Bayesian insights into diabetes: Revolutionizing risk prediction

— Project Report —
Advanced Bayesian Data Analysis

Sabarish Suriya Swaminathan Ravichandran [242653]
Christy Johns [261910]
Firosh Mohan Chalissery [269605]

March 13, 2025

TU Dortmund University

Contents

1 Introduction

1.1 Motivation

One of the most common diseases in the United States (US) is Diabetes, which has a significant impact on the economy and quality of life in society. This disease is caused by abnormal glucose levels in the blood and can lead to serious health problems such as heart disease and vision impairment (Kaggle Diabetes Dataset).

1.2 Problem

Although there is no definitive cure for diabetes, early detection of the disease is vital to minimize its effects on quality of life and public health. As a result, robust predictive models are necessary to identify people at risk of developing diabetes.

1.3 Modeling Idea

This project applies a dataset from the Behavioral Risk Factor Surveillance (BRFSS). The response variable in the dataset is binary. Thus, the project uses several techniques, including Bayesian Logistic Regression, and Multilevel (Hierarchical) Intercept models, to model the dataset for predicting diabetes, due to their suitability for handling binary response variables.

1.4 Illustrative Figure

Figure 1 illustrates how the models are used in this project to predict diabetes. In the beginning, a suitable dataset is chosen from Kaggle. Then, the dataset is explored to understand data structure, distributions, and relationships between features. Next, the models suitable for the dataset are selected and trained with the data. After that, the models are evaluated, compared, and the best models are chosen. Finally, the selected models are interpreted to find the significant factors affecting diabetes risk.

Figure 1: Development Steps for Diabetes Prediction Model [self-made, using PowerPoint]

2 Data

2.1 Data Collection

In this project, a dataset from the Behavioral Risk Factor Surveillance (BRFSS) is used. The BRFSS is an annually conducted telephone-based survey by the Centers for Disease Control and Prevention (CDC), and it collects responses from more than 400,000 Americans (Kaggle Diabetes Dataset).

The dataset includes three files (Kaggle Diabetes Dataset) as follows:

- Diabetes_012_health_indicators_BRFSS2021.csv: This is a clean dataset of 236,378 survey responses to the CDC's BRFSS2021. The target variable has 3 classes including 0 for no diabetes or only during pregnancy, 1 is for pre-diabetes, and 2 is for diabetes.
- Diabetes_binary_5050split_health_indicators_BRFSS2021.csv: This is a clean and balanced dataset of 67,136 survey responses to the CDC's BRFSS2021. The target variable is binary, where 0 shows no diabetes, and 1 indicates pre-diabetes or diabetes.
- Diabetes_binary_health_indicators_BRFSS2021.csv: This is a clean and imