Posterior Summary

Jonathan Ma

2025-12-01

Table of Contents

1 Posterior Summary

This section summarizes the estimated posterior distributions of the model parameters, including fixed and random effects. We compute point estimates, credible intervals, and derive odds ratios and intra-class correlation metrics to assess variance partitioning.

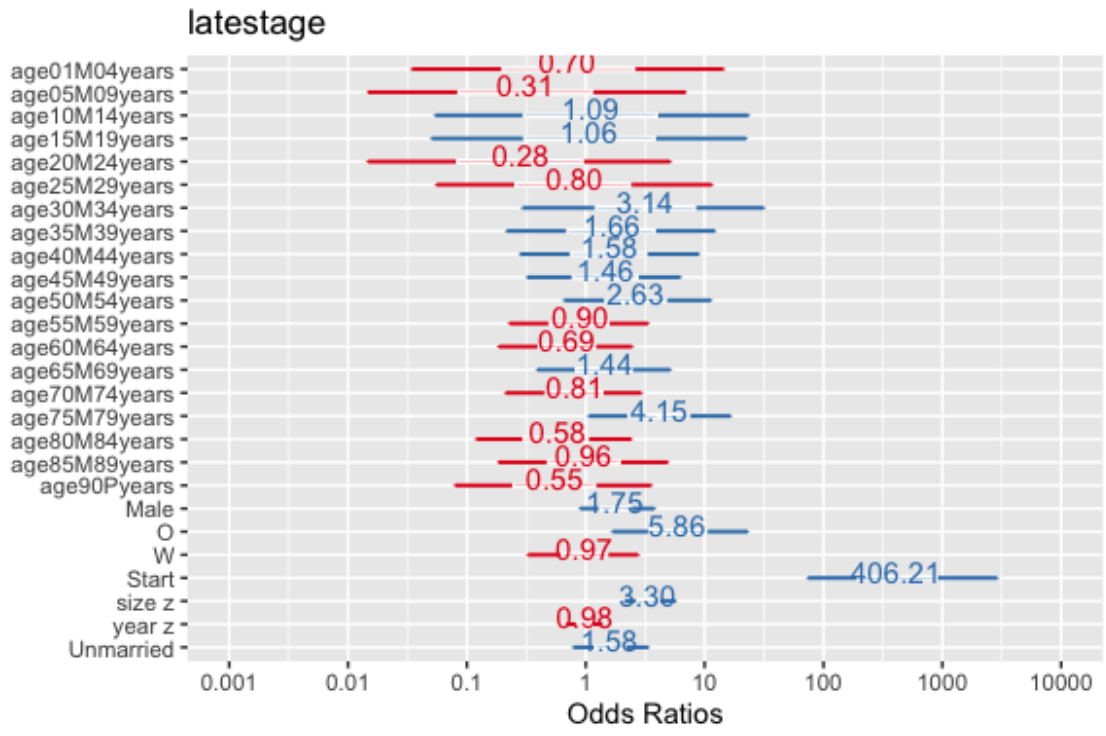
1.1 Summary

The following shows the first few rows of the full posterior summary for all parameters, including fixed effects, group-level effects, and auxiliary parameters.

```
##                     Estimate Est.Error      Q2.5     Q97.5
## b_Intercept     -4.32250000  1.362928 -7.174176 -1.752077
## b_age01M04years -0.33931969  1.886682 -4.046930  3.323349
## b_age05M09years -1.16955713  1.915065 -4.824186  2.580967
## b_age10M14years  0.08948779  1.882779 -3.473392  3.905725
## b_age15M19years  0.07176377  1.904186 -3.653903  3.866177
## b_age20M24years -1.27660843  1.826520 -4.950319  2.311217
```

1.2 Fixed Effects

Below we visualize the fixed effect posterior estimates and 95% credible intervals. Coefficients are on the log-odds scale. Positive values indicate increased odds of late-stage diagnosis.



