DS 6040 Bayesian Machine Learning
Prof. Donald E. Brown
December 12th, 2021
Hannah Schmuckler (mmc4cv), Karolina Naranjo-Velasco(kn3cs), Aparna Marathe (am7ad), Beza Delelegn (bmd5bc)

# Modeling Uncertainty in Diabetes Incidence by Race in the United States

## I. Abstract:

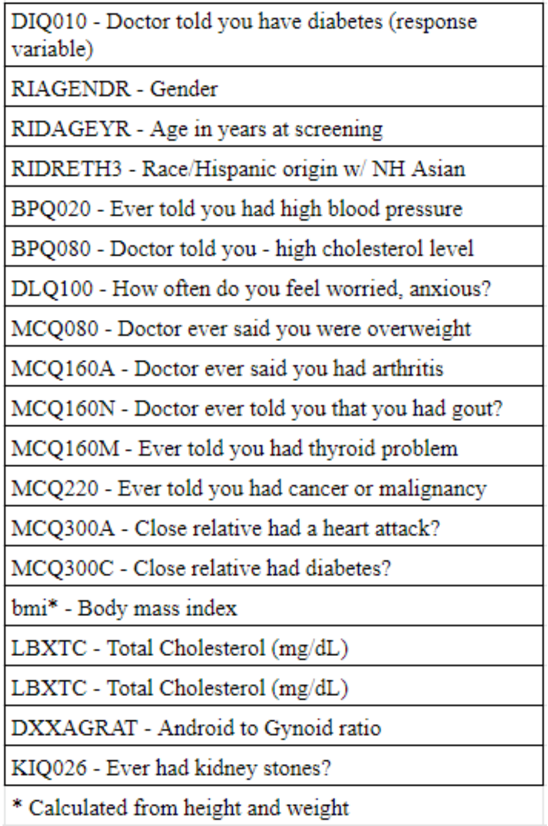
This paper discusses Bayesian machine learning methods to model uncertainty in diabetes prediction. Three different Bayesian regression models were developed using various medical features from patient surveys to predict the incidence of diabetes. Of the 34.2 million adults with diabetes, 7.3 million were undiagnosed as of 2018[1]. Moreover, diabetes disproportionately affects racial/ethnicity minority populations. When compared to white adults, the risk of having diabetes is 77% higher among African Americans, 66% higher among Latinos/Hispanics, and 18% higher among Asian Americans[2]. Three models using 6 variables were developed to examine the possibility of using prior health data in predicting the incidence of diabetes. Since race/ethnicity is believed to have a significant effect on diabetes outcomes for individuals, a partially-pooled and hierarchical model were developed to examine these differences, in addition to a pooled model to examine the overall population.

## II. Methodology:

### A. *Data Collection and Processing*

The data used for these models was collected from the National Health and Nutrition Examination Survey (NHANES)[3]. This survey includes information about whether a patient was diagnosed with diabetes, as well as other physical health statistics. It also includes information about mental health, which may be linked to the presence of diabetes. For this analysis, 2015-2016 data were used. Based on initial research, the variables in Figure 1 were selected for further consideration.

Figure 1



| Figure 1 |
|---|
| DIQ010 - Doctor told you have diabetes (response variable) |
| RIAGENDR - Gender |
| RIDAGEYR - Age in years at screening |
| RIDRETH3 - Race/Hispanic origin w/ NH Asian |
| BPQ020 - Ever told you had high blood pressure |
| BPQ080 - Doctor told you - high cholesterol level |
| DLQ100 - How often do you feel worried, anxious? |
| MCQ080 - Doctor ever said you were overweight |
| MCQ160A - Doctor ever said you had arthritis |
| MCQ160N - Doctor ever told you that you had gout? |
| MCQ160M - Ever told you had thyroid problem |
| MCQ220 - Ever told you had cancer or malignancy |
| MCQ300A - Close relative had a heart attack? |
| MCQ300C - Close relative had diabetes? |
| bmi* - Body mass index |
| LBXTC - Total Cholesterol (mg/dL) |
| LBXTC - Total Cholesterol (mg/dL) |
| DXXAGRAT - Android to Gynoid ratio |
| KIQ026 - Ever had kidney stones? |
| * Calculated from height and weight |

The NHANES data had a large number of missing values, so after selecting variables, all cases with missing values were dropped. The quantitative variables were standardized using a z-score, and binary categorical variables were recoded.. DIQ010, the response variable, was reduced from three categories to two, grouping borderline individuals with those who have diabetes so that logistic regression could be performed. After these processing steps, the cleaned data set contained 2,392 observations. Finally, the data set was split into an 80% training and 20% test set.

### *B. Feature Selection with WAIC*

Based on the features listed above, a correlation matrix was created to evaluate multicollinearity and to inform decisions on predictors to drop in our reduced models (Appendix A, Figure 2.).

Highest Correlations:
(1) bmi and WHD020 have the highest correlation = 0.87.
(2) DLQ140 and DLQ100 are highly correlated (0.6).
(3) bmi and MCQ080 are moderately correlated (0.57).
(4) WHD020 and MCQ080 are also moderately correlated (0.5).
(5) DXXAGRAT and RIAGENDR are moderately correlated (0.49).
(6) DXXAGRAT and WHD020 are moderately correlated (0.49).

From here, one model with all predictors and four reduced models were created. and a WAIC analysis was performed (Appendix B, figure 5). Each model was a pooled model where Sigma (HC~(5)), β0 (N~(0, 1)), β_uninformative (N~(0, 100)), and β_informative (N~(1, 2)) each fed into a logistic y_hat. Figure 3 in appendix C shows this relationship, though the number of predictors on each plate varies by model. BPQ080, MCQ160M, and MCQ160N were informative priors, and all other predictor variables were uninformative priors.

To compare the 5 models and prevent overfitting, a WAIC analysis was performed (Appendix B, Figure 5). Reduced model 4, with 6 predictors, performed the best, with a WAIC score of 125.68, so those predictors will be used to create a partially pooled and hierarchical model as well. The final predictors chosen include 4 using uninformative priors (BPQ020, RIDAGEYR, MCQ300C, and bmi) and two using informative priors (BPQ080 and MCQ160N).

### *C. Models*

#### *a. Pooled Model*

Reduced model 4, described above, is a pooled model. The mathematical visualization of this pooled model is displayed in the Bayesian network in Appendix C, Figure 6. The uninformative priors, with a distribution of N~(0, 100), are on a plate with dimension 4, and the informative priors (N~(1, 2)) are on a plate with dimension 2. The intercept (β0) also has an informative prior of N~(0, 1), as it is likely to be close to 0. The model variance sigma is represented by a Half Cauchy with beta 5. Together, these parameters act as priors on our likelihood, resulting in a logistic posterior. The resulting trace plots and forest plots from the model are shown below the graphical display in Appendix C, figures 7 & 8. Looking at the trace plots, there is good convergence with a tight peak for two out of four of the coefficients on the uninformed predictors: age and bmi. There is also good convergence with a tight peak for one out of two of coefficients on the informed predictors: high blood pressure. The coefficient on gout has a high variance.

Since pooled models cannot provide estimates for groups in data, it is imperative to create a model that can discern any differences that exist between racial groups. Partially pooled and hierarchical models should be able to incorporate the differences between groups in the model.

#### *b. Partially-Pooled Model*

The graphical representation of the partially-pooled model is shown in Appendix D, Figure 10, along with the resulting trace plots (Figure 11) and forest plots (Figure 12). These plots look nearly identical to those of the complete-pooled model. Age, bmi, and high blood pressure show the best convergence. Gout has the highest variance.

This model uses the same parameters and priors as the pooled model, with the addition of a variable for race/ethnicity (RIDRETH3), which interacts with the intercept. This creates a different intercept for each model, while the rest of the regression equation remains the same. Groups with less observations will carry less weight and their average will shrink toward the overall average, as compared with an unpooled model with the same parameters. Groups with bigger weights will carry more weight and, as a result, be closer to the unpooled mean estimates. By using partial pooling, small groups may be estimated better than an unpooled model.

### *c. Hierarchical Model*

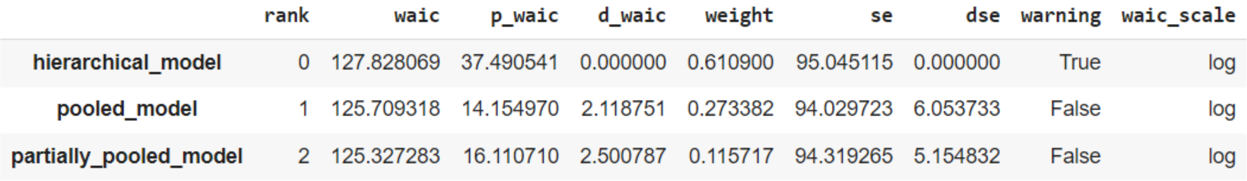
Hierarchical models use all the information in the data to better estimate group target or response variables even with groups of small sample size. Since a hierarchical model is a special case of a multilevel model, prior distributions were placed on the hyperparameters of the prior distributions of the base model. The same variables as before were used as predictors.

The graphical representation of the hierarchical model is shown in Appendix E, Figure 14. Note how the hierarchical priors feed into the priors of the base model. The model error, sigma_y, is HC~(5). The informed parameters have a prior on mu of N~(1,2) and a prior on sigma of HC~(10e5), while the parameters with uninformed priors and the intercept have a prior on their mu of N~(0, 10e5), and on sigma of HC~(10e5). All parameters and the intercept are interacted with race. The resulting trace plots and forest plots are shown in Appendix E, Figures 15 and 16. Again note that the best convergence in the coefficients on the predictors: age, bmi and high blood pressure. While there were some divergences while the trace was running, the traces look fine.

## III. Results and Conclusions:

From the plots, the most influential predictors seem to be similar for all of the models. To further analyze which model performs the best, a second WAIC comparison table was computed.

Figure 2: WAIC Comparison Table



| | rank | waic | p_waic | d_waic | weight | se | dse | warning | waic_scale |
|---|---|---|---|---|---|---|---|---|---|
| **hierarchical_model** | 0 | 127.828069 | 37.490541 | 0.000000 | 0.610900 | 95.045115 | 0.000000 | True | log |
| **pooled_model** | 1 | 125.709318 | 14.154970 | 2.118751 | 0.273382 | 94.029723 | 6.053733 | False | log |
| **partially_pooled_model** | 2 | 125.327283 | 16.110710 | 2.500787 | 0.115717 | 94.319265 | 5.154832 | False | log |

Looking at the above WAIC table, the hierarchical model ranks the best, though all three are well within standard error of each other. Figure 18 in Appendix F illustrates the similarity of all three models: it plots all three model's nearly indistinguishable posteriors for one individual.

