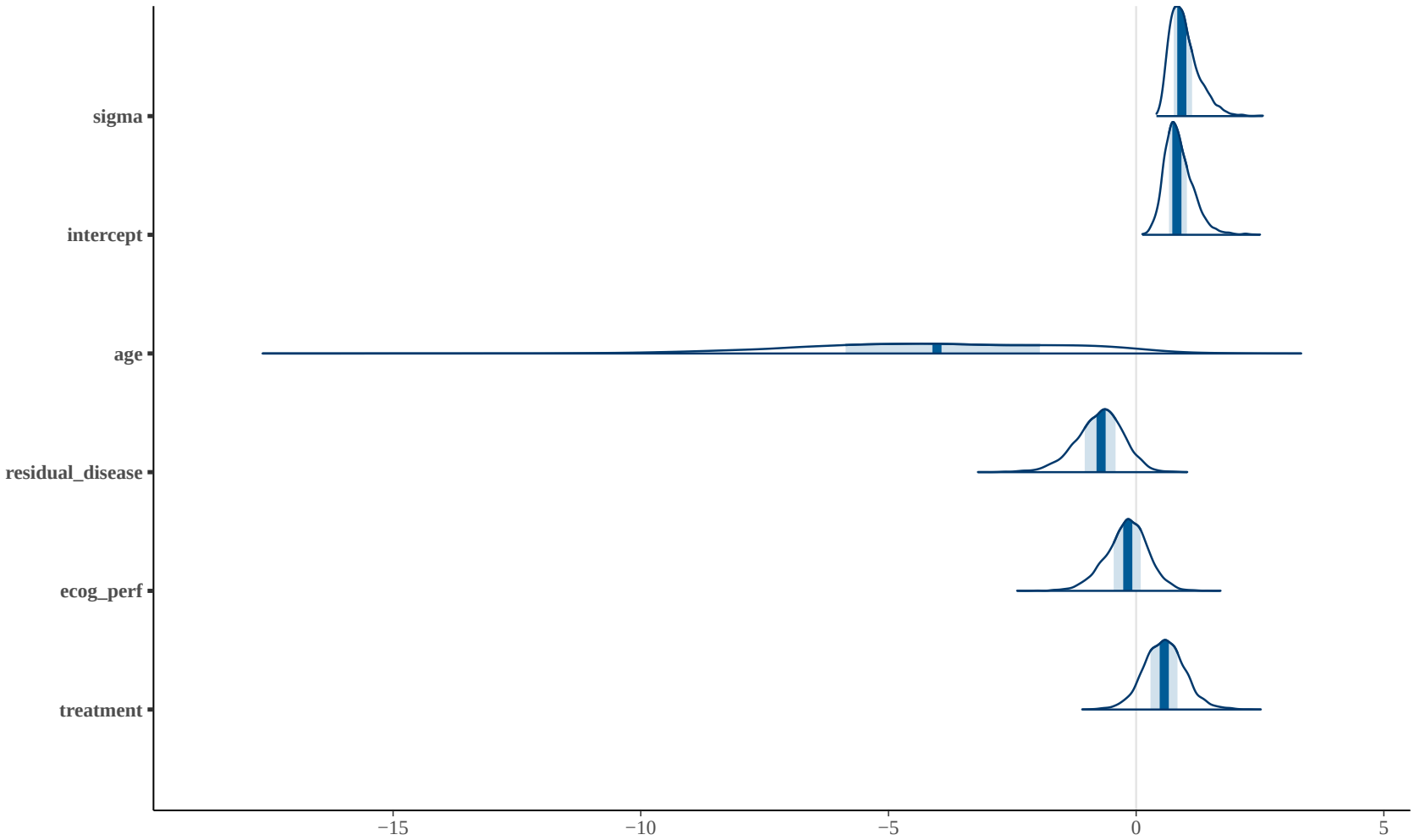
Table 3: Model comparison with Stacking, Pseudo-BMA and Pseudo-BMA with Bayesian Bootstrap.

model	stacking	pseudo_bma	pseudo_bma_bb
weibull	0	0.014	0.042
lognormal	1	0.772	0.742
gamma	0	0.214	0.216

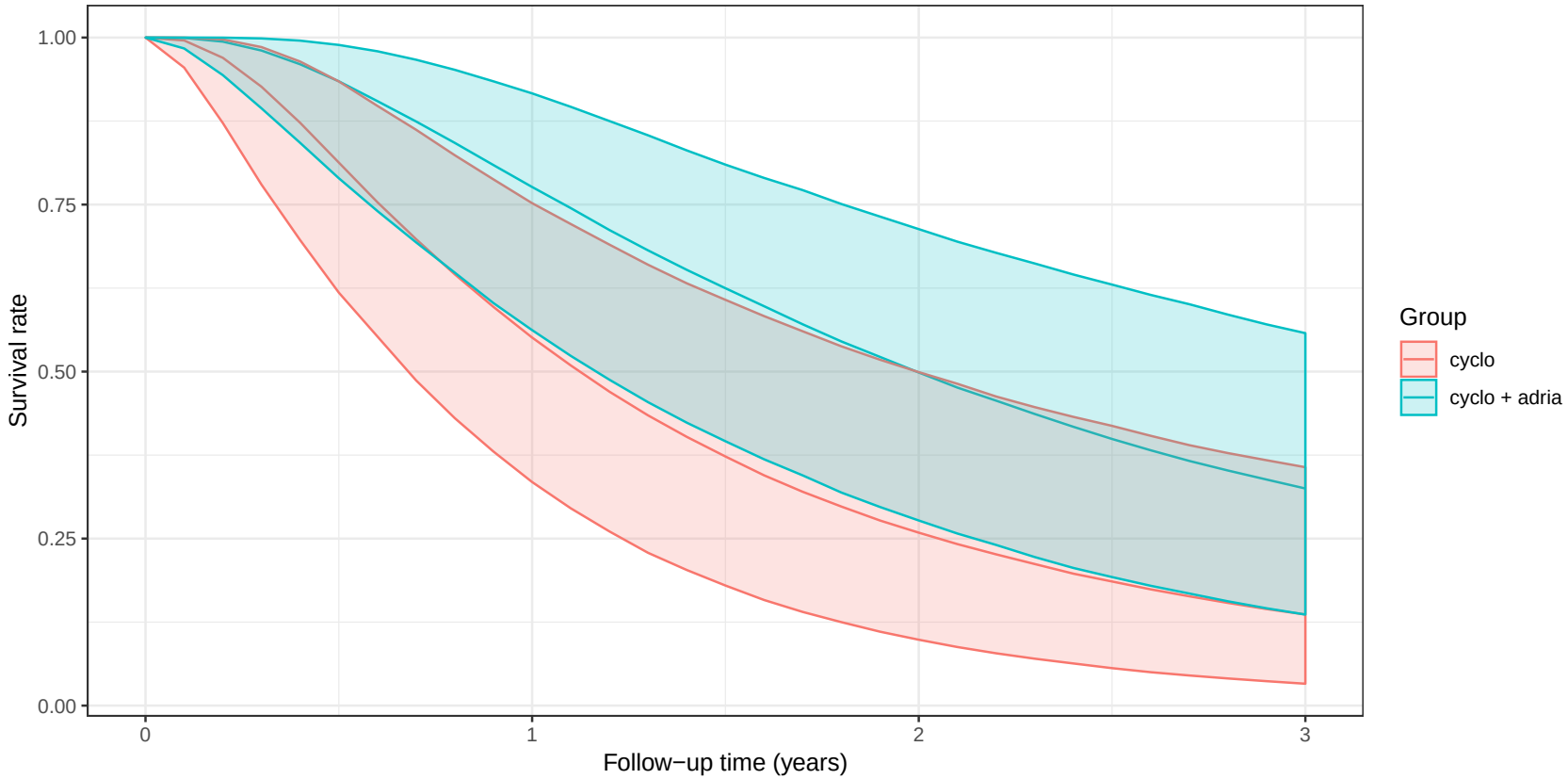
Parameters of the model (1)