The odds ratio plot displays the estimated likelihood of being diagnosed at a late stage of cancer across various predictors, with 95% confidence intervals. Values above 1 indicate higher odds of late-stage diagnosis, while values below 1 suggest lower odds. Younger patients—particularly those aged 01–04, 05–09, 10–14, and 15–19 years—have odds ratios well below 1, indicating significantly lower risk of late-stage diagnosis. In contrast, certain middle-aged groups, such as 35–39 and 50–54 years, show elevated odds ratios above 2 or even 3, suggesting increased risk. Gender and marital status also play a role: male and unmarried patients are more likely to be diagnosed late. Notably, one covariate possibly related to tumor size or diagnosis year has an extremely high odds ratio of over 400, indicating a strong association with late-stage detection, though such a large value may suggest issues like rare events or data sparsity.

1.3 Random Effects (group level variation)

These posterior summaries indicate that region-level effects are weak or negligible in influencing late-stage diagnosis once individual-level variables are accounted for. The wide intervals reflect uncertainty, possibly due to small sample sizes within regions or limited regional variation.

```
## , , Intercept
##
##      Estimate Est.Error      Q2.5    Q97.5
## 1 -0.03564271 0.7260108 -1.559148 1.469797
## 2 -0.08212961 0.7922132 -1.831757 1.453806
## 3  0.15199696 0.7598713 -1.249647 1.824940
```

1.4 Odds Ratios

The odds ratio (OR) is the exponentiated posterior mean (or median) of each coefficient. This table highlights the most influential predictors in the model. Strong, well-estimated effects include early tumor grade, other race, and tumor size. Age effects are mixed, with older adults generally at higher risk, but credible intervals widen in extreme age groups due to fewer observations.

```
##                    Estimate Est.Error         Q2.5        Q97.5
## gradeStart    427.47742764  3.080237 5.260676e+01 4472.3118442
## raceth0         5.99385960  2.256193 1.337940e+00   30.9401709
## age75M79years   4.13938554  2.364600 7.846040e-01   22.6481428
## size_z          3.37119134  1.355896 1.988640e+00    6.4890874
## age30M34years   3.11891155  4.312735 1.632972e-01   59.1595807
## age50M54years   2.65375774  2.435053 4.823600e-01   15.8427305
## Intercept       0.01326668  3.907619 7.661164e-04    0.1734133
## sexMale         1.78551454  1.575105 7.537673e-01    4.5340126
## age20M24years   0.27898188  6.212231 7.081147e-03   10.0866975
## age05M09years   0.31050442  6.787379 8.033092e-03   13.2099102
```