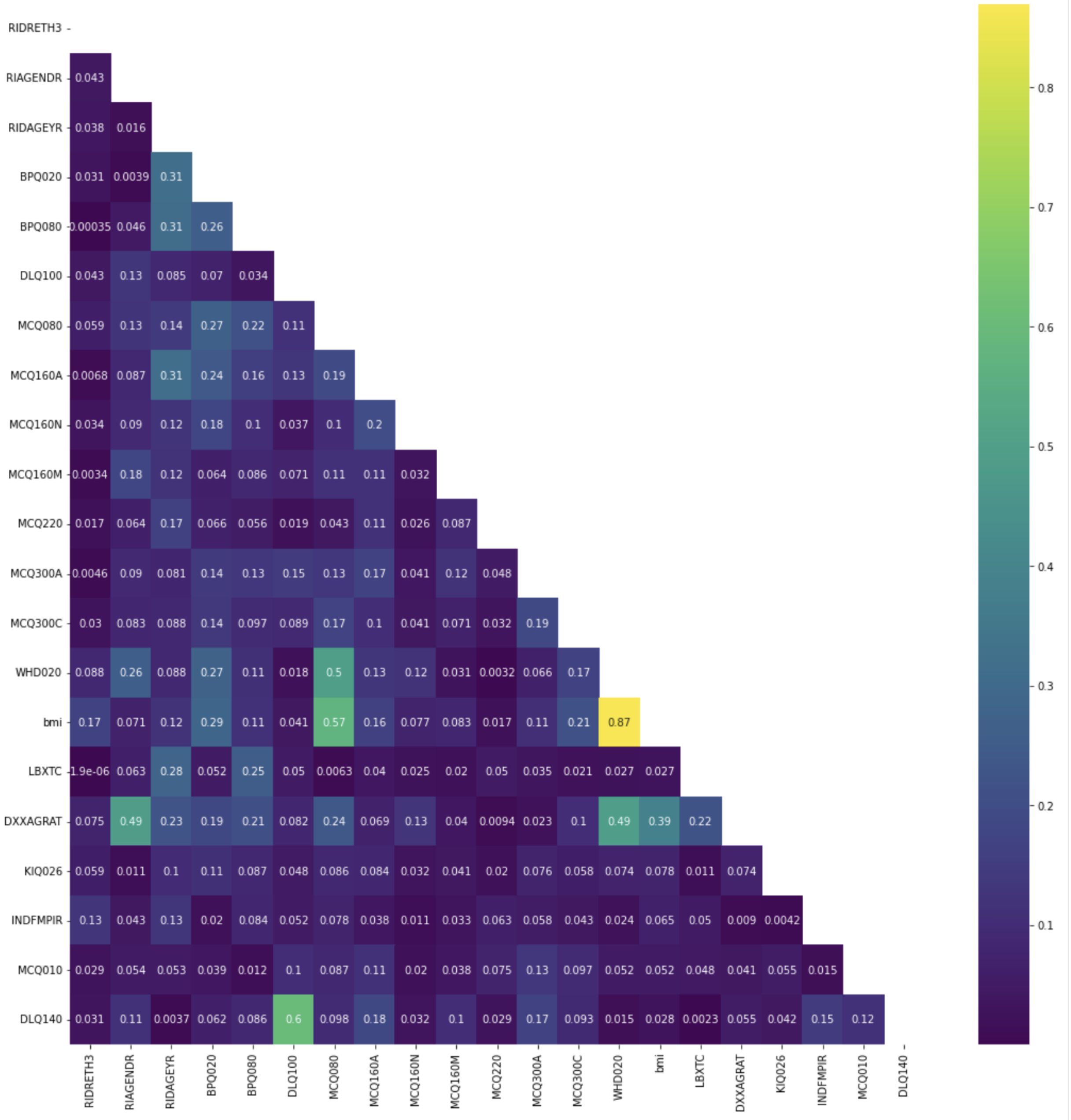
Posterior predictions for hold-out data were generated, 1000 samples for each individual in the test set. For interpretability, these logits were converted to probabilities and plotted. Appendix C, Figure 9 shows the results of the pooled model. Though it appears that test set diabetics do, on average, have higher average posterior classification, the two groups are well within each other's credible intervals (Appendix G, Figure 19). The results of the partially pooled model by race are shown in Appendix D,

Figure 13. Though this graph shows that black diabetics are classed slightly higher than other groups, all posterior means are well within each other's credible intervals (Appendix G, Figure 20). Appendix E, Figure 17 shows the same plot for the hierarchical model, with the same results. Perhaps the racial difference in diabetes diagnoses was not reflected in the original data.

For classification purposes, the mean of each individual's posterior was found and assigned as their final probability of being diabetic, and confusion matrices were created (Appendix H, Figure 22). For all models, a threshold of 0.5 was used. The pooled model had a false negative rate of 1.92% and false negative rate of 68.15%. The partially pooled model had a false negative rate of 1.92% and a false negative rate of 67.68%. The hierarchical model had a 0% false negative rate and a 69.56% false positive rate. A model with performance like this would be useful for identifying individuals who are at high risk of diabetes.

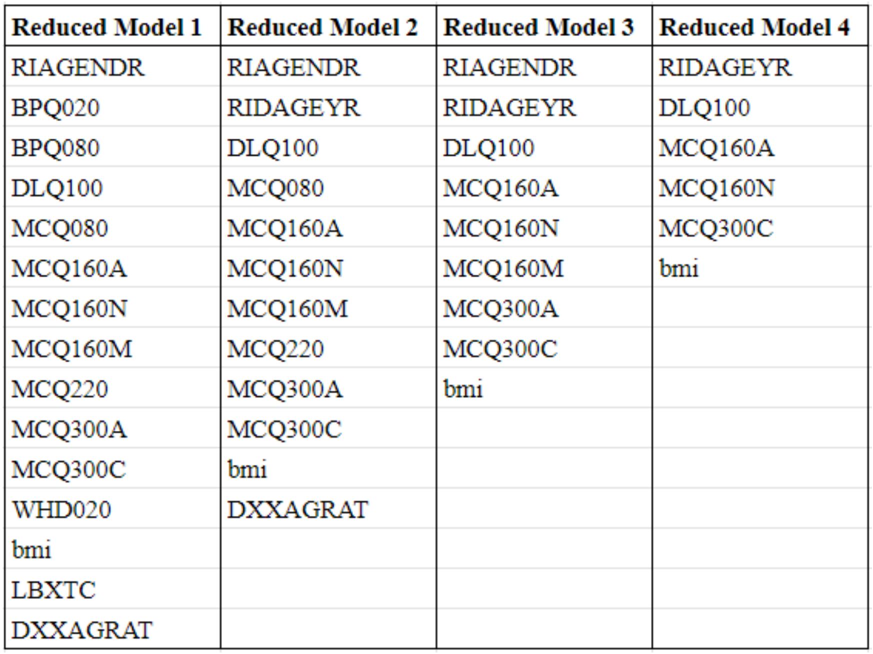
**Appendix A**

*Figure 1*.Correlation Matrix



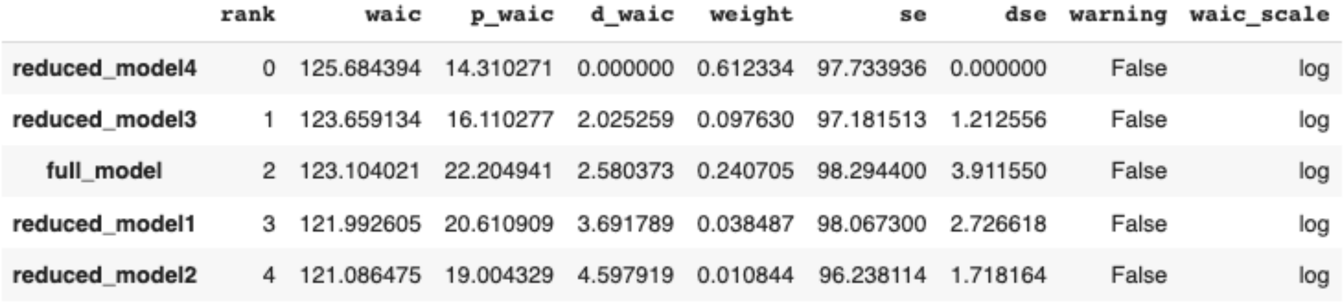
**Appendix B**

*Figure 4*. Model Breakdown



| Reduced Model 1 | Reduced Model 2 | Reduced Model 3 | Reduced Model 4 |
|---|---|---|---|
| RIAGENDR | RIAGENDR | RIAGENDR | RIDAGEYR |
| BPQ020 | RIDAGEYR | RIDAGEYR | DLQ100 |
| BPQ080 | DLQ100 | DLQ100 | MCQ160A |
| DLQ100 | MCQ080 | MCQ160A | MCQ160N |
| MCQ080 | MCQ160A | MCQ160N | MCQ300C |
| MCQ160A | MCQ160N | MCQ160M | bmi |
| MCQ160N | MCQ160M | MCQ300A | |
| MCQ160M | MCQ220 | MCQ300C | |
| MCQ220 | MCQ300A | bmi | |
| MCQ300A | MCQ300C | | |
| MCQ300C | bmi | | |
| WHD020 | DXXAGRAT | | |
| bmi | | | |
| LBXTC | | | |
| DXXAGRAT | | | |

Figure 5: WAIC Analysis for Feature Selection



| | rank | waic | p_waic | d_waic | weight | se | dse | warning | waic_scale |
|---|---|---|---|---|---|---|---|---|---|
| **reduced_model4** | 0 | 125.684394 | 14.310271 | 0.000000 | 0.612334 | 97.733936 | 0.000000 | False | log |
| **reduced_model3** | 1 | 123.659134 | 16.110277 | 2.025259 | 0.097630 | 97.181513 | 1.212556 | False | log |
| **full_model** | 2 | 123.104021 | 22.204941 | 2.580373 | 0.240705 | 98.294400 | 3.911550 | False | log |
| **reduced_model1** | 3 | 121.992605 | 20.610909 | 3.691789 | 0.038487 | 98.067300 | 2.726618 | False | log |
| **reduced_model2** | 4 | 121.086475 | 19.004329 | 4.597919 | 0.010844 | 96.238114 | 1.718164 | False | log |