Parameters of the model (2)



Posterior predictive survival curves



Logistic Regression Case Study

Influenza vaccine case study

- Hospitalized adults with acute respiratory disease tested for influenza with laboratory test (RT-PCR) (Talbot et al. 2013)
- The aim of the study was to estimate vaccine effectiveness in reducing risk of influenza
- Case-positive, control-negative study design
- Low prevalence of influenza ($\approx 10\%$)

The data

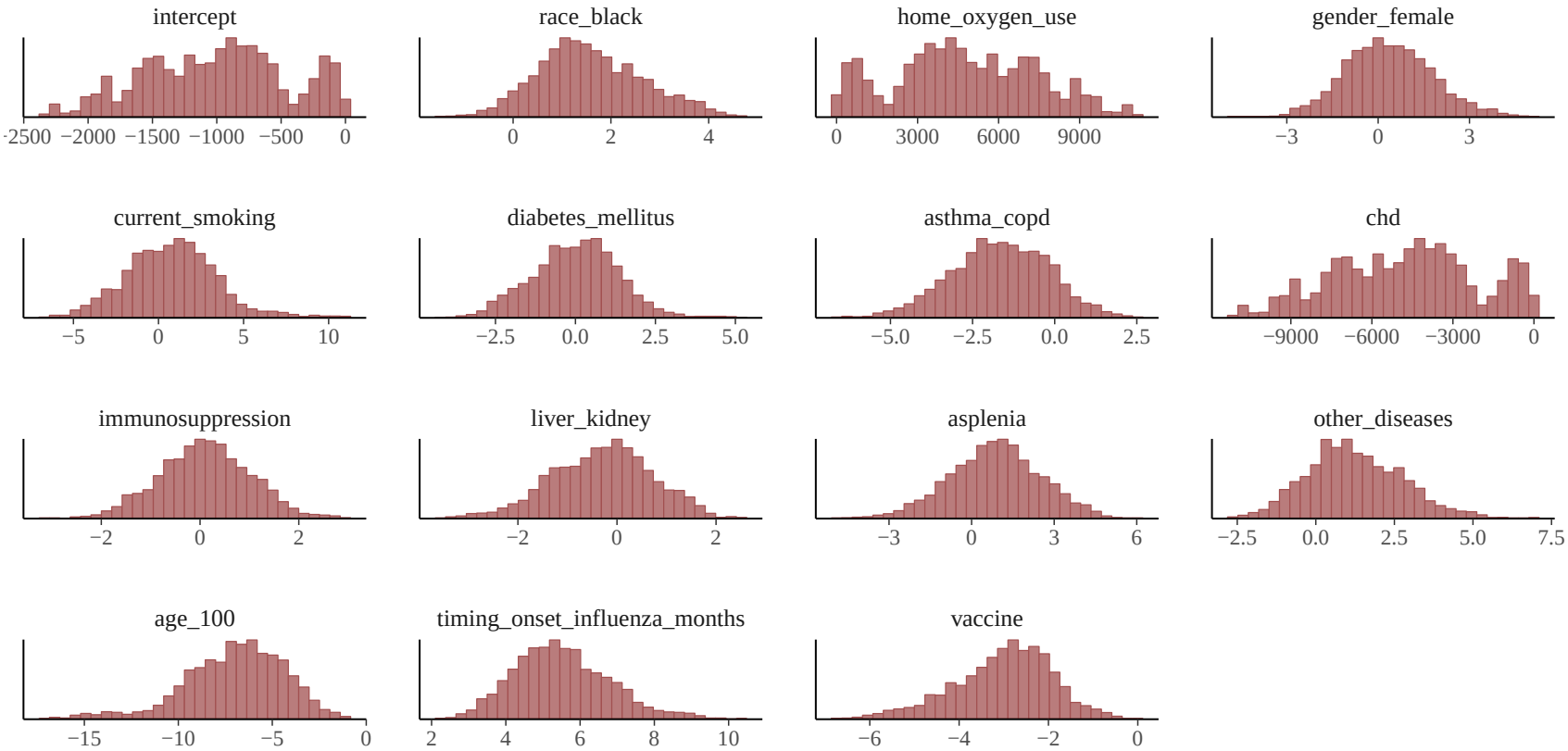
- Data were simulated from the information provided by *Chen et al. (2016)* (Chen et al., n.d.):
- 200 subjects
- 19 with positive influenza status and 119 with verified vaccination status
- 13 confounders: race, home oxygen use, current smoking status, diabetes mellitus, asthma chronic obstructive pulmonary disease, chronic heart disease, immunosuppression, chronic liver or kidney disease, asplenia, and other type of disease

