Project 2

Kent Codding

2025-03-31

read in packages, data

```
library(rjags)
```

```
## Warning: package 'rjags' was built under R version 4.4.2
```

```
## Loading required package: coda
```

```
## Warning: package 'coda' was built under R version 4.4.2
```

```
## Linked to JAGS 4.3.1
```

```
## Loaded modules: basemod,bugs
```

```
library(coda)
library(jagsUI)
```

```
## Warning: package 'jagsUI' was built under R version 4.4.2
```

```
##
## Attaching package: 'jagsUI'
```

```
## The following object is masked from 'package:coda':
##
##     traceplot
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.4.2
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```r
library(ggridges)
```

```
## Warning: package 'ggridges' was built under R version 4.4.3
```

```r
path <- "CXCL14_dat_7.txt"
df <-
read.delim2(path, header = TRUE, sep = " ", dec = ".")
```

Project Questions

1.

Do the studies support the conclusion that CXCL14 is a useful biomarker?

2.

In a hypothetical new study, what is the probability that a patient with a CXCL14 expression level of your choice will have an optimal surgery? Please make your own necessary assumptions about parameters in the new study, and state them clearly.

3.

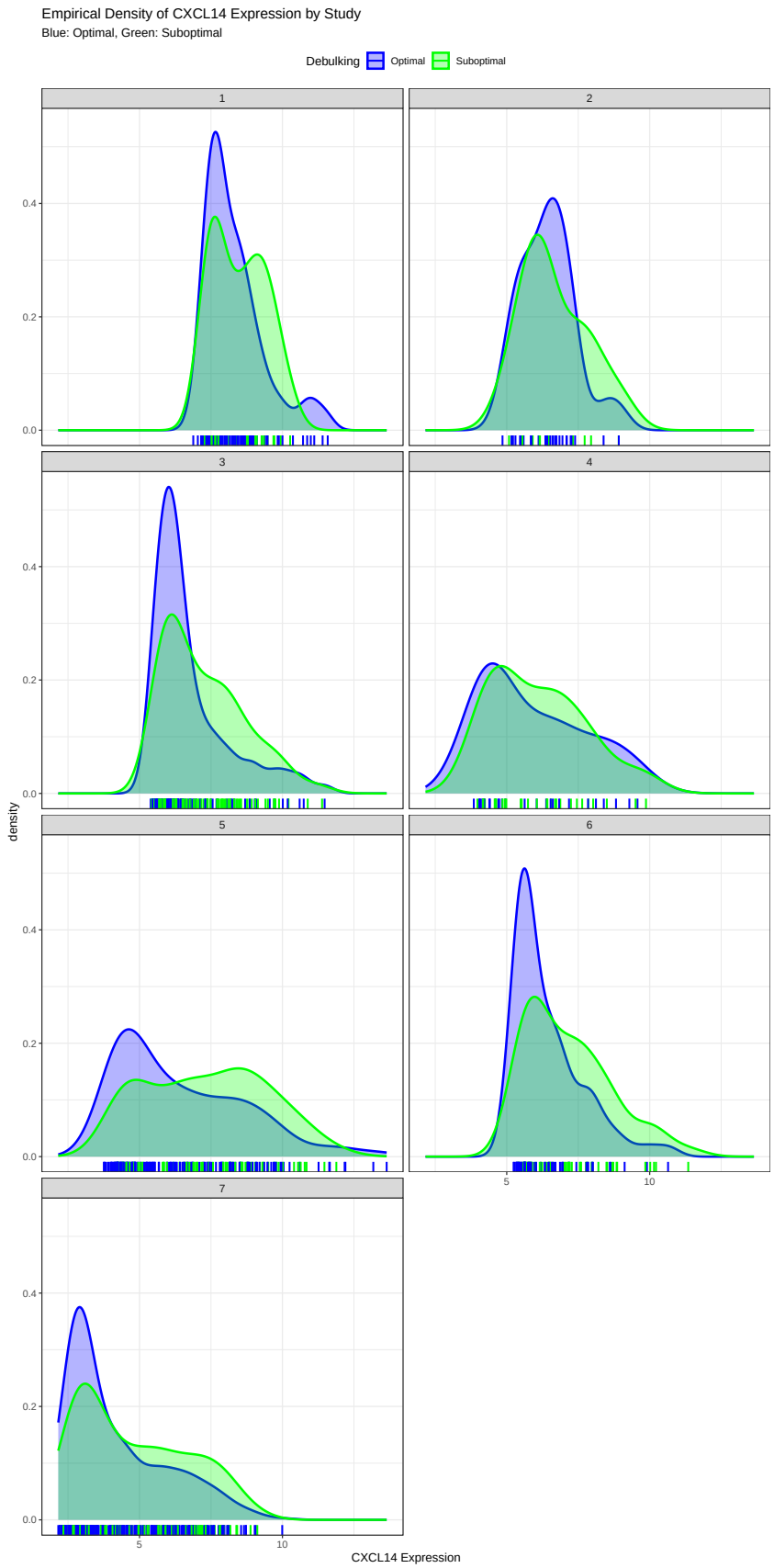
What is the probability that a hypothetical new study with 100 patients will show a significant p-value ($<$ 0.05) for the difference between the two surgical outcomes?