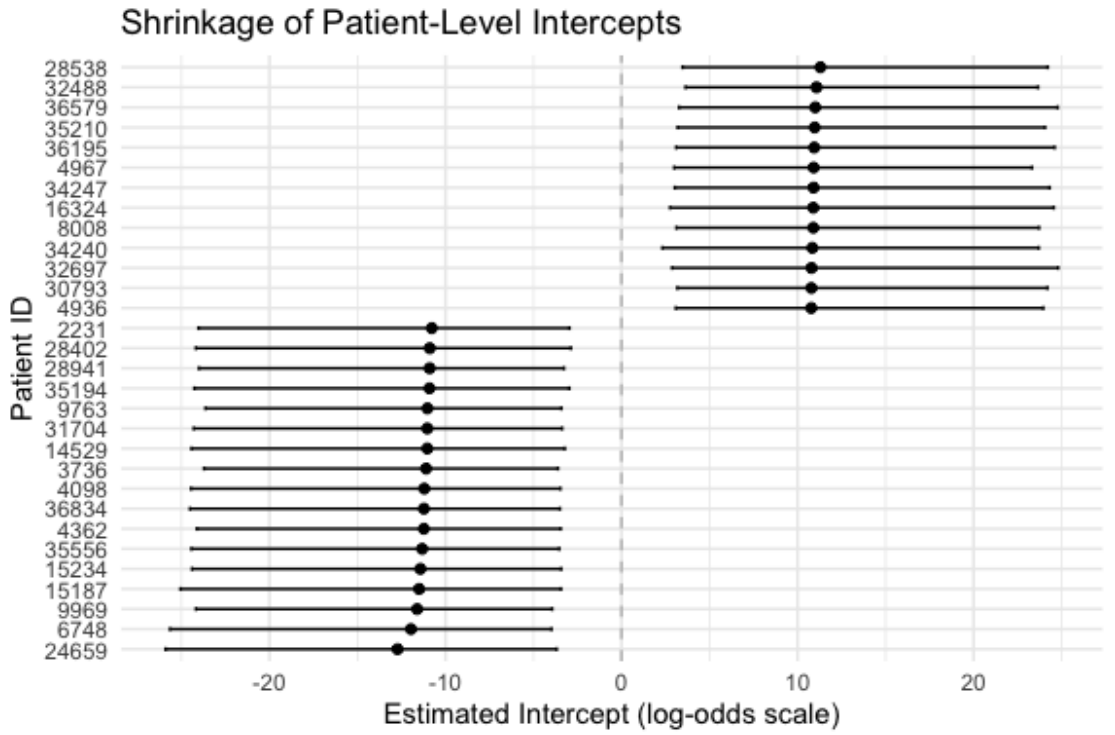
1.5 Marginal Effects

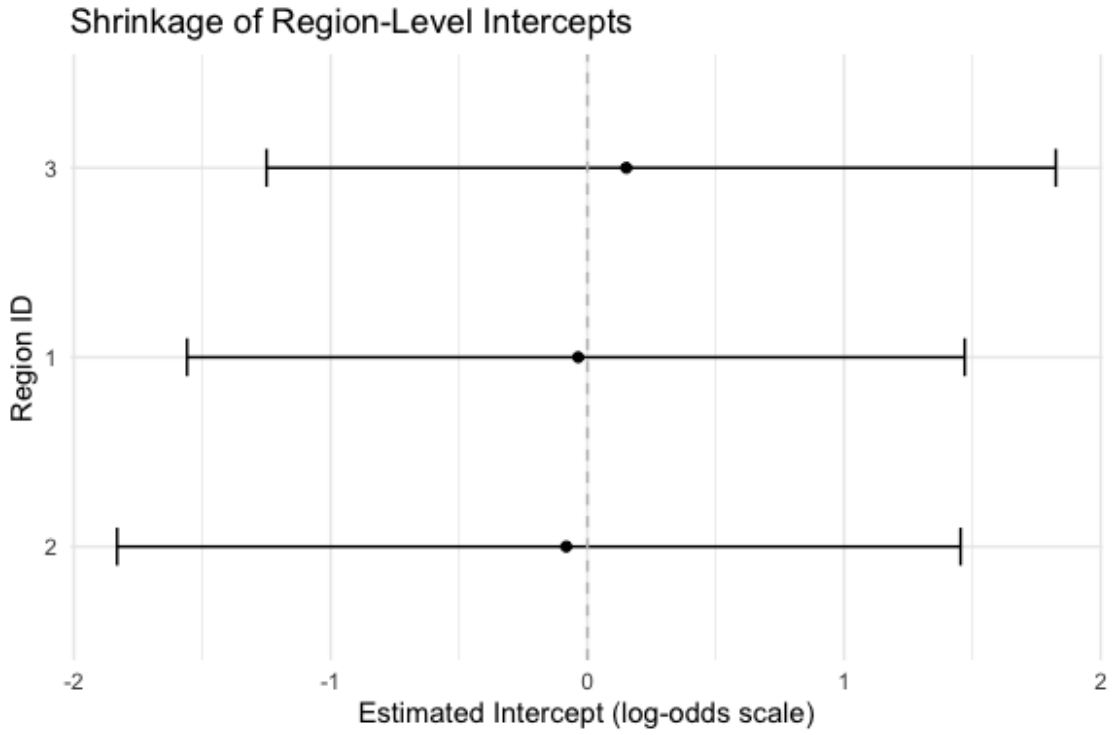
The marginal effects table confirms that tumor size and grade are the strongest predictors of the outcome. Age and race are meaningful, but less dominant. Sex, marital status, and year exert smaller effects. The magnitude and range of these effects help prioritize which variables matter most for prediction and interpretation.

```
## # A tibble: 7 × 4
##   variable avg_range avg_estimate n_levels
##   <chr>        <dbl>        <dbl>    <int>
## 1 size_z       0.470       0.227       100
## 2 grade        0.384       0.429         2
## 3 age          0.286       0.0174       20
## 4 raceth       0.241       0.0337        3
## 5 sex          0.183       0.0188        2
## 6 marry        0.171       0.0177        2
## 7 year_z       0.146       0.0136      100
```