The goal

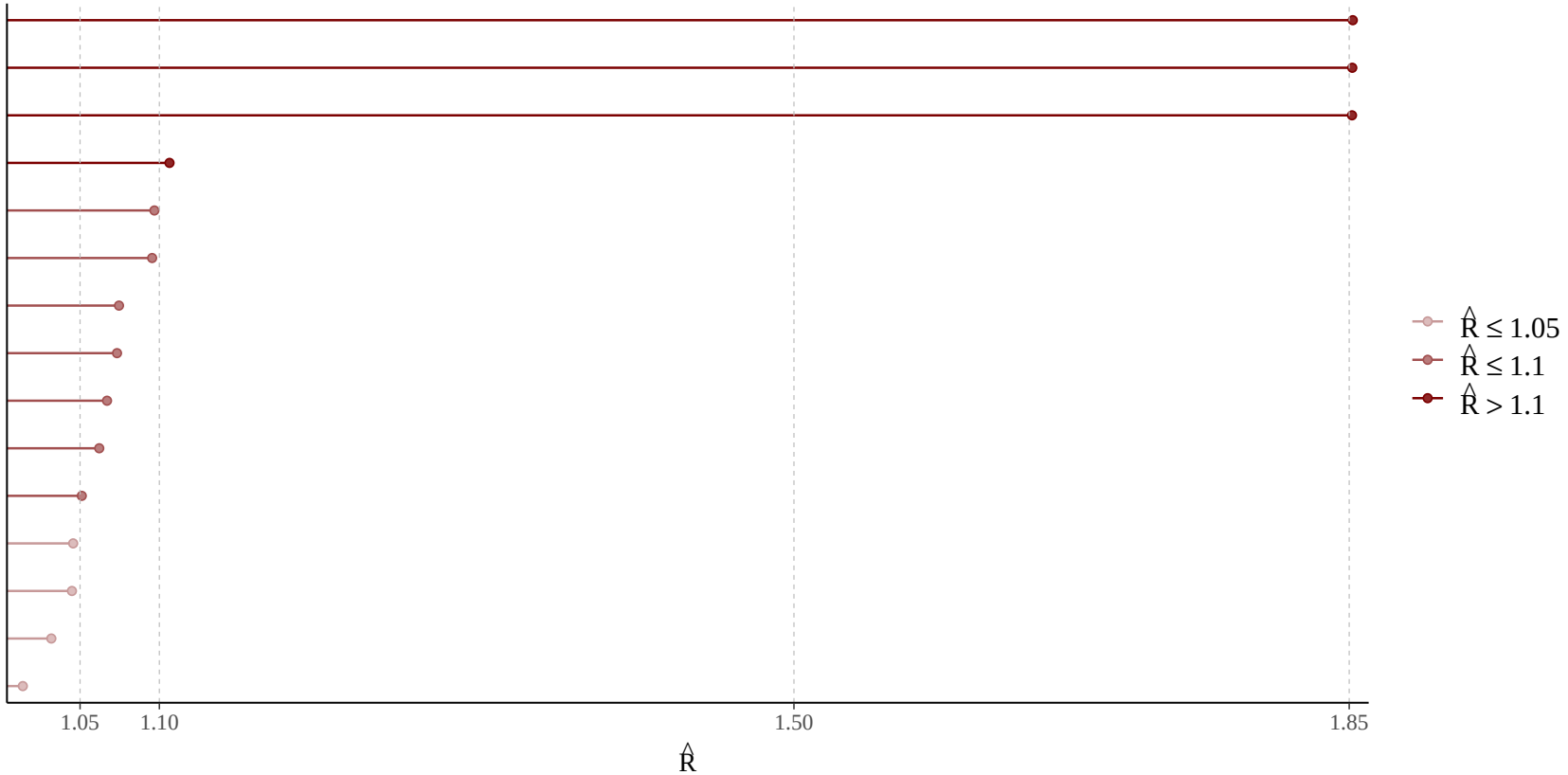
- With low number of cases and high number of covariates a standard logistic regression would overfit the data
- In their study, *Chen et al. (2016)* showed the benefits of penalizing maximum likelihood estimates (MLE) of all the terms in the model but the one related to the vaccination status
- Control both for overfitting and bias in the exposure estimate
- How can prior distributions help in such situations?