Methods

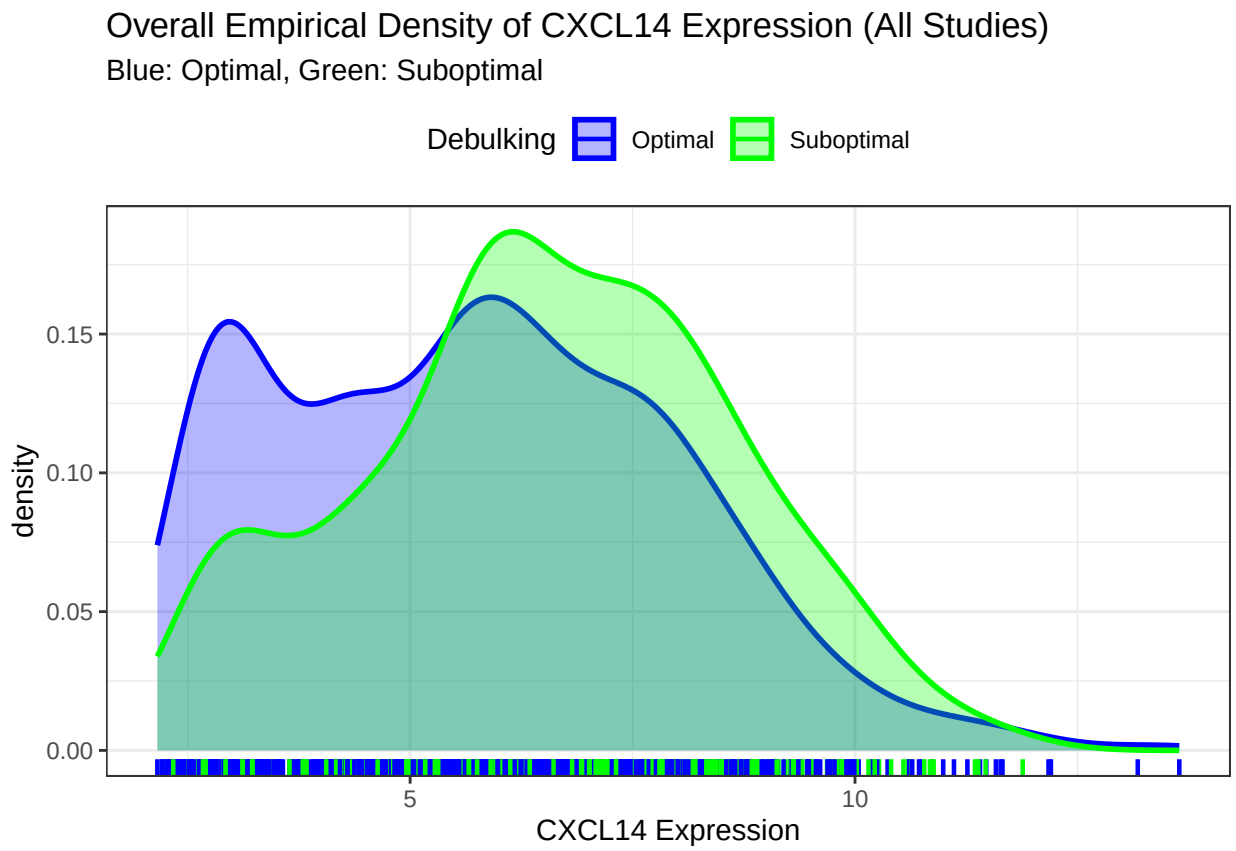
EDA

Here, I perform exploratory data analysis to plot gene expression across studies in order to inform prior choice, patient-level distributions, and study-level distributions. I have chosen to stack the denisty curves for each study on the same x-axis.

```r
# new grouping variable for debulking
df <- df %>%
  mutate(Debulking = ifelse(YY == 1, "Optimal", "Suboptimal"))

ggplot(df, aes(x = XX, color = Debulking, fill = Debulking)) +
  geom_density(alpha = 0.3, linewidth = 1) +
  geom_rug(aes(color = Debulking), linewidth = 0.8) +
  facet_wrap(~ZZ, ncol = 2, scales = "fixed") +
  labs(x = "CXCL14 Expression",
       title = "Empirical Density of CXCL14 Expression by Study",
       subtitle = "Blue: Optimal, Green: Suboptimal") +
  scale_color_manual(values = c("Optimal" = "blue", "Suboptimal" = "green")) +
  scale_fill_manual(values = c("Optimal" = "blue", "Suboptimal" = "green")) +
  theme_bw() +
  theme(legend.position = "top",
        strip.text = element_text(size = 10),
        plot.margin = unit(c(0.5, 0.5, 0.5, 0.5), "cm"))
```


Empirical Density of CXCL14 Expression by Study
Blue: Optimal, Green: Suboptimal
Debulking
Optimal
Suboptimal
1
2
3
4
5
6
7
density
0.0
0.2
0.4
5
10
CXCL14 Expression

```r
ggplot(df, aes(x = XX, color = Debulking, fill = Debulking)) +
  geom_density(alpha = 0.3, linewidth = 1) +
  geom_rug(aes(color = Debulking), linewidth = 0.8) +
  labs(x = "CXCL14 Expression",
       title = "Overall Empirical Density of CXCL14 Expression (All Studies)",
       subtitle = "Blue: Optimal, Green: Suboptimal") +
  scale_color_manual(values = c("Optimal" = "blue", "Suboptimal" = "green")) +
  scale_fill_manual(values = c("Optimal" = "blue", "Suboptimal" = "green")) +