1.6 Shrinkage EFfects

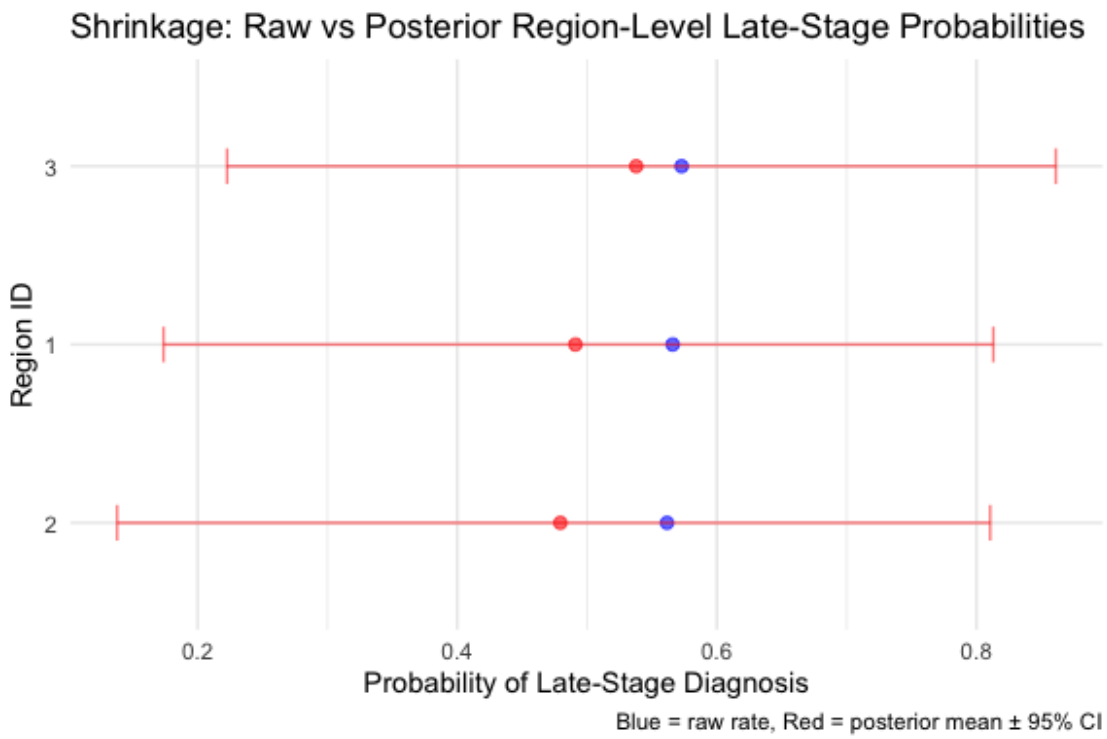
The shrinkage plots illustrate the estimated patient- and region-level random intercepts on the log-odds scale, capturing unobserved heterogeneity after accounting for covariates. At the patient level, there is substantial variation, with some individuals showing highly negative intercepts (indicating much lower odds of the outcome) and others showing strongly positive intercepts (indicating much higher odds), suggesting that latent individual-level factors significantly influence outcome risk. These estimates are subject to shrinkage, where extreme values are pulled toward the population mean to stabilize inference. In contrast, the region-level intercepts exhibit minimal variability: all three regions have posterior means near zero with wide credible intervals that span zero, indicating no strong regional effects after adjusting for other variables. Together, the plots reinforce that most of the residual variation lies at the patient level, not the regional level, justifying the hierarchical structure of the model.





1.7 Raw vs Shrunk Estimates

Red points represent posterior-predicted probabilities, which incorporate both observed data and model priors. Blue points are empirical averages (raw rate). Posterior estimates are shrunk toward the global mean, especially for small regions.



The plot visualizes the effect of Bayesian shrinkage on region-level estimates of late-stage diagnosis probabilities. Each region's raw proportion (blue dots) is plotted alongside the corresponding posterior mean and 95% credible interval (red dots and error bars). Across all three regions, the raw rates of late-stage diagnosis vary slightly, but the posterior estimates are drawn toward the global average due to partial pooling. This is especially evident in regions with smaller sample sizes or greater uncertainty, where the posterior means are more conservative and the intervals wider.