## Appendix C

*Figure 6:* Bayesian Graphical Visualization for Pooled Model

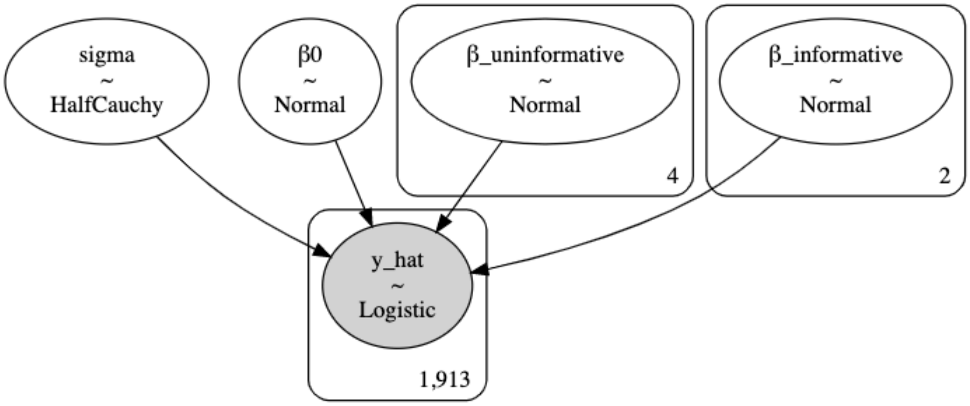


*Figure 7*: Trace Plots for Pooled Model

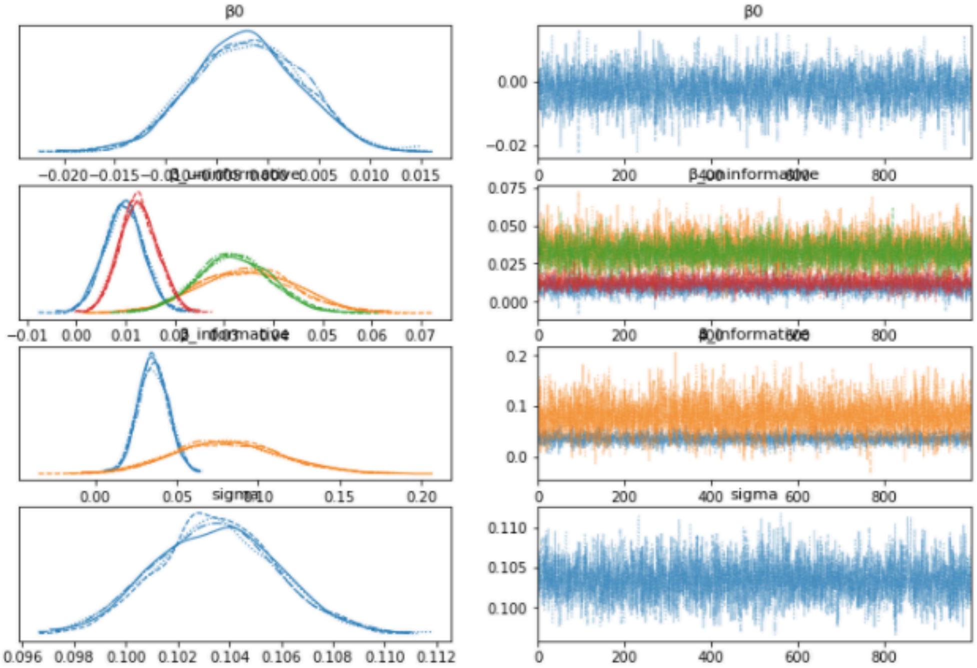


*Figure 8:* Forest Plots for Pooled Model

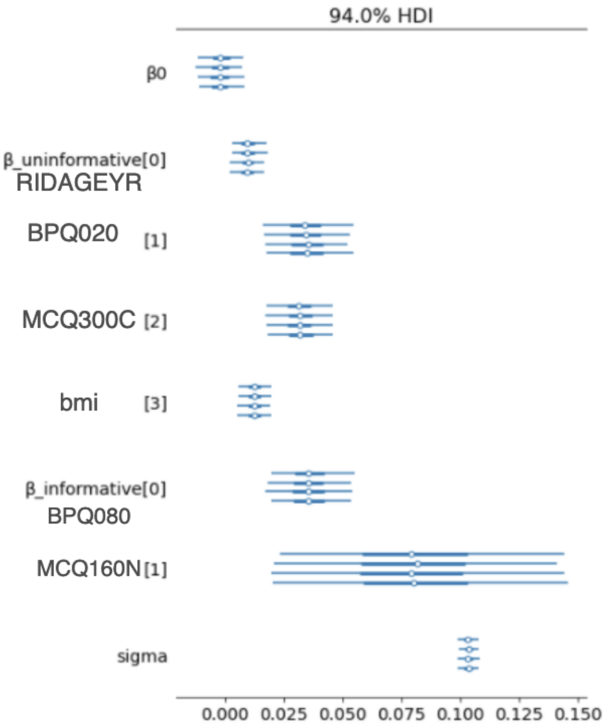
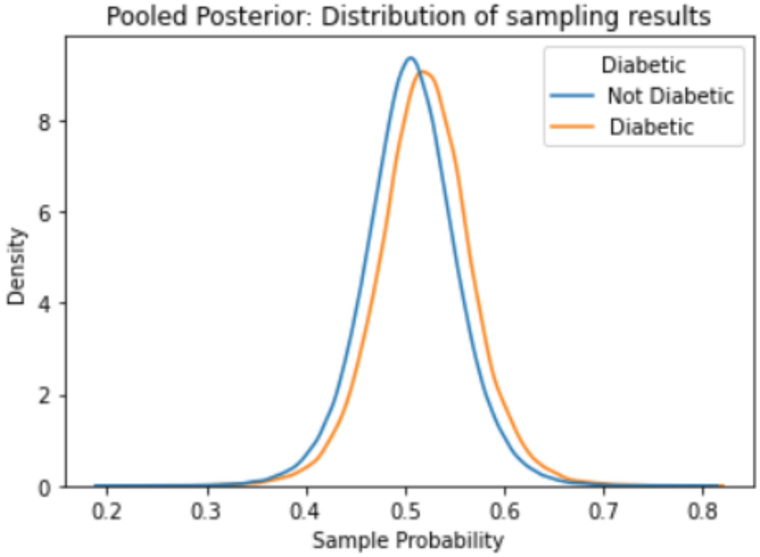


*Figure 9*: Posterior Distribution of sampling results for pooled model



**Appendix D**

*Figure 10:* Bayesian Graphical Visualization for Partially-Pooled Model

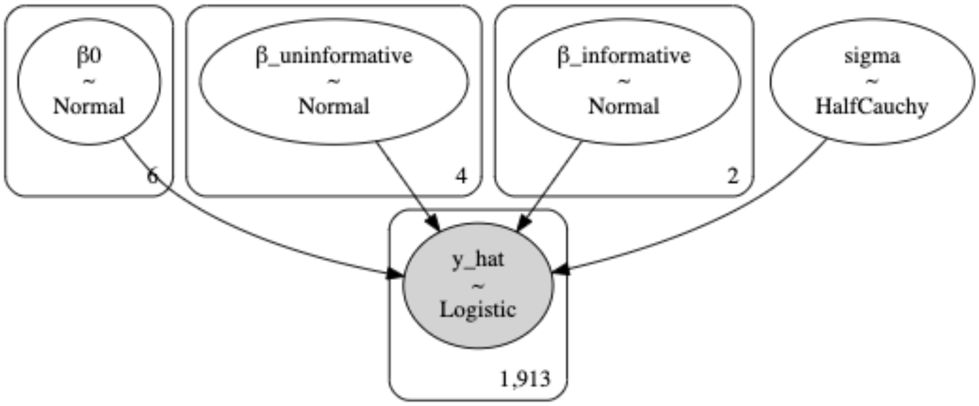


*Figure 11:* Forest Plot for partially pooled model

B0 is on a plate.
Key:
0: Mexican American
1: Non-Hispanic Black
2: Other Hispanic
3: Non-Hispanic White
4: Non-Hispanic Asian
5: Other Race - Including Multi-Racial

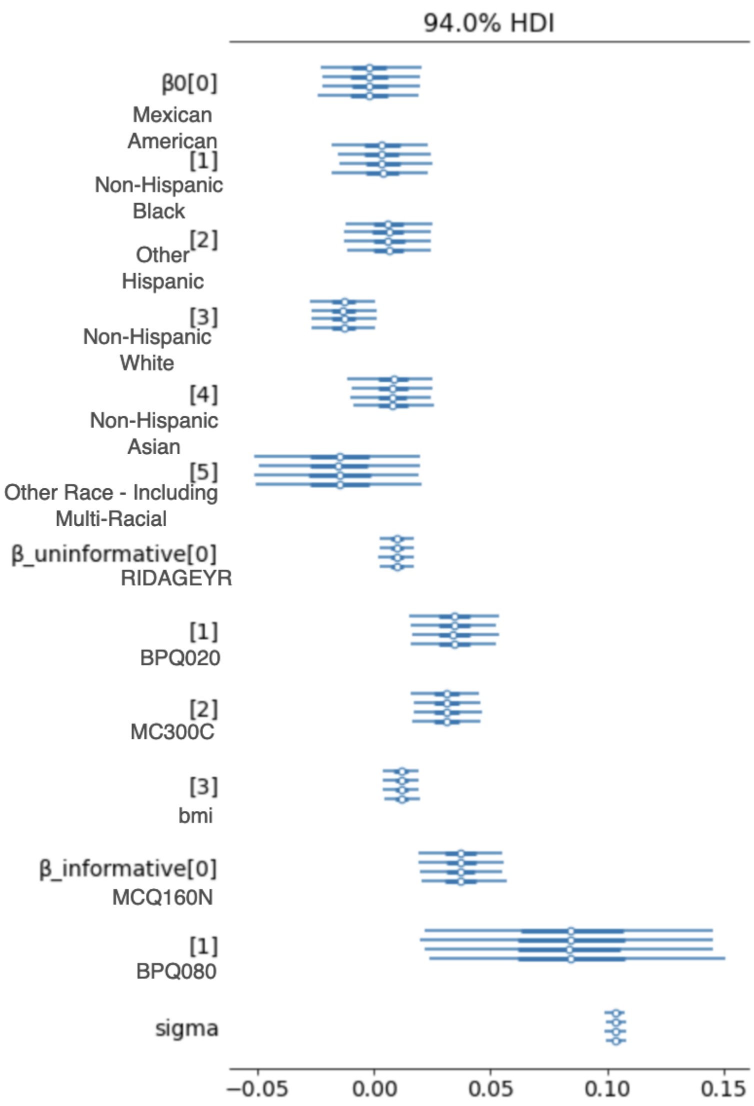
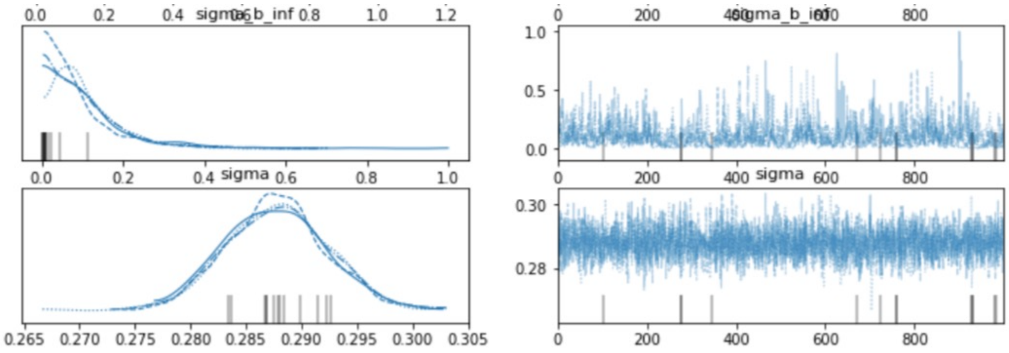