  theme_bw() +
  theme(legend.position = "top")
```



Model Bayesian MCMC

a) Choice of patient-level distributions and their parameterizations

Pooling patient-level distributions assumes that all 1221 patients have the same relationship of CXCL14-debulking status. That is, all patients are treated as the same population. The assumption of homogeneity allows for efficiency and maximization of statistical power to estimate an overall effect. The study-level distributions can account for the heterogeneity that occurs from different studies analyzing different representative samples of the population

b) Choice of study-level distributions; decisions about which parameters are model hierarchically versus independently across studies (e.g. in the binary case we modeled lambda's hierarchically and theta0's independently)

For study-level distributions, partial pooling is used in order to account for the aforementioned heterogeneity that likely arises due to population and specific scientific method differences among the 7 studies. For hierarchically modeling effects, λ_j will represent the study-specific CXCL14 on the binary outcome of optimal debulking where $\lambda_j \sim \mathcal{N}(\mu_\lambda, \tau^{-2})$. Partial pooling also performs shrinkage on study-specific estimates, reducing overfitting by incorporating global information. Lastly, the global intercept parameter will allow interpretation to focus on the varying-slope parameter λ_j, which is the individual-study effect on increase in log-odds of optimal debulking.

Observed Data:

For each patient $i = 1, \ldots, N,$ in study $j[i],$ observe:

$X_i:$ CXCL14 expression (continuous),

$Y_i:$ Debulking outcome (1 = optimal, 0 = suboptimal).