Non-informative uniform priors

Non-informative uniform priors



Non-informative uniform priors

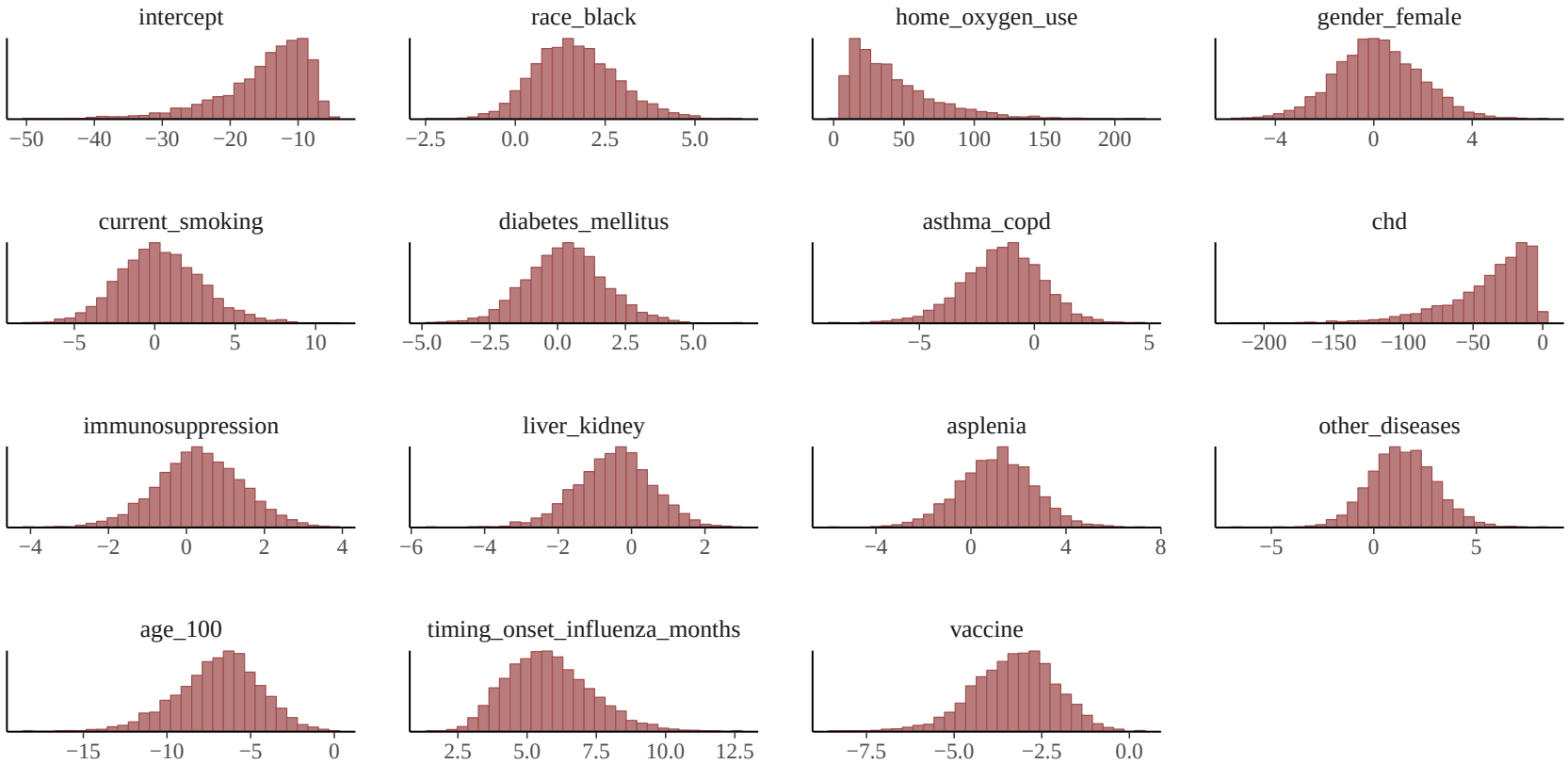


Non-informative uniform priors

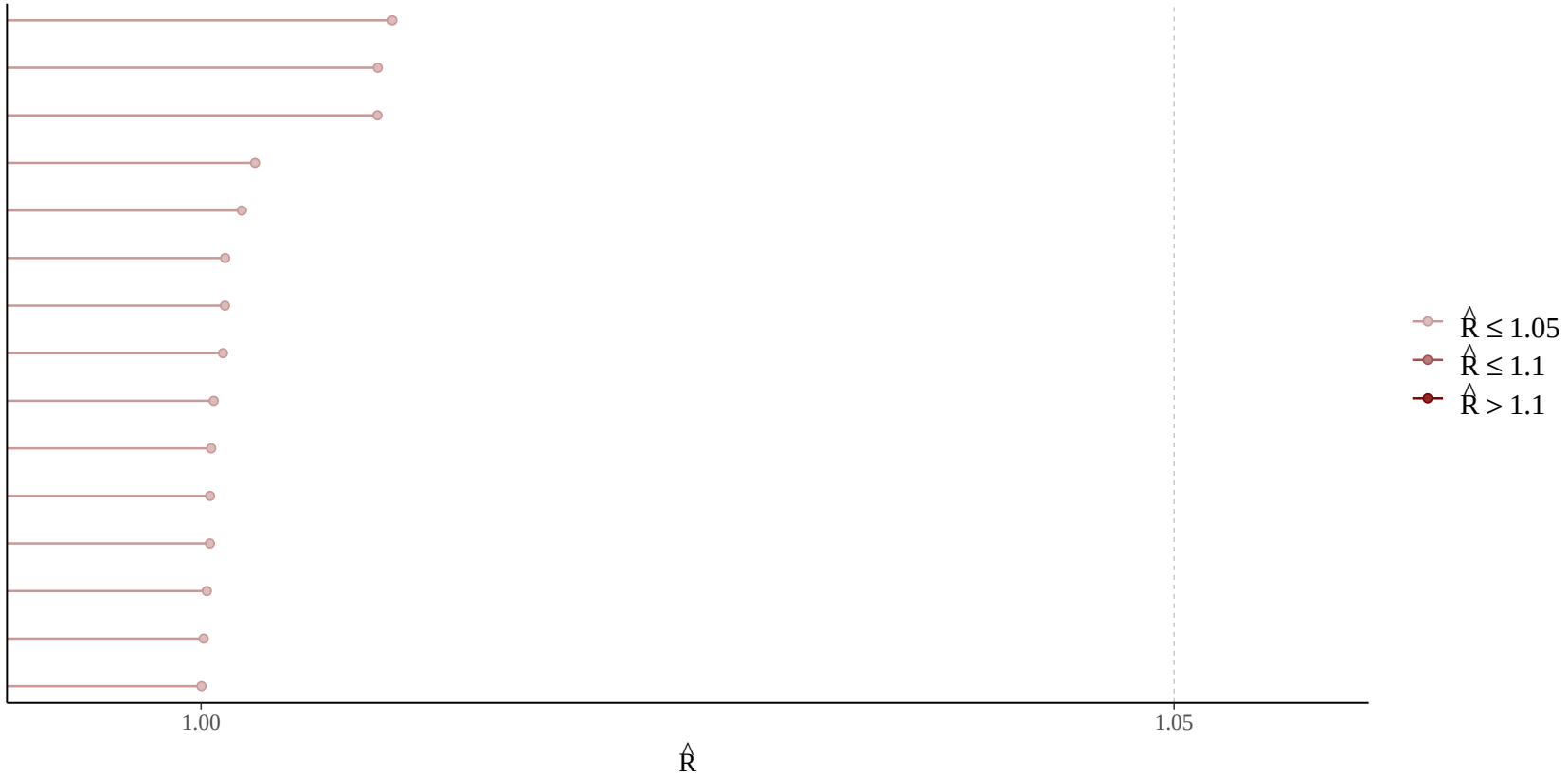
- Coefficients estimates are too extreme to be plausible
- The algorithm did not perform a good sample of the posterior:
  - Many divergent transitions
  - Low R_{hat} and ESS values
- The model likely overfitted the data
- Priors that allow for less extreme values may help