*Figure 12*: Trace Plots for partially pooled model



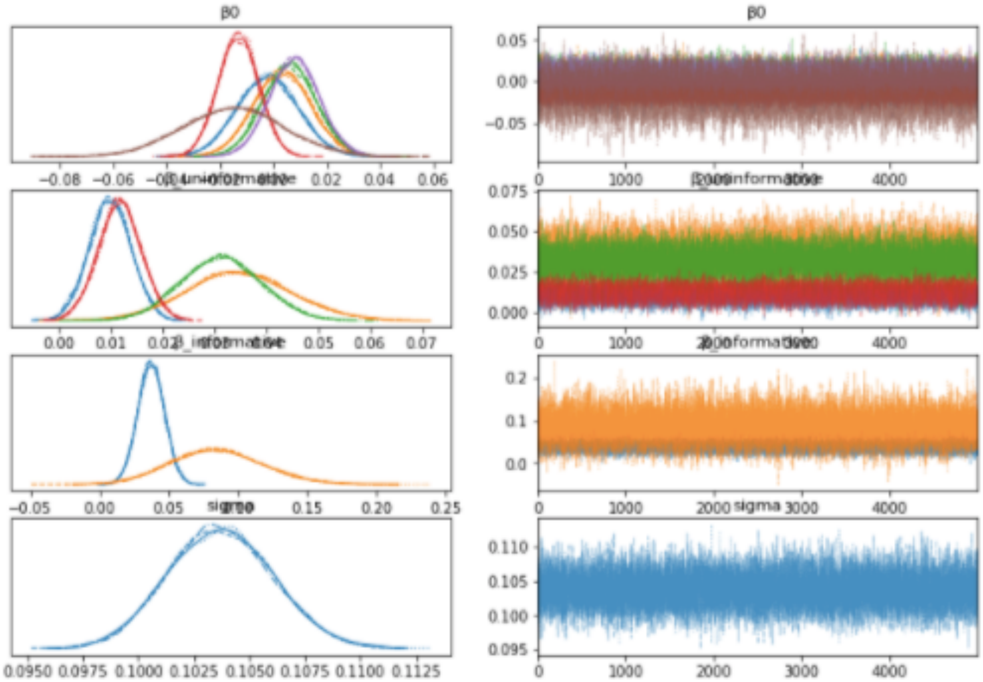
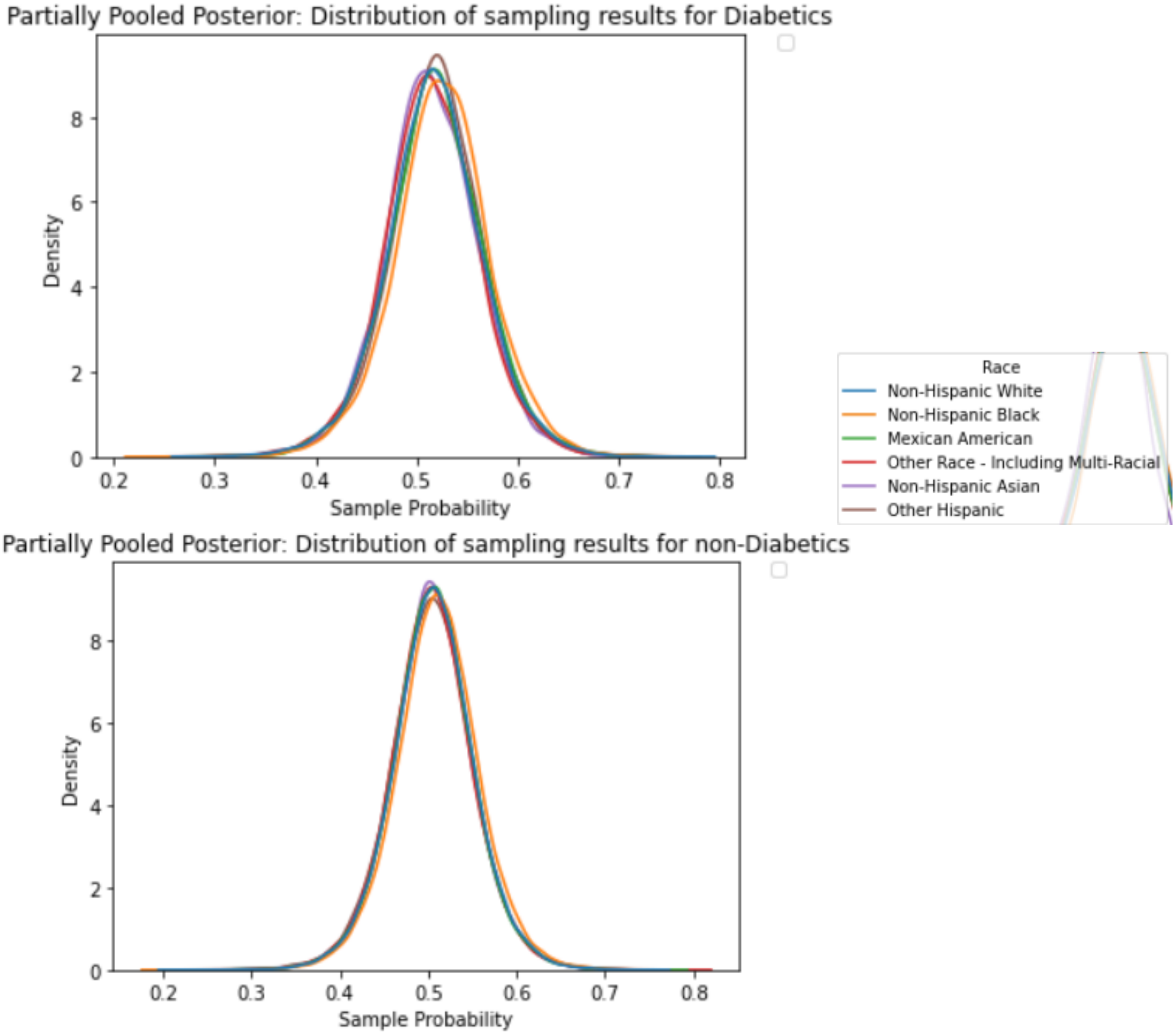

β0
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
β0
0.05
0.00
−0.05
0
1000
2000
3000
4000
β_uninformative
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
β_uninformative
0.075
0.050
0.025
0.000
β_informative
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
β_informative
0.2
0.1
0.0
sigma
0.0950
0.0975
0.1000
0.1025
0.1050
0.1075
0.1100
0.1125
sigma
0.110
0.105
0.100
0.095

*Figure 13:* Plots of partially pooled posterior by race & diabetes status



## Appendix E

*Figure 14*: Bayesian Graphical Visualization for Hierarchical Model

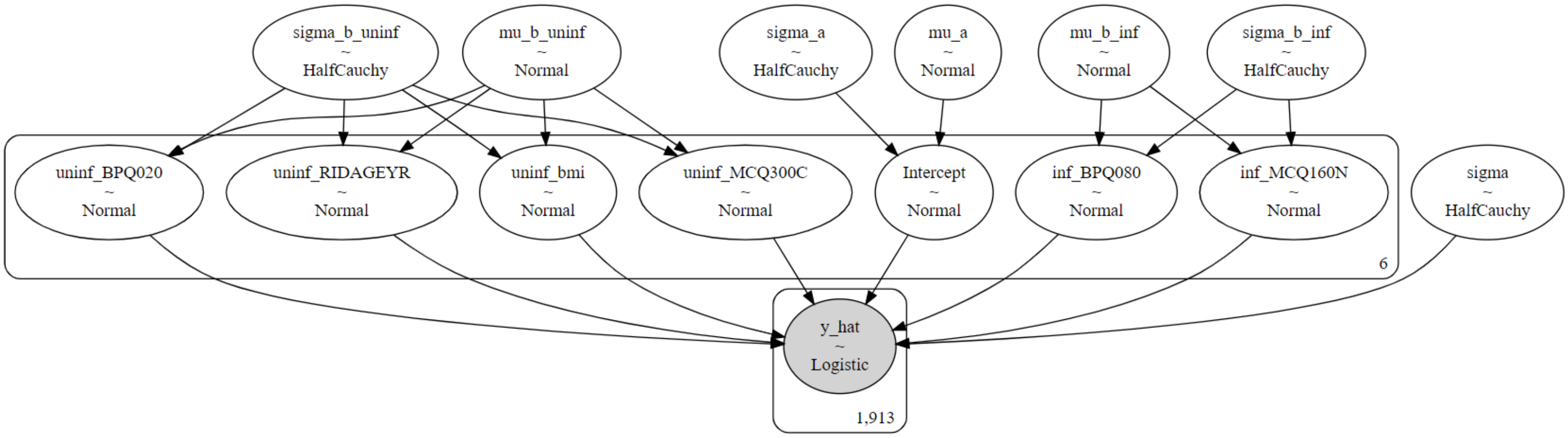
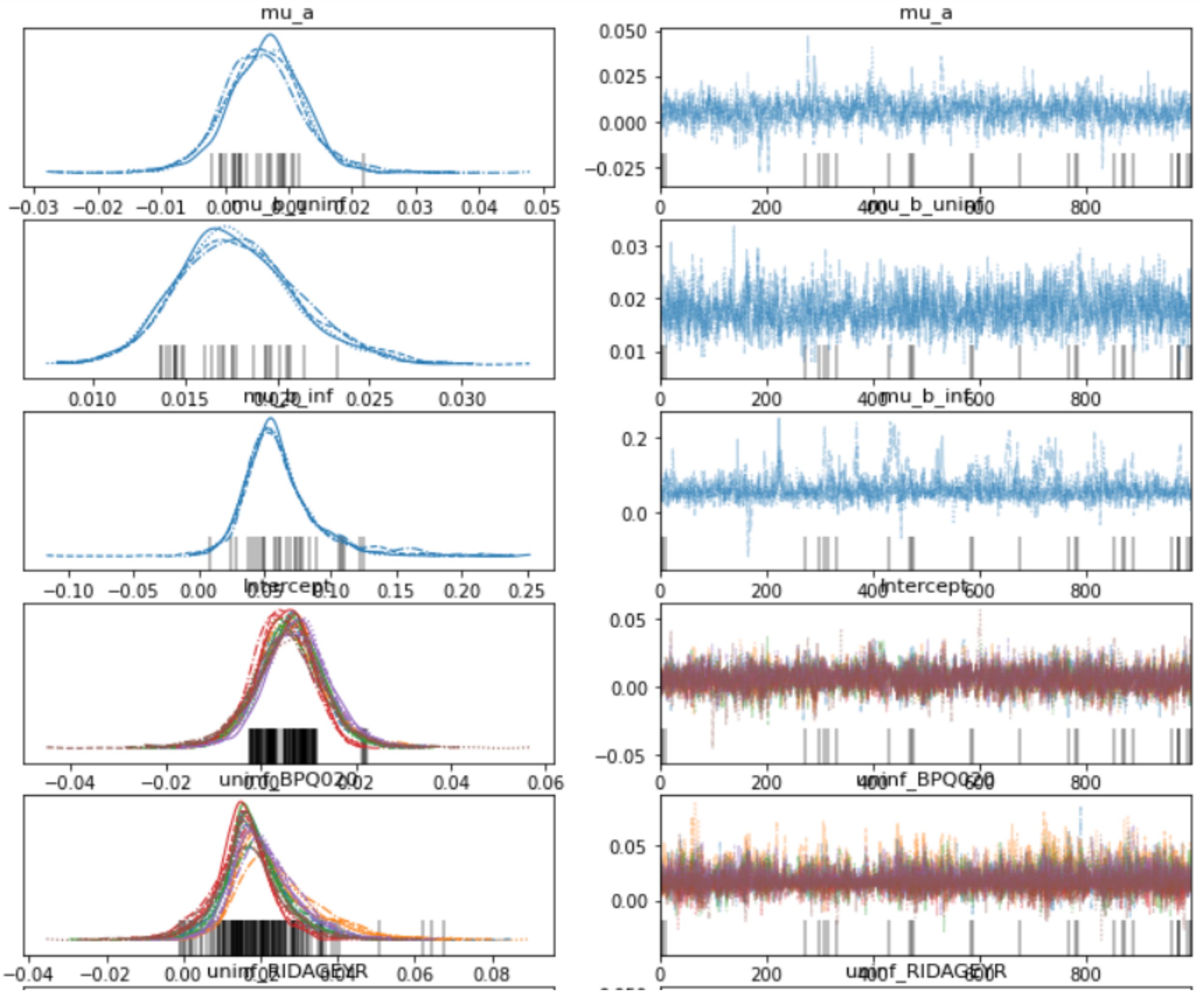
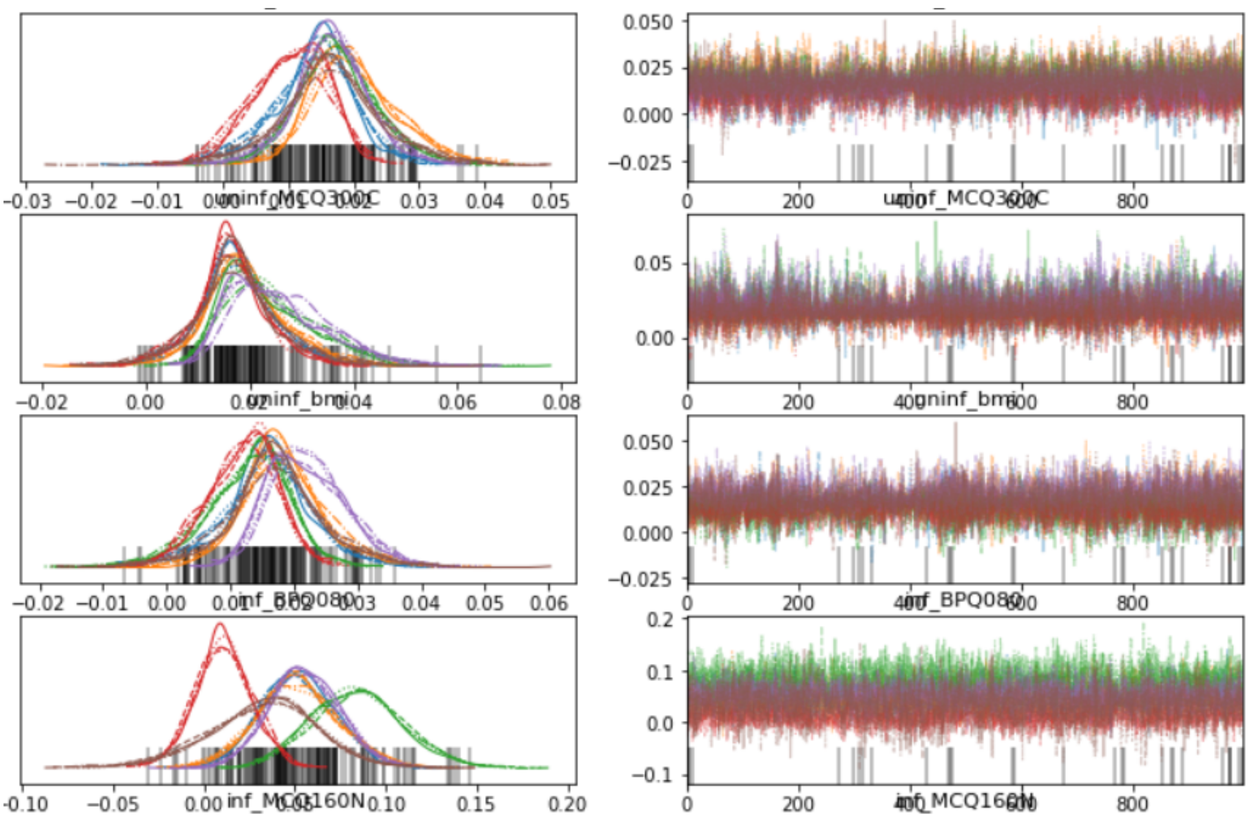