Model Structure:

$$Y_i \sim \text{Bernoulli}(p_i),$$

$$\text{logit}(p_i) = \alpha + \lambda_{j[i]} X_i,$$

$$\lambda_j \sim \mathcal{N}(\mu_\lambda,\, \tau^{-2}), \quad j = 1, \ldots, J.$$

Hyperpriors:

$$\alpha \sim \mathcal{N}(0,\, 0.01),$$

$$\mu_\lambda \sim \mathcal{N}(0,\, 0.01),$$

τ (precision) via continuous prior; high precision was chosen

to reflect small between-study differences.

Interpretation:

α is a global intercept (baseline log-odds of optimal debulking).

λ_j is the study-specific slope for CXCL14,

allowing partial pooling across studies (shrinkage).

μ_λ is the overall mean effect of CXCL14,

while τ controls between-study heterogeneity.

c) Choice of prior distributions

Since each study's parameters in the partial pooling model come from an overarching prior distribution, I assign each study-specific slope λ_j a Normal prior which can be justified by the Central Limit Theorem that is centered around a common mean μ_λ. The parameter $lambda_j$ represents the increase in the log-odds of optimal debulking with each unit increase in CXCL14 expression, and this parameter is drawn from the overarching population distribution. Partial pooling also shrinks studies with fewer data points towards the overall mean μ_λ, only allowing strong evidence to pull the posterior $lambda_j$ away from μ_λ.

The hyperprior $\mu_\lambda \sim \mathcal{N}(0,\, 0.01)$ centers the global log-odds effect at 0 with a weakly informative variance, allowing the data to play a large role in the posterior. The discrete prior for τ constrains the model to

different levels of heterogeneity as done in Section. Since the normal distribution in Jags is parameterized by precision, the same logic was employed with the global α intercept parameter.

```
cxcl14_model <- "
model {
  for (i in 1:N) {
    # Bernoulli likelihood for each patient
    Y[i] ~ dbern(p[i])

    # Logit link with single global intercept (alpha)
    # and a study-specific slope (lambda[study[i]]) for X[i].
    logit(p[i]) <- alpha + lambda[study[i]] * X[i]
  }

  # prior for study-specific slopes
  for (j in 1:N_studies) {
    lambda[j] ~ dnorm(mu_lambda, precision_tau)
  }
  # hyperprior: mean baseline effect (intercept)
  alpha ~ dnorm(0.0, 0.01)

  # hyperprior: mean effect of CXCL14
  mu_lambda ~ dnorm(0.0, 0.01)

  # discrete prior for tau (similar to section 8 example)
  # this controls for between-study heterogeneity in lambda.
  tau.int.prior <- rep(0.03, 25)
  tau.int ~ dcat(tau.int.prior)
  tau <- tau.int * 0.01
  precision_tau <- 1 / (tau * tau)
}
"
```

```
df <- df %>%
  mutate(study_index = as.numeric(as.factor(ZZ)))

model_data <- list(
  N = nrow(df),
  N_studies = length(unique(df$study_index)),
  X = df$XX,
  Y = df$YY,
  study = df$study_index
)


jags_mod <- jags.model(textConnection(cxcl14_model),
                       data = model_data,
                       n.chains = 2,
                       n.adapt = 500,
                       quiet = TRUE)

# PARAMETER reminders
# 'lambda': the vector of study-specific slopes
# 'alpha' : the global intercept
# 'mu_lambda': the overall mean slope (distribution where lambda is drawn from )
```

```r
# 'tau.int', 'tau' etc. for the discrete prior
params <- c("lambda", "alpha", "mu_lambda", "tau.int", "tau")
jags_fit <- coda.samples(jags_mod, params, n.iter = 5000, thin = 5)

summary(jags_fit)
```