Vague priors: $N \sim (0, 100)$

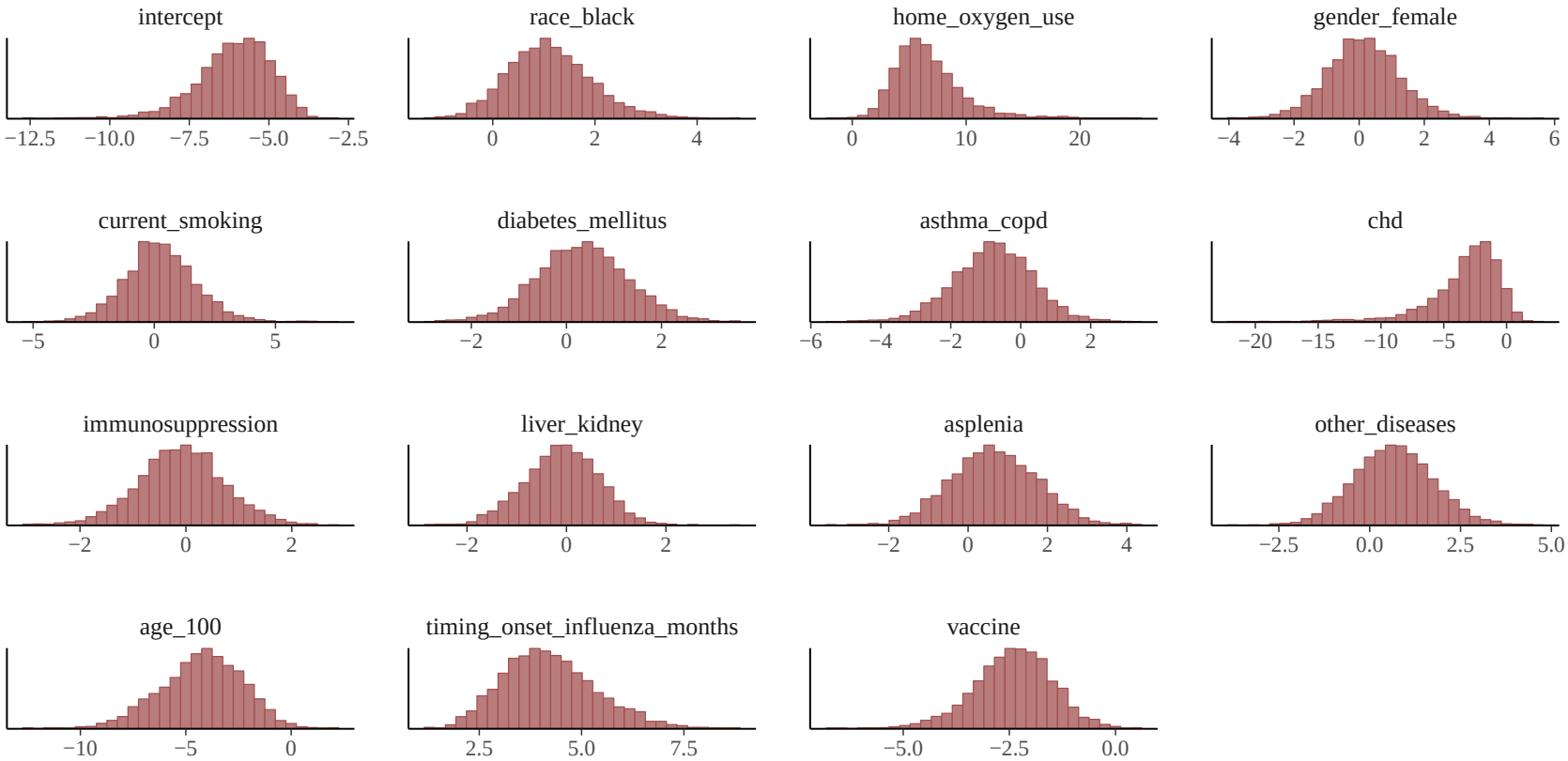
Vague priors: $N \sim (0, 100)$



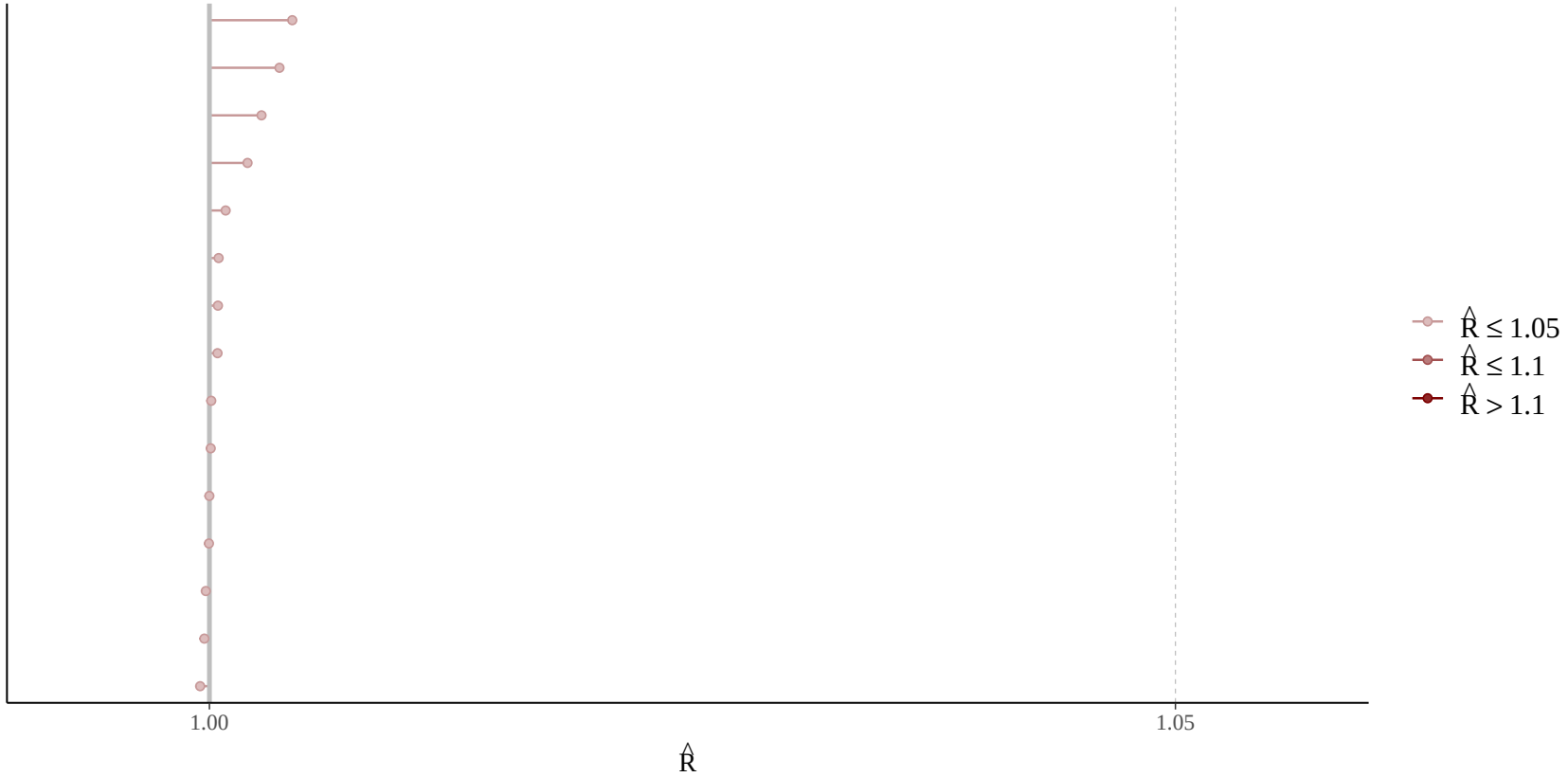
Vague priors: $N \sim (0, 100)$



$Cauchy \sim (0, 2.5)$



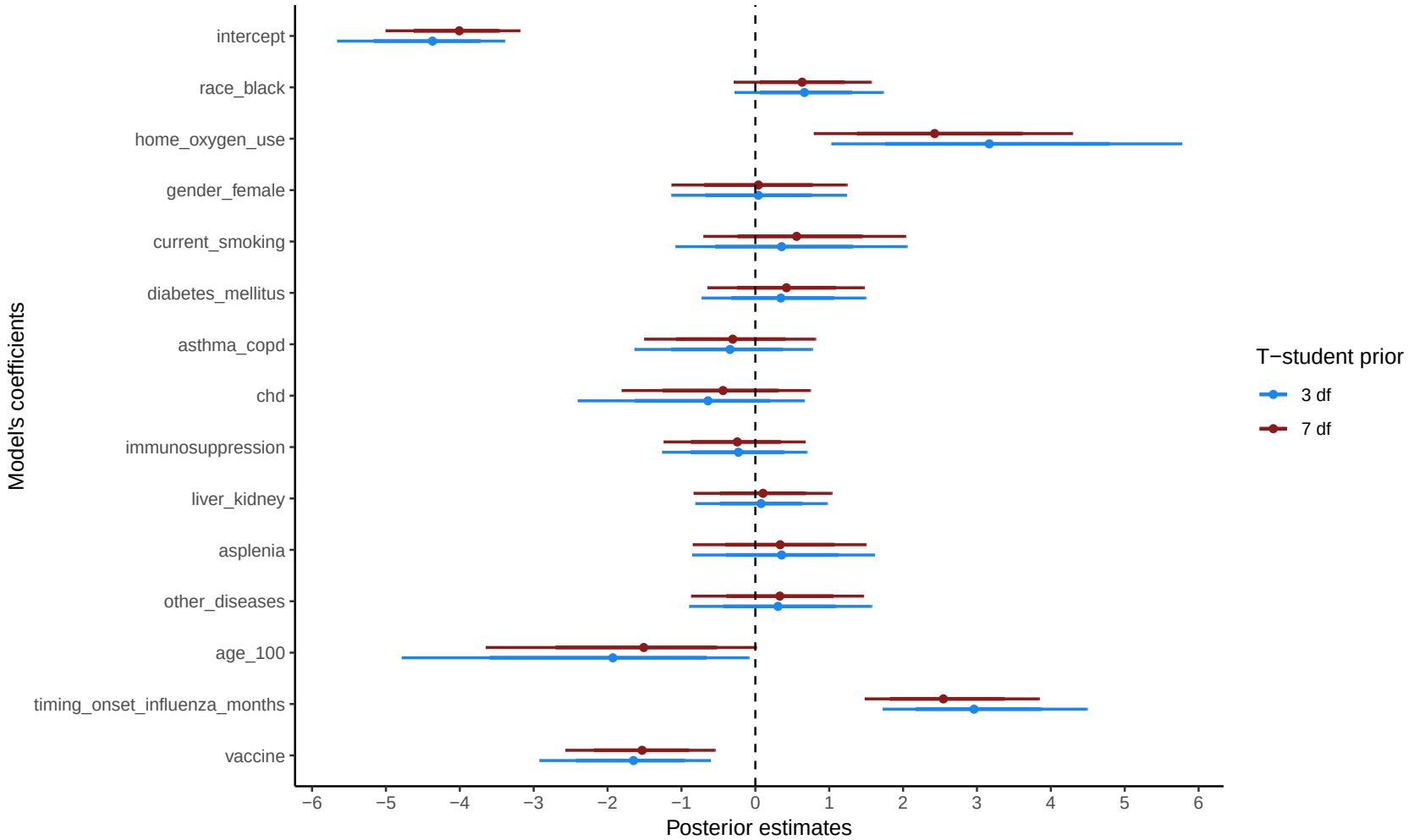
$Cauchy \sim (0, 2.5)$



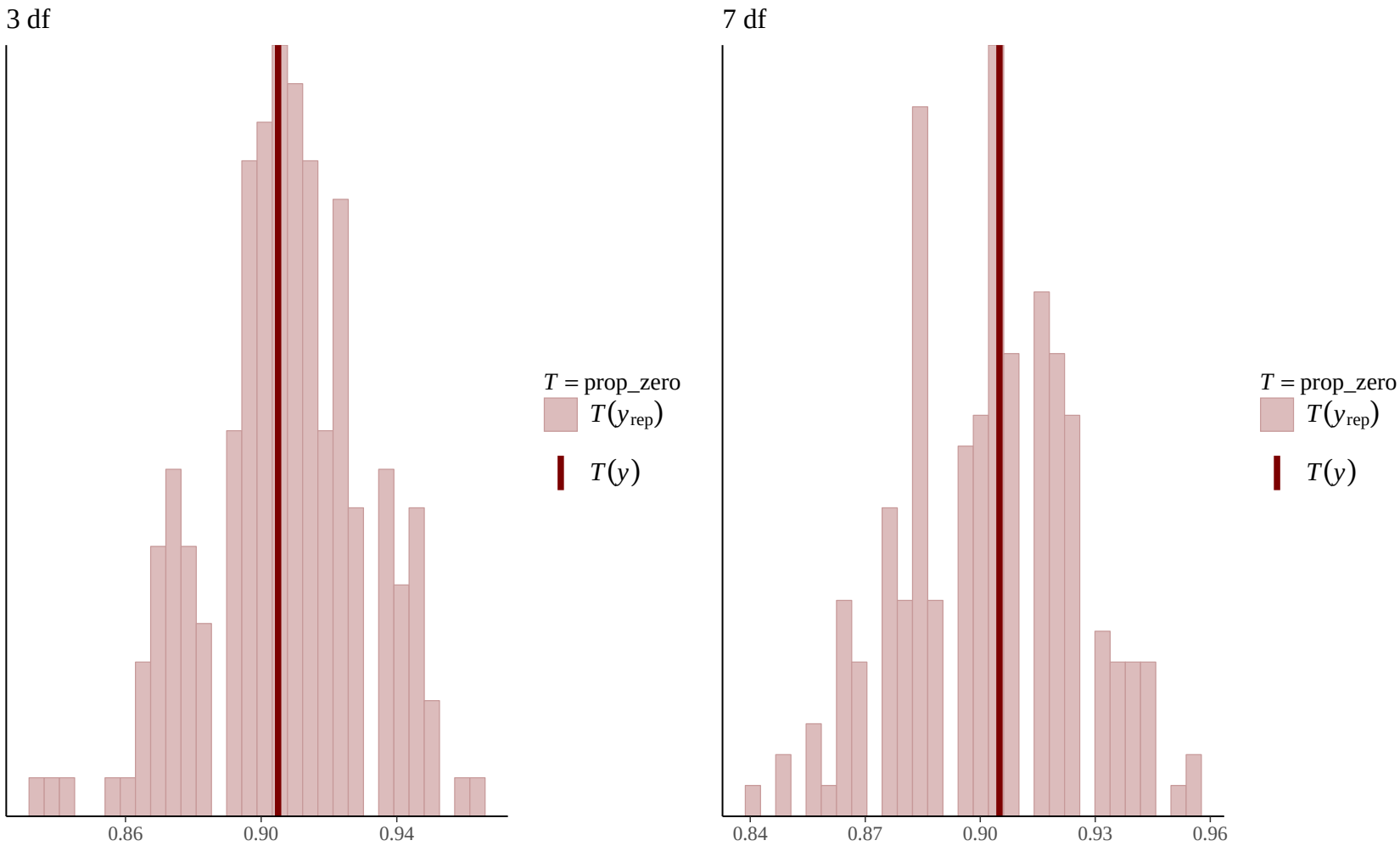
Weakly informative priors

- *Cauchy* $\sim (0, 2.5)$ improved the fit, but overfitting may be still present given the high values of coefficients
- Weakly informative priors may help in such situations to regularize inference by shrinking regression coefficients to 0
- The idea is to give more probability to values near the 0 while giving at the same time some chances to higher values
- If covariates are roughly on unit scale, $t - student \sim (df, 0, 1)$ with $3 \leq df \leq 7$ is a reasonable choice for logistic regression models

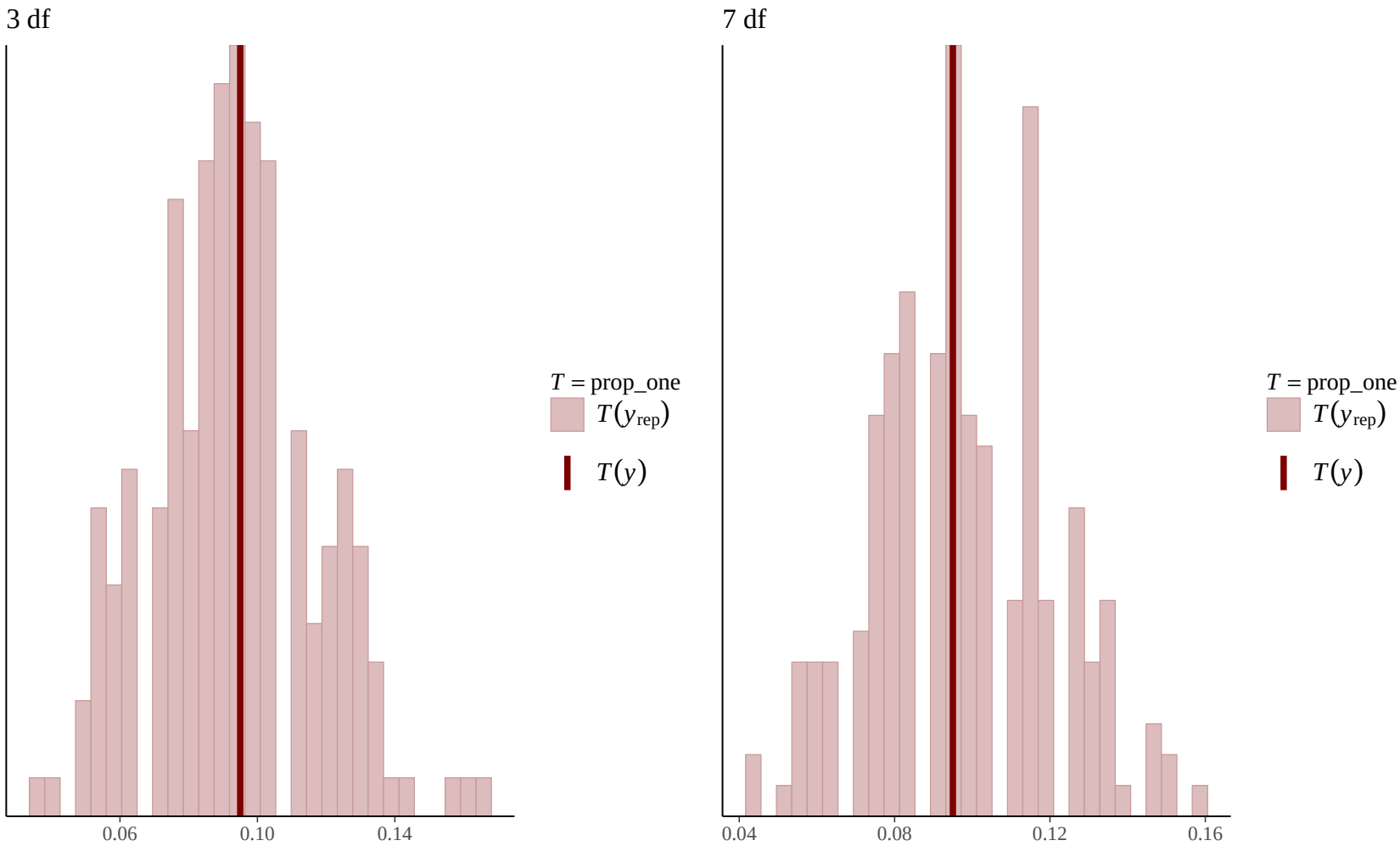
T-student priors: coefficients



T-student priors: posterior checks (1)