*Figure 15:* Trace Plots for Hierarchical Model




uninf_MCQ300C
uninf_bmi
inf_BPQ080
inf_MCQ160N
−0.03
−0.02
−0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.08
−0.10
−0.05
0.10
0.15
0.20
0.050
0.025
0.000
−0.025
0.05
0.00
0.2
0.1
0.0
−0.1
0
200
400
600
800

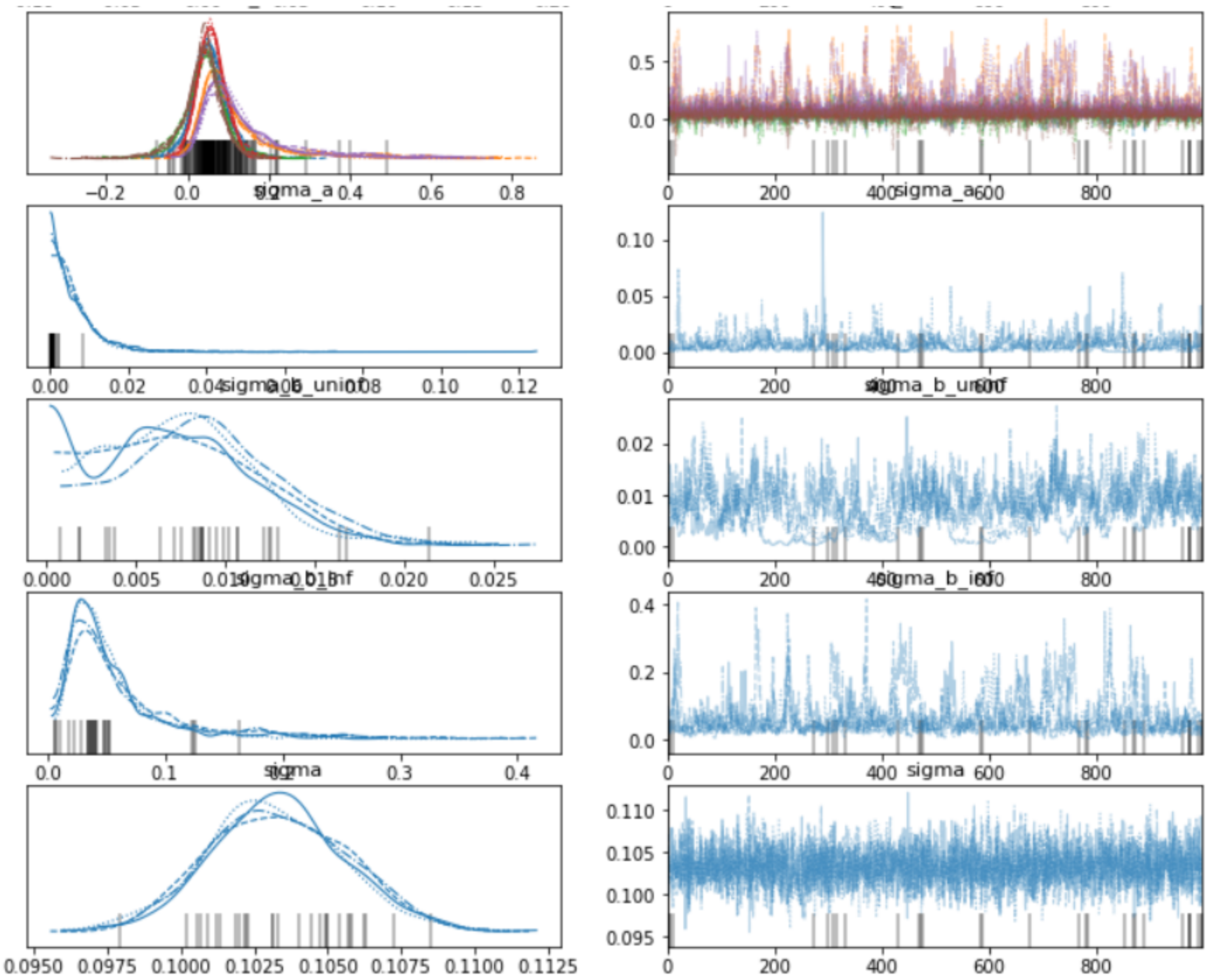
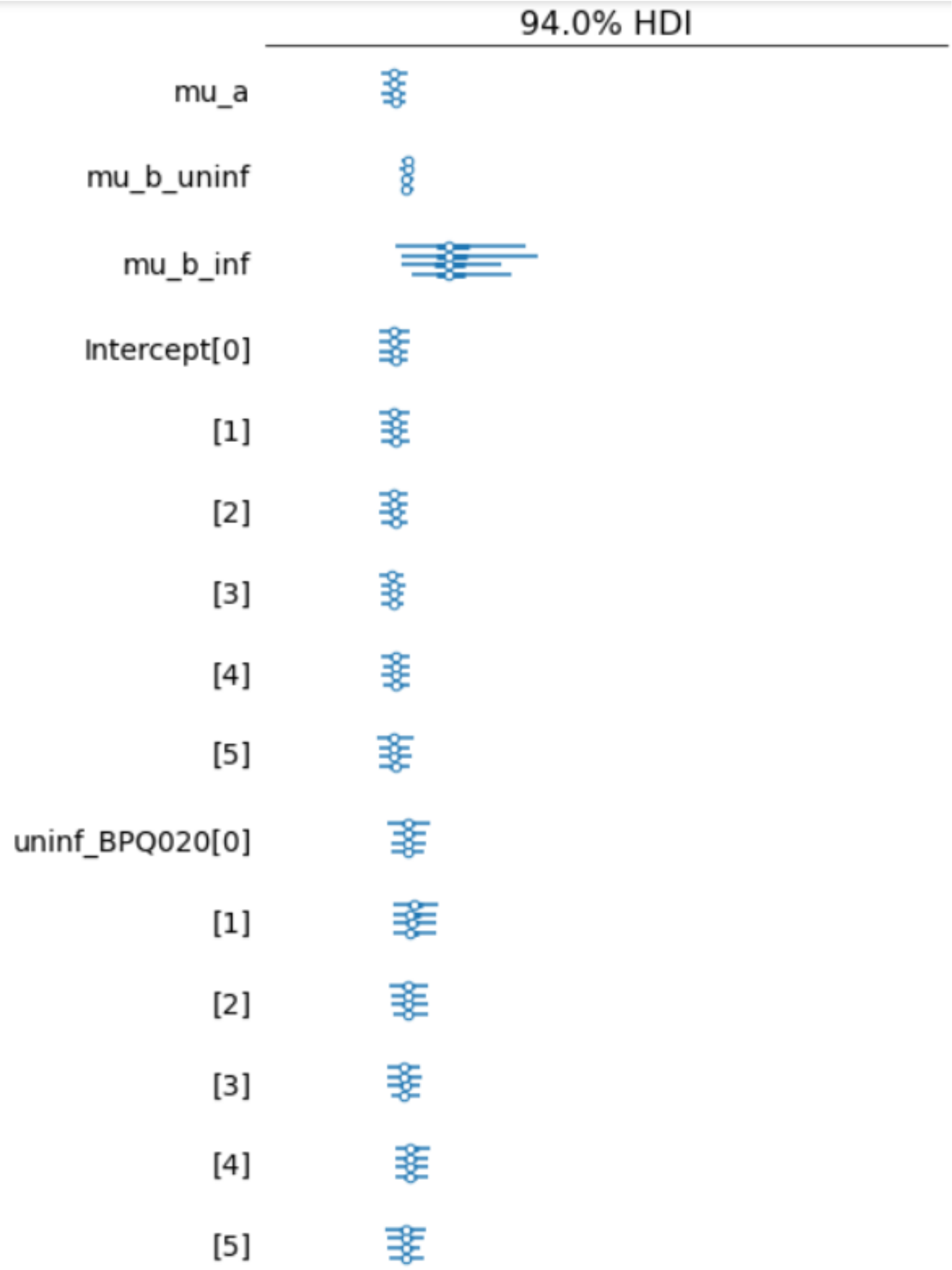
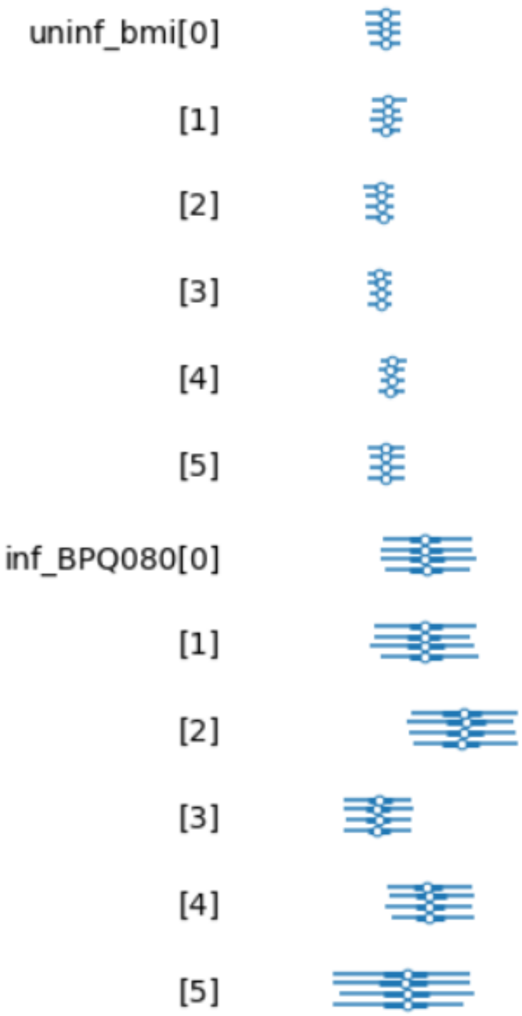