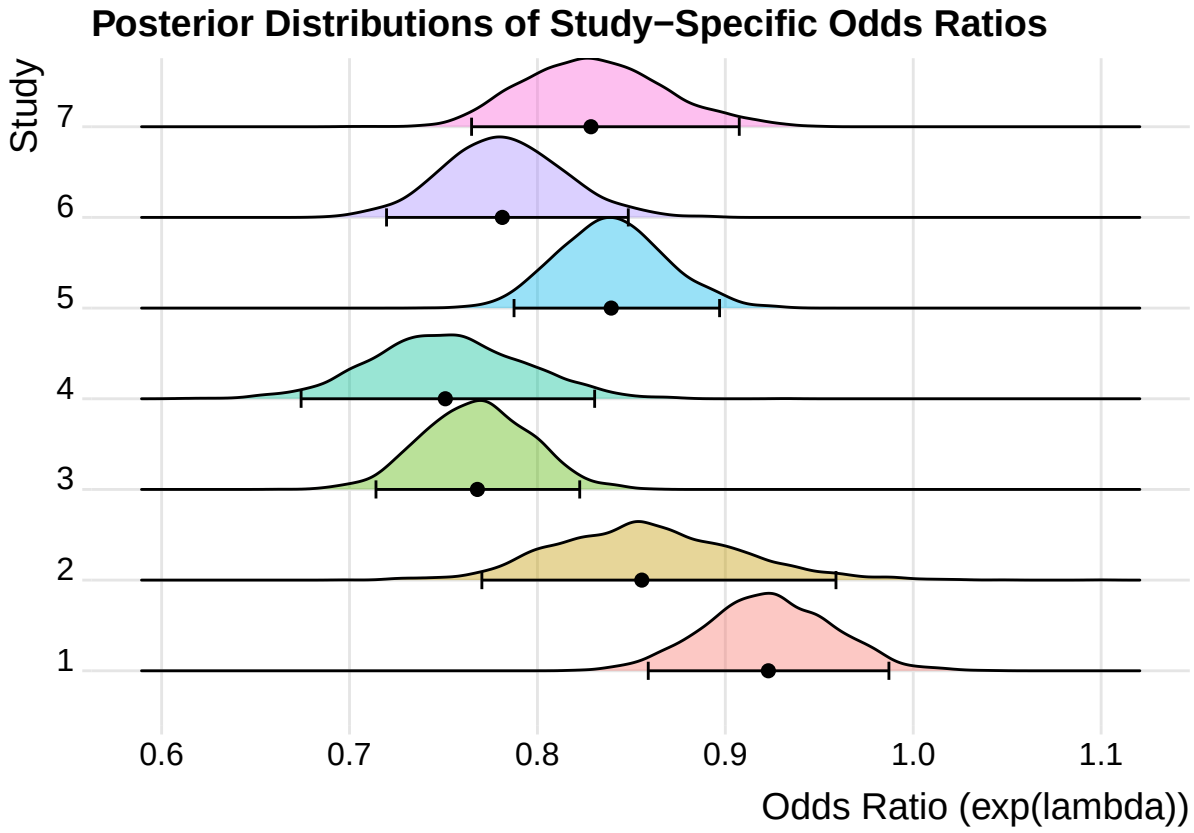
```
##
## Iterations = 505:5500
## Thinning interval = 5
## Number of chains = 2
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##               Mean      SD  Naive SE Time-series SE
## alpha      1.83503 0.21841 0.0048837      0.0136852
## lambda[1] -0.08057 0.03633 0.0008124      0.0019033
## lambda[2] -0.15455 0.05683 0.0012709      0.0021806
## lambda[3] -0.26480 0.03703 0.0008280      0.0018963
## lambda[4] -0.28665 0.05386 0.0012043      0.0022534
## lambda[5] -0.17519 0.03371 0.0007537      0.0018141
## lambda[6] -0.24636 0.04108 0.0009186      0.0020866
## lambda[7] -0.18699 0.04421 0.0009887      0.0026040
## mu_lambda -0.19984 0.05193 0.0011612      0.0021542
## tau        0.09784 0.03873 0.0008660      0.0008662
## tau.int    9.78350 3.87287 0.0866000      0.0866201
##
## 2. Quantiles for each variable:
##
##              2.5%     25%      50%      75%    97.5%
## alpha      1.4101  1.6927  1.83285  1.98559  2.26023
## lambda[1] -0.1520 -0.1044 -0.08023 -0.05528 -0.01302
## lambda[2] -0.2608 -0.1932 -0.15601 -0.11704 -0.04198
## lambda[3] -0.3368 -0.2899 -0.26391 -0.23920 -0.19540
## lambda[4] -0.3942 -0.3218 -0.28638 -0.25068 -0.18578
## lambda[5] -0.2388 -0.1985 -0.17522 -0.15363 -0.10877
## lambda[6] -0.3289 -0.2735 -0.24678 -0.21985 -0.16445
## lambda[7] -0.2679 -0.2183 -0.18814 -0.15709 -0.09716
## mu_lambda -0.3016 -0.2324 -0.20056 -0.16763 -0.09549
## tau        0.0500  0.0700  0.09000  0.12000  0.20000
## tau.int    5.0000  7.0000  9.00000 12.00000 20.00000
```