1.8 Intra-Class Correlation

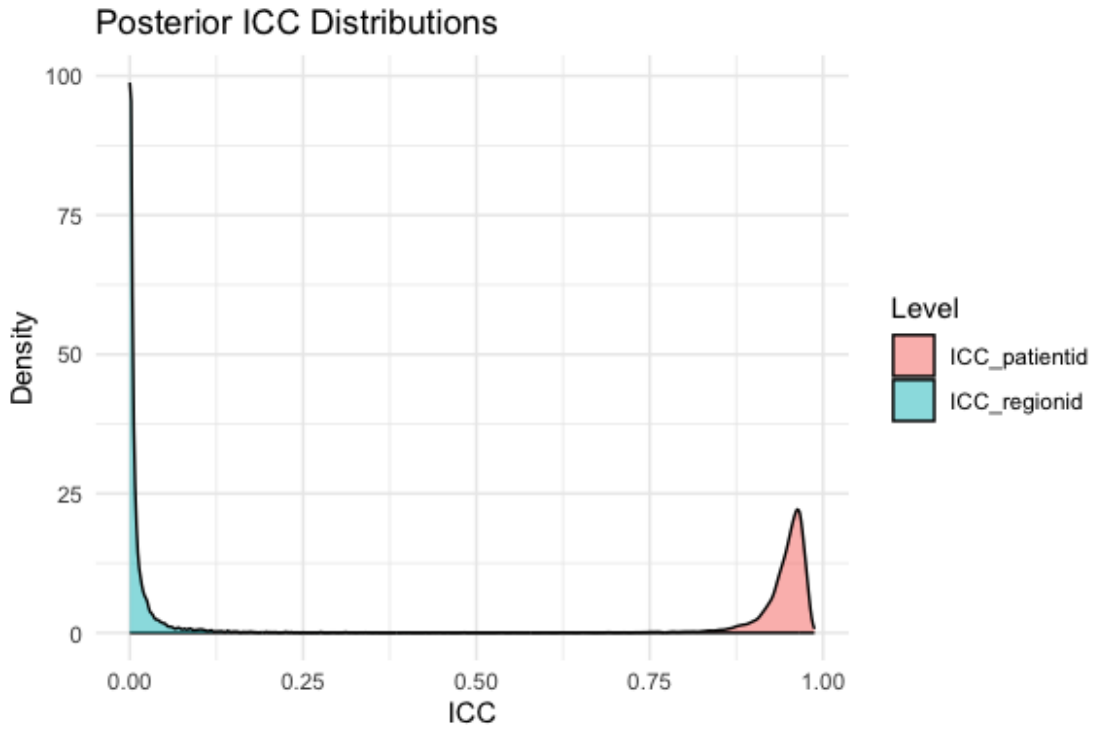
ICC (intra-class correlation) quantifies the proportion of total variance attributable to grouping structure. Here we compute ICC for both patients and regions.

Posterior Intra-Class Correlation Estimates

Group	SD	ICC	CI
Patient	9.1081563	0.9427376	[0.841, 0.978]
Region	0.7869934	0.0150378	[0.000, 0.112]

The Patient-Level ICC (0.943) indicates that approximately 94.3% of the total variance in outcomes is due to differences between patients, with a narrow 95% credible interval of [0.841, 0.978]. This suggests strong patient-specific heterogeneity, meaning individual-level characteristics explain most of the variability in diagnosis outcomes. In contrast, the region-level ICC is 1.5%, with a wider credible interval of [0.000, 0.112], suggesting that very little variance is explained by differences between regions. The lower bound of zero

and the upper bound just above 11% indicate weak and uncertain regional effects, possibly due to regional homogeneity or limited sample size at the region level.



The posterior ICC distribution plot vividly illustrates the stark contrast in variance explained at the patient and region levels. The red density curve, representing patient-level ICCs, is sharply concentrated near 1.0, indicating that for nearly all posterior samples, a very high proportion of outcome variability is attributed to individual patients. This reinforces strong within-patient clustering and heterogeneity in late-stage diagnosis risk. Conversely, the blue density curve for region-level ICCs is concentrated near 0, suggesting that region explains virtually none of the outcome variation, with most samples falling below 0.05. The non-overlapping and polarized nature of these distributions confirms that individual-level effects dominate, while regional variation is minimal and likely negligible in the context of this hierarchical model.

2 Preparing Data for Clustering

We want to do BGMM, so we need to standardizee values after extracting. Since we are interested in patient clustering, we will use dfRE_P, the summary for patient level random effects.