*Figure 16:* Forest Plots for Hierarchical Model

Each predictor and the intercept are split by race.
Key:
0: Mexican American
1: Non-Hispanic Black
2: Other Hispanic
3: Non-Hispanic White
4: Non-Hispanic Asian
5: Other Race - Including Multi-Racial


94.0% HDI
mu_a
mu_b_uninf
mu_b_inf
Intercept[0]
[1]
[2]
[3]
[4]
[5]
uninf_BPQ020[0]
[1]
[2]
[3]
[4]
[5]

uninf_RIDAGEYR[0]
[1]
[2]
[3]
[4]
[5]
uninf_MCQ300C[0]
[1]
[2]
[3]
[4]
[5]


uninf_bmi[0]
[1]
[2]
[3]
[4]
[5]
inf_BPQ080[0]
[1]
[2]
[3]
[4]
[5]

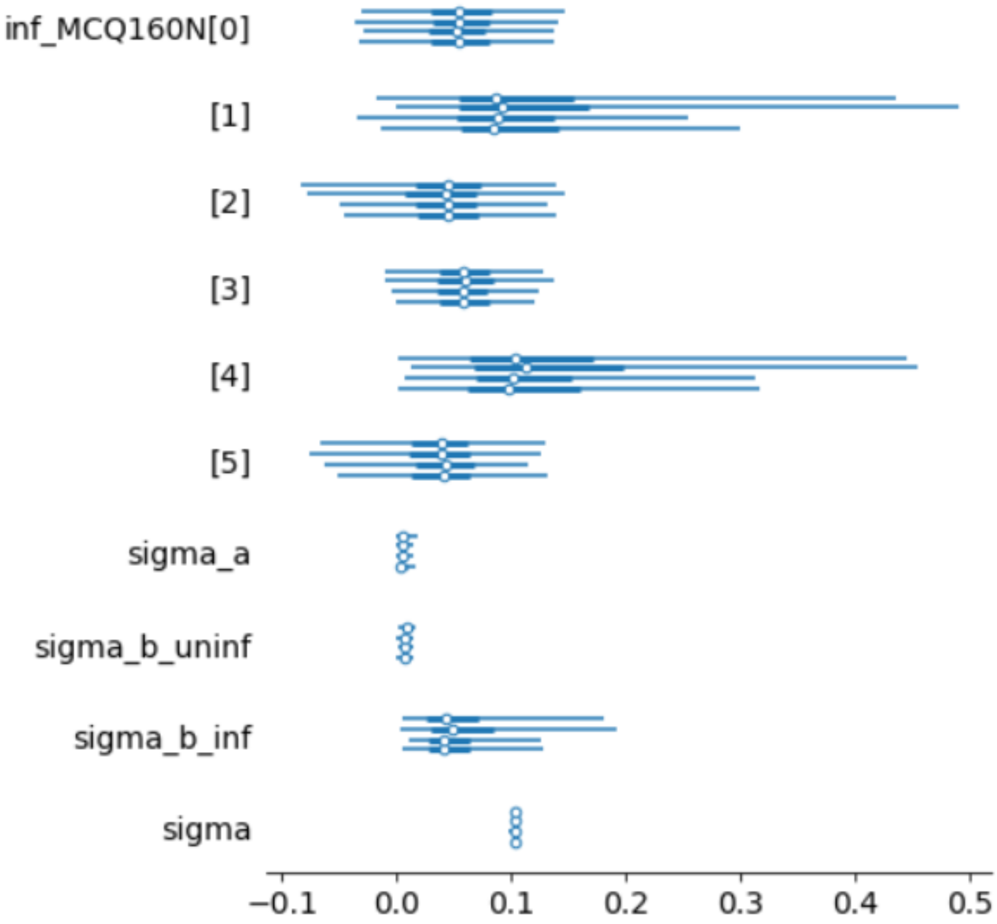
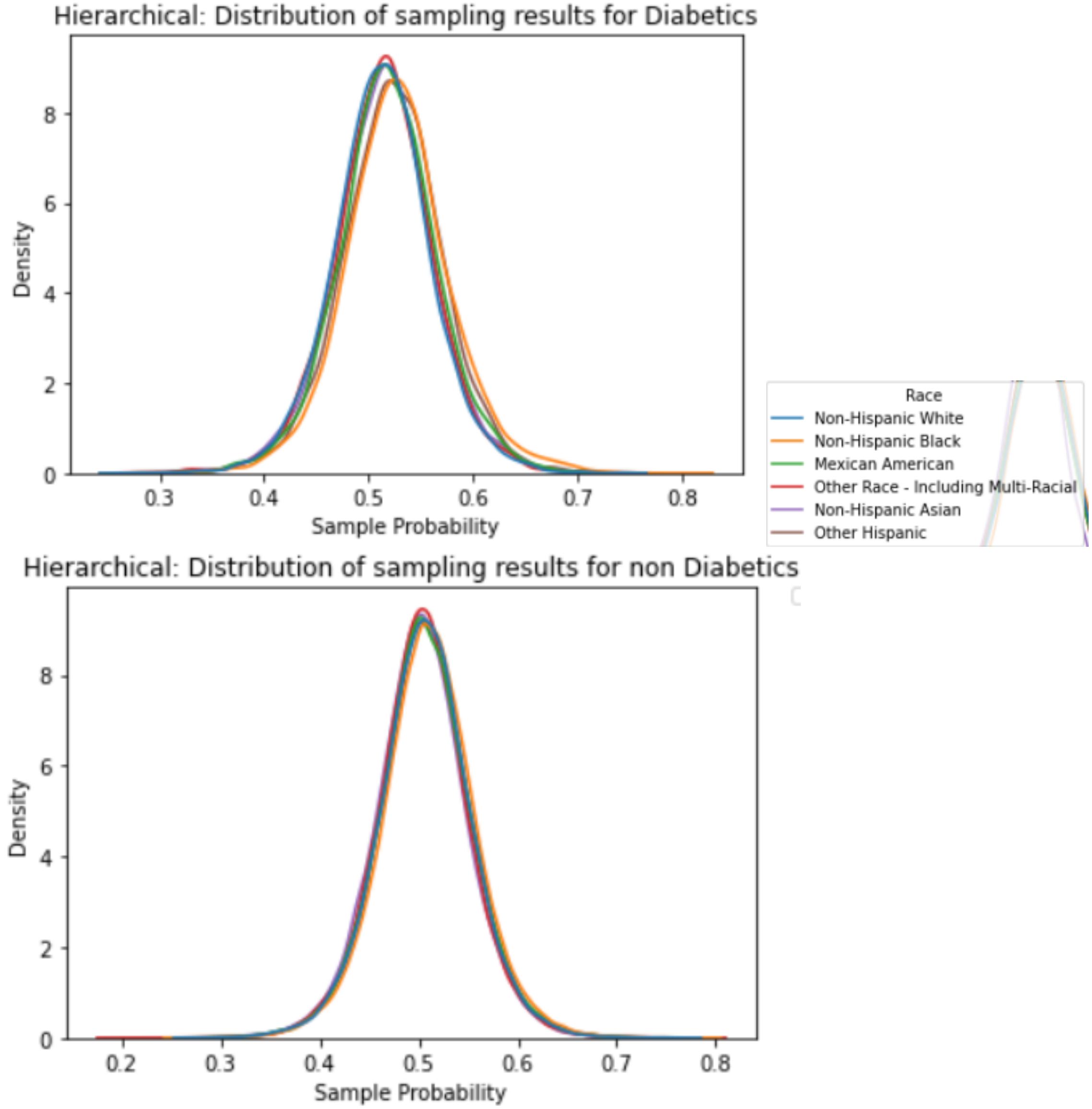
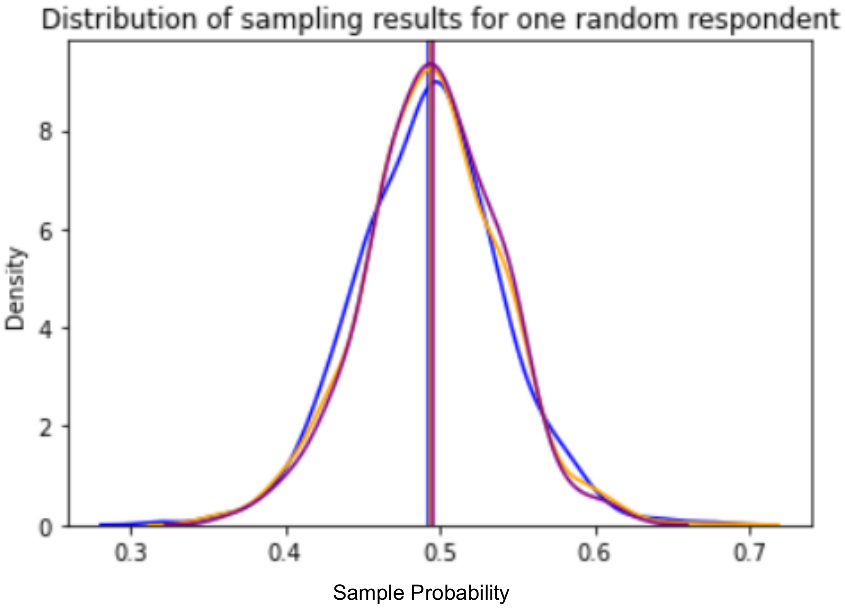

inf_MCQ160N[0]
[1]
[2]
[3]
[4]
[5]
sigma_a
sigma_b_uninf
sigma_b_inf
sigma
−0.1
0.0
0.1
0.2
0.3
0.4
0.5