Results

```r
posterior_matrix <- as.matrix(jags_fit)

# cols for study-specific estimates
lambda_cols <- grep("lambda\\[", colnames(posterior_matrix), value = TRUE)

# translate log-odds by exp
or_df <- as.data.frame(exp(posterior_matrix[, lambda_cols]))
colnames(or_df) <- gsub("lambda\\[|\\]", "", lambda_cols)
```

```r
or_long <- or_df %>%
  pivot_longer(cols = everything(),
               names_to = "study",
               values_to = "OR")

or_summary <- or_long %>%
  group_by(study) %>%
  summarize(
    rr_low  = quantile(OR, 0.025),
    rr_med  = quantile(OR, 0.50),
    rr_high = quantile(OR, 0.975)
  )

ggplot(or_long, aes(x = OR, y = study, fill = study)) +
  geom_density_ridges(alpha = 0.4, scale = 1.0) +
  geom_errorbarh(data = or_summary,
                 aes(xmin = rr_low, xmax = rr_high, y = study),
                 height = 0.2, color = "black", inherit.aes = FALSE) +
  geom_point(data = or_summary,
             aes(x = rr_med, y = study),
             color = "black", size = 2, inherit.aes = FALSE) +
  labs(
    x = "Odds Ratio (exp(lambda))",
    y = "Study",
    title = "Posterior Distributions of Study-Specific Odds Ratios"
  ) +
  theme_ridges() +
  theme(legend.position = "none")
```

```
## Picking joint bandwidth of 0.00693
```



1.

95% credible intervals do not include an odds-ratio of 1.0 for any study - which represents the null hypothesis that a unit increase in CXCL14 expression has no effect on the log-odds of optimal debulking, implying that CXCL14 is a useful biomarker to predict debulking status given the data under the given model and prior assumptions. That is, each study suggests that higher CXCL14 expression consistently decreases the odds of optimal surgery.

2.

In order to perform this calculation, I will make the assumption that the study effect size in the new study is the global effect size (slope μ_λ). I will assume tht the new patient's CXCL14 expression level is 15 and will use the global intercept and slope value to compute the predicted probability of optimal debulking.

$$\text{logit}(p^*) = \alpha + \lambda_{\text{new}}\, X^*, \quad \text{with } \lambda_{\text{new}} \approx \mu_\lambda,$$

$$p^* = \frac{1}{1 + \exp\Big\{-\big(\alpha + \mu_\lambda\, X^*\big)\Big\}}.$$

```
jags_summary <- summary(jags_fit)$statistics
alpha_est <- jags_summary["alpha", "Mean"]
mu_lambda_est <- jags_summary["mu_lambda", "Mean"]
X_star <- 15
```

```r
p_star <- 1 / (1 + exp(-(alpha_est + mu_lambda_est * X_star)))
cat('the probability of optimal debulking for a patient with CXCL14 expression of 15 is', p_star)
```

```
## the probability of optimal debulking for a patient with CXCL14 expression of 15 is 0.2381993
```

3.