T-student priors: posterior checks (2)

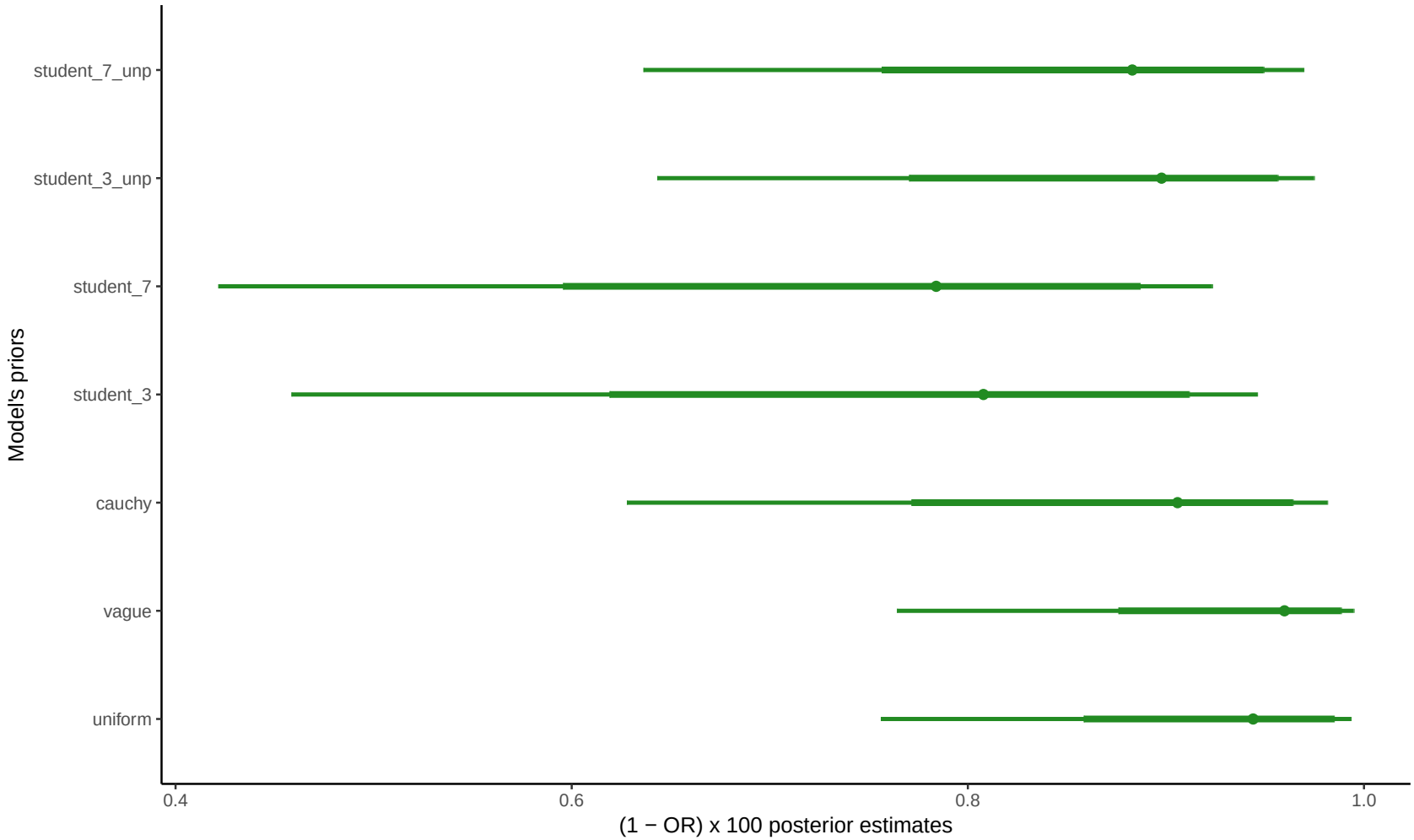


T-student priors: model comparison

Table 4: Model comparison with Stacking, Pseudo-BMA and Pseudo-BMA with Bayesian Bootstrap.

model	stacking	pseudo_bma	pseudo_bma_bb
student_t_3	0.579	0.333	0.307
student_t_7	0.001	0.179	0.202
student_t_3_unp	0.419	0.300	0.320
student_t_7_unp	0.000	0.188	0.171

Vaccine effectiveness



Additional information

- Stan's website at http://mc-stan.org/. Here you can find the reference manual, videos, tutorials, case studies and so on
- Here's a list of R packages that interface with Stan:
  - **rstan**
  - **bayesplot**
  - **loo**
  - **brms**
  - **rstanarm**
  - **trialr**
  - **RBesT**
  - **survHE**
- The slides of the presentation, the R and Stan codes used for the case studies are at https://github.com/danielebottigliengo/IBIG_2018

References

Chen, Qingxia, Hui Nian, Yuwei Zhu, H. Keipp Talbot, Marie R. Griffin, and Frank E. Harrell. n.d. "Too Many Covariates and Too Few Cases? – A Comparative Study." *Statistics in Medicine* 35 (25): 4546–58. doi:10.1002/sim.7021.

Edmonson, J. H., T. R. Fleming, D. G. Decker, G. D. Malkasian, E. O. Jorgensen, J. A. Jefferies, M. J. Webb, and L. K. Kvols. 1979. "Different Chemotherapeutic Sensitivities and Host Factors Affecting Prognosis in Advanced Ovarian Carcinoma Versus Minimal Residual Disease." *Cancer Treatment Reports* 63 (2): 241–47.

Talbot, H. Keipp, Yuwei Zhu, Qingxia Chen, John V. Williams, Mark G. Thompson, and Marie R. Griffin. 2013. "Effectiveness of Influenza Vaccine for Preventing Laboratory-Confirmed Influenza Hospitalizations in Adults, 2011–2012 Influenza Season." *Clinical Infectious Diseases* 56 (12): 1774–7. doi:10.1093/cid/cit124.