*Figure 17:* Plots of hierarchical posterior by race & diabetes status



**Appendix F:**

*Figure 18 :* Density of posterior results for pooled, partially pooled, and hierarchical models for one random individual



**Appendix G:**

*Figure 19:* Pooled Posterior Credible Intervals

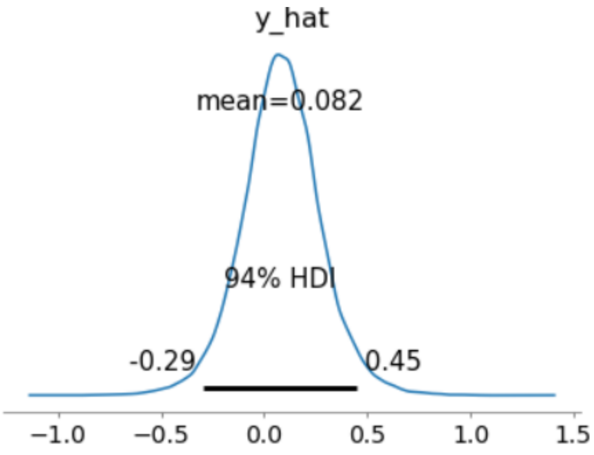
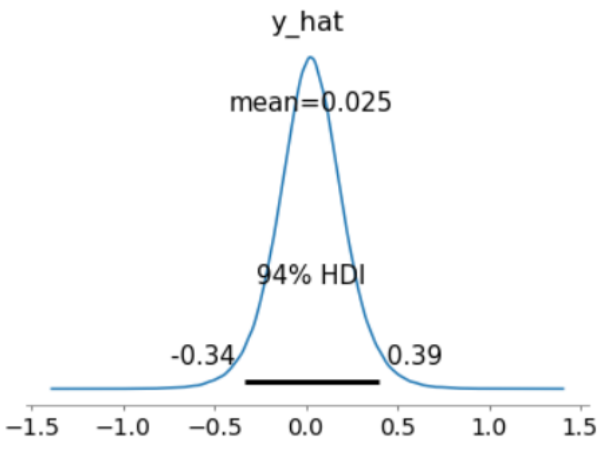
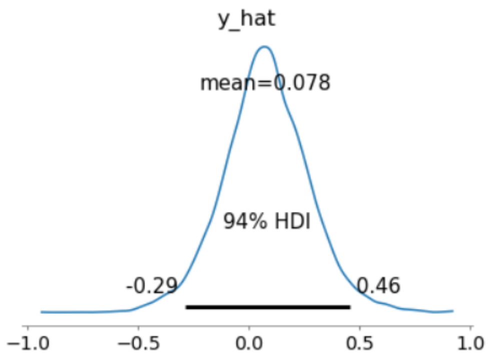
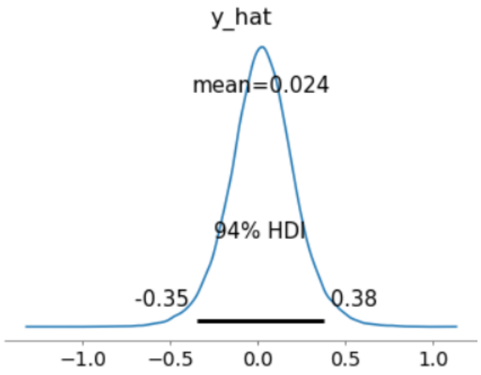
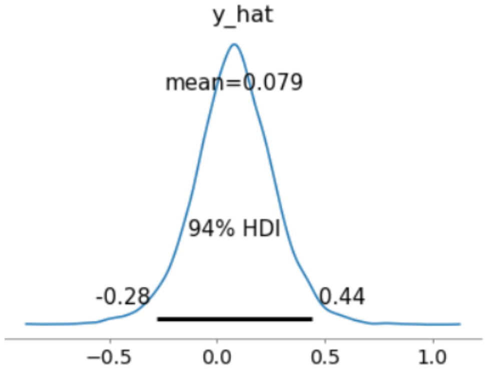
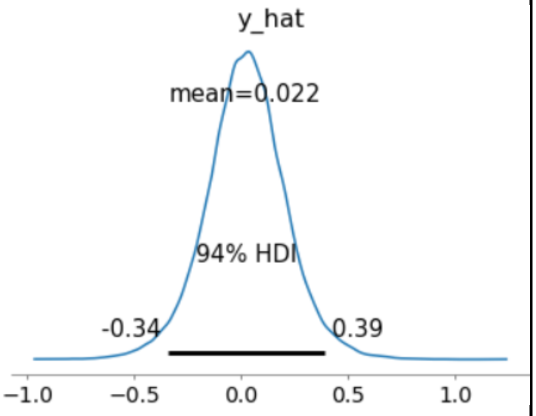
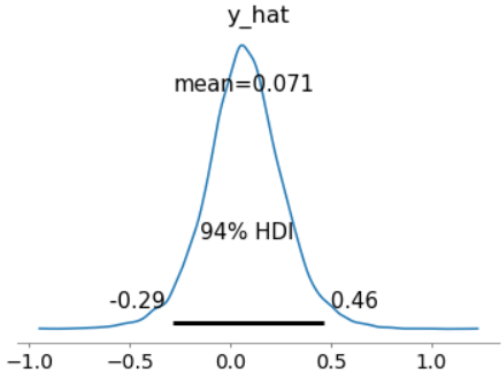
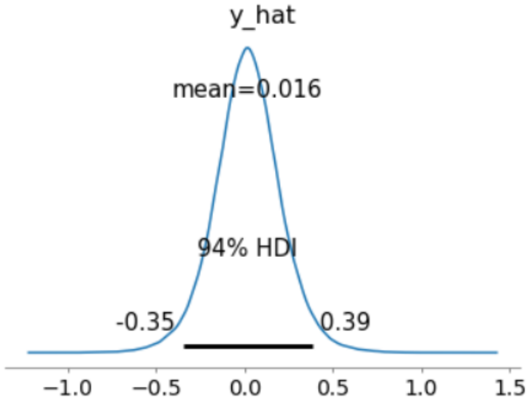
| Diabetic - All Races | Not Diabetic - All Races |
|---|---|
|  |  |

*Figure 20:*
Partially Pooled Credible Intervals:

| | Diabetic | Not Diabetic |
|---|---|---|
| Mexican American |  |  |
| Other Hispanic |  |  |
| Non-Hispanic White |  |  |