The following uses the same assumptions from question *2* that the new study-specific effect equals the global effect and keeps the global intercept and that the mean of the new study is normally distributed. For each simulation, parameters are drawn from a random posterior sample. Then, I fit a new logistic regression model to predict optimal debulking from the simulated XX and randomly generated binary YY data (with probability using hierarchical logistical model with specified assumptions above).

```r
n_sims <- 1000
n_patients <- 100

jags_summary <- summary(jags_fit)$statistics
alpha_post <- jags_fit[[1]][, "alpha"]
mu_lambda_post <- jags_fit[[1]][, "mu_lambda"]

X_mean <- mean(df$XX)
X_sd <- sd(df$XX)

signif_results <- numeric(n_sims)

for(i in 1:n_sims){
  idx <- sample(1:length(alpha_post), 1)
  alpha_i <- alpha_post[idx]
  mu_lambda_i <- mu_lambda_post[idx]

  X_new <- rnorm(n_patients, mean = X_mean, sd = X_sd)
  p_new <- 1 / (1 + exp(-(alpha_i + mu_lambda_i * X_new)))
  Y_new <- rbinom(n_patients, size = 1, prob = p_new)
  model_new <- glm(Y_new ~ X_new, family = binomial(link = "logit"))
  p_val <- summary(model_new)$coefficients["X_new", "Pr(>|z|)"]
  signif_results[i] <- ifelse(p_val < 0.05, 1, 0)
}

prob_signif <- mean(signif_results)
cat('Estimated probability that a new study will have p < 0.05, ', prob_signif)
```

```
## Estimated probability that a new study will have p < 0.05,  0.496
```

Discussion

Some advantages of this analysis include reliance on logit links which convey a linear relationship on the log-odds scale. Unlike the section, where we used dichotomized data and relative risk as an effect size calculation, less biological signal is lost from using a continuous independent variable as we have observed throughout the semester. The straightforward interpretation—that an increase in log-odds translates directly into a change in outcome—allows for easy comparison across studies. Additionally, the symmetry of the odds ratio with respect to optimal and suboptimal debulking enables transformation invariance, which simplifies the interpretation of the effect size across different settings

However, there are also disadvantages associated with using the odds ratio as an effect size. When outcomes

are common, the odds ratio can exaggerate the effect size relative to the relative risk, which can lead to misinterpretation if not properly adjusted for baseline risk. Unlike relative risk, which directly indicates that patients are, for example, "RR times as likely" to experience optimal debulking, the odds ratio does not convey a direct change in the absolute probability of the outcome. This means that additional steps are needed to translate the odds ratio into a more intuitive measure, making it less straightforward for those without a statistical background. For instance, the calculation above for question 2 to arrive at the probability of optimal debulking for a patient with CXCL14 expression of 15 (0.2410442) is much more nuanced than a hypothetical relative risk calculation. Compared with the above code in question 2, the probability of optimal debulking for a patient with CXCL14 expression of 15 could be calculated by

```
predicted_risk <- baseline_risk * (RR^15)
```

Lastly, some major disadvantages of this analysis include a meta-analysis of only one biomarker. Regardless of the effect size used, recent methods like Top Scoring Pairs or mas-o-menos as we have discussed in class are simple methods with high statistical power and clinical implication primarily because of their ability to observe multiple biomarkers. Moreover, these methods are comparable with more complex ML models without using extensive compute power. Question *2* above shows that even for the extremely high gene expression value of 15, which is negatively correlated with optimal debulking both based on posterior output and density of expression in the *EDA* section, the logistic regression model cannot discriminate between optimal and suboptimal debulking with a high degree of certainty. Thus, although the results suggest that CXCL14 is a useful biomarker, this meta-analysis of seven studies is not succinct enough to inform clinical decisions without analyzing CXCL14 in conjunction with other biomarkers that may increase discrimination performance.

Bibliography

Generative AI was used to convert written section notes and advice from Iris into Latex in order to give a mathematical explanation of the logic behind partial pooling, prior, and parameter choices and clean up data visualization techniques to best visualize study-level variation. In addition, Generative AI was used to debug code and primarily Latex compilation issues and edit discussion for grammar.