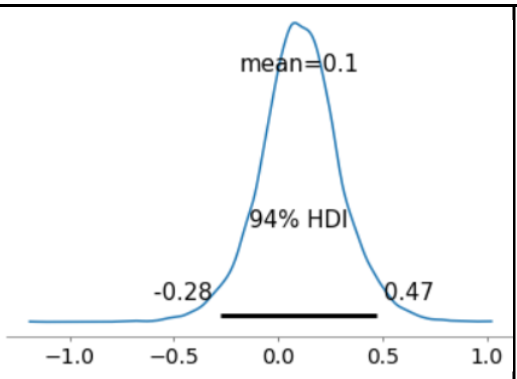
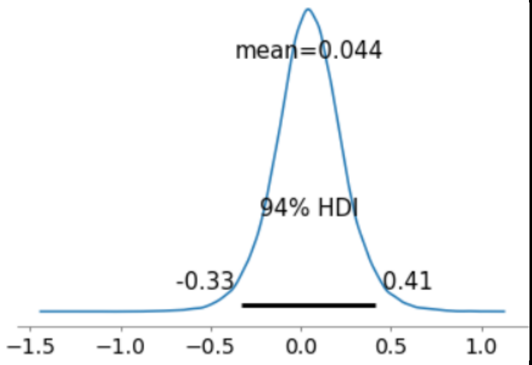
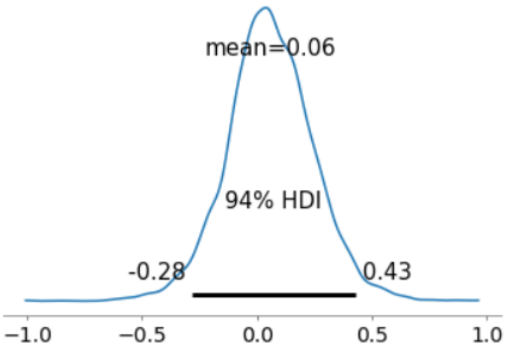
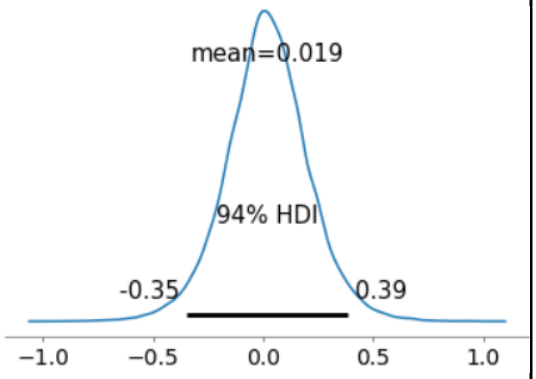
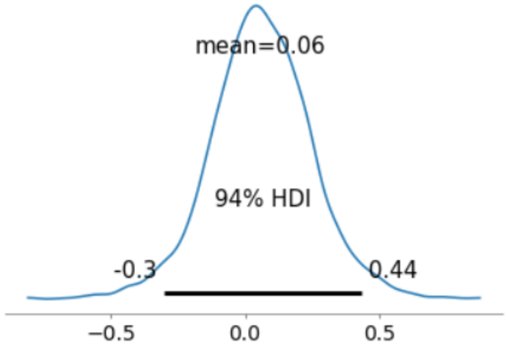
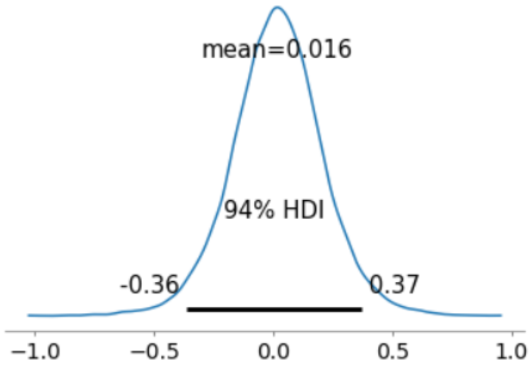
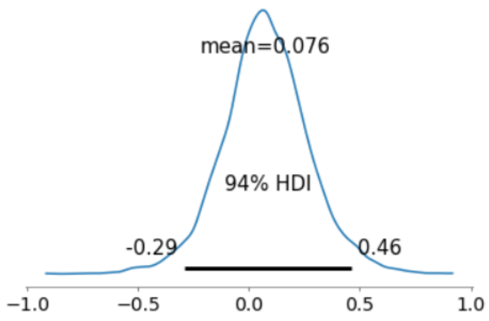
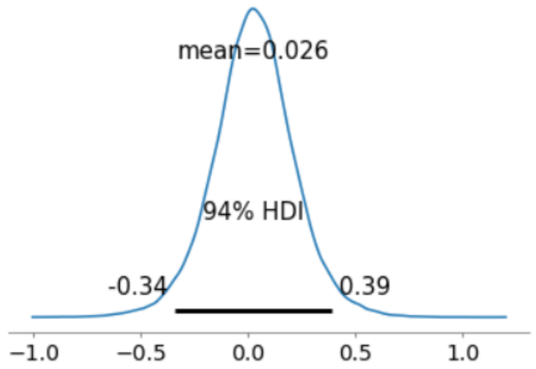
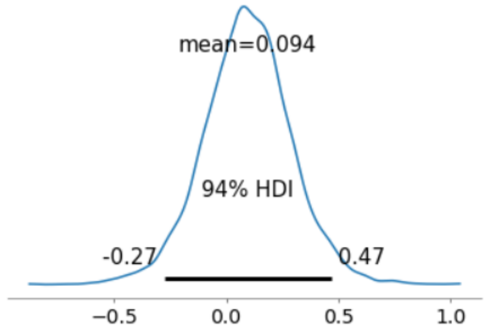
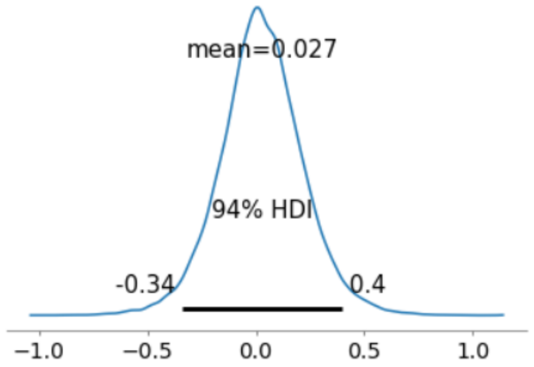
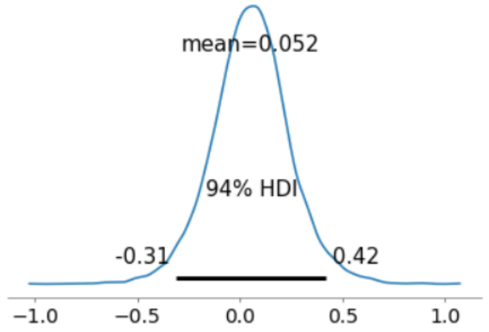
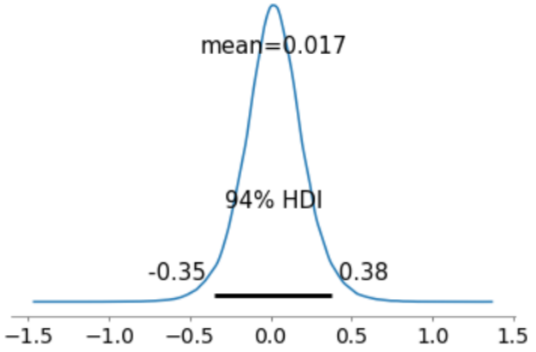
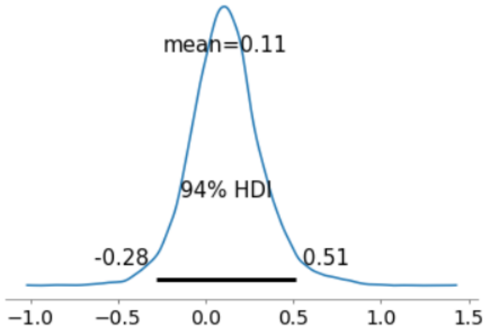
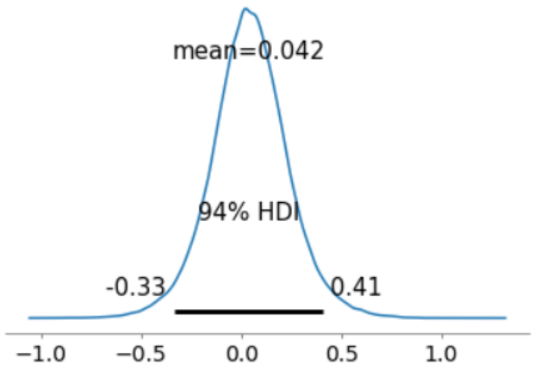
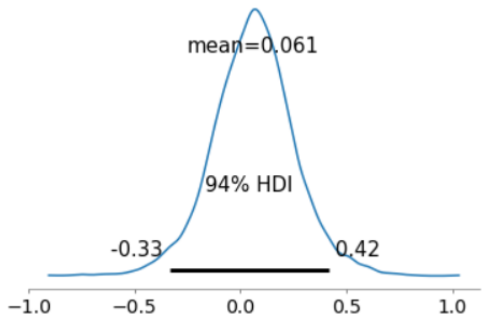
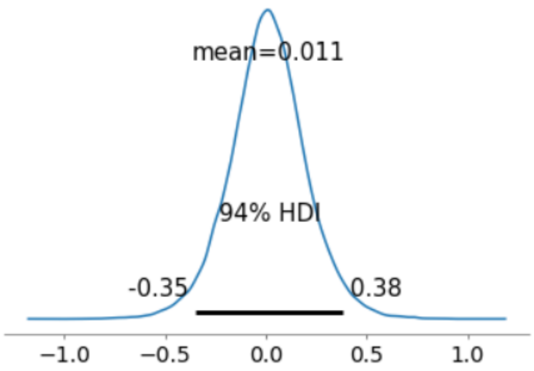
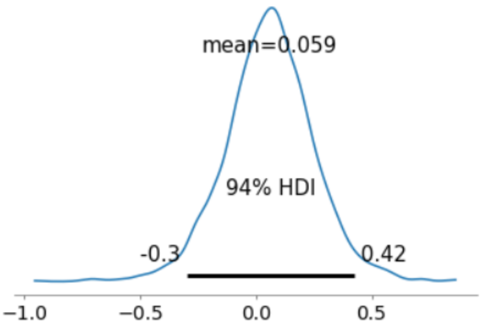
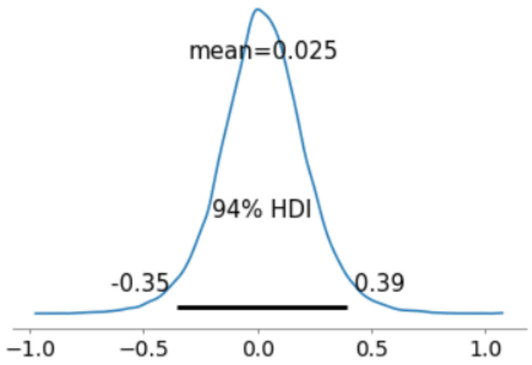
| | | |
|---|---|---|
| Non-Hispanic Black |  |  |
| Non-Hispanic Asian |  |  |
| Other Race - Including Multi-Racial |  |  |

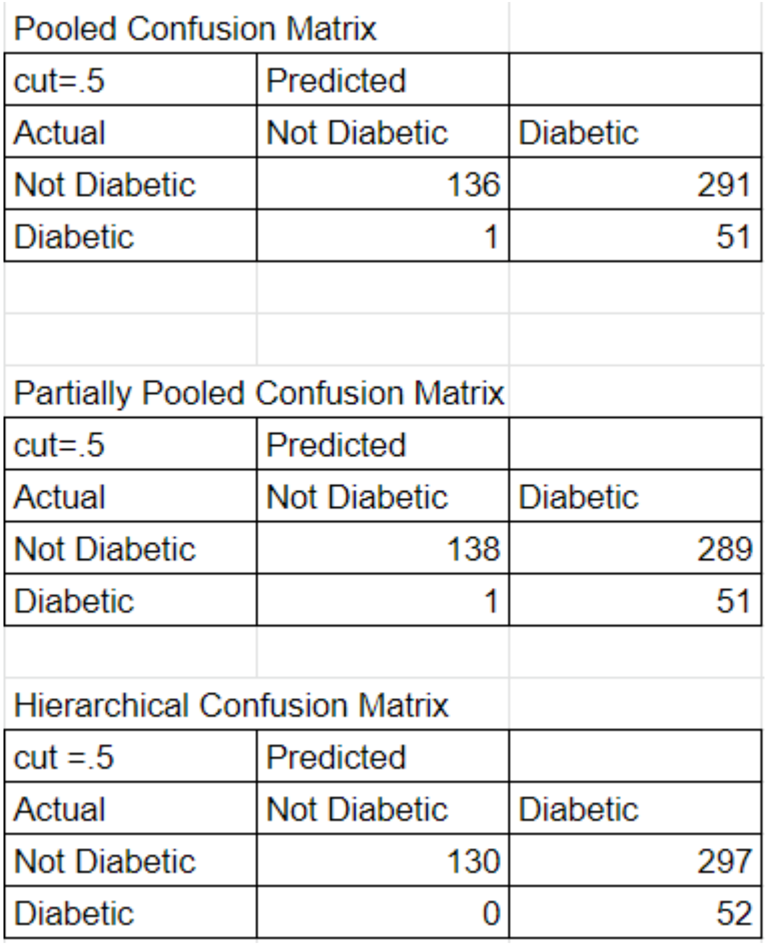
*Figure 21:* Hierarchical Credible Intervals

| | Diabetic | Not Diabetic |
|---|---|---|
| Mexican American |  |  |
| Other Hispanic |  |  |
| Non-Hispanic White |  |  |
| Non-Hispanic Black |  |  |

| | | |
|---|---|---|
| Non-Hispanic Asian |  |  |
| Other Race - Including Multi-Racial |  |  |

**Appendix H**

*Figure 22:* Confusion Matrices



Pooled Confusion Matrix

| cut=.5 | Predicted | |
|---|---|---|
| Actual | Not Diabetic | Diabetic |
| Not Diabetic | 136 | 291 |
| Diabetic | 1 | 51 |

Partially Pooled Confusion Matrix

| cut=.5 | Predicted | |
|---|---|---|
| Actual | Not Diabetic | Diabetic |
| Not Diabetic | 138 | 289 |
| Diabetic | 1 | 51 |

Hierarchical Confusion Matrix

| cut =.5 | Predicted | |
|---|---|---|
| Actual | Not Diabetic | Diabetic |
| Not Diabetic | 130 | 297 |
| Diabetic | 0 | 52 |

# Works Cited:

[1]"Statistics about diabetes," *Statistics About Diabetes | ADA*. [Online]. Available: https://www.diabetes.org/resources/statistics/statistics-about-diabetes. [Accessed: 13-Dec-2021].

[2]Y.-Y. Meng, A. Diamant, J. Jones, W. Lin, X. Chen, S.-H. Wu, N. Pourat, D. Roby, and G. F. Kominski, "Racial and ethnic disparities in diabetes care and impact of vendor-based disease management programs," *Diabetes Care*, vol. 39, no. 5, pp. 743–749, 2016.

[3]"NHANES 2005-2006: Diabetes Data Documentation, codebook, and frequencies," *Centers for Disease Control and Prevention*. [Online]. Available: https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DIQ_D.htm. [Accessed: 13-Dec-2021].

[4]"Cholesterol and diabetes," *www.heart.org*. [Online]. Available: https://www.heart.org/en/health-topics/diabetes/diabetes-complications-and-risks/cholesterol-abnormalities--diabetes. [Accessed: 13-Dec-2021].

[5]Y.-C. Tung, S.-S. Lee, W.-C. Tsai, G.-T. Lin, H.-W. Chang, and H.-P. Tu, "Association between gout and incident type 2 diabetes mellitus: A retrospective cohort study," *The American Journal of Medicine*, vol. 129, no. 11, 2016.

[6]B. Biondi, G. J. Kahaly, and R. P. Robertson, "Thyroid dysfunction and diabetes mellitus: Two closely associated disorders," *Endocrine Reviews*, vol. 40, no. 3, pp. 789–824, 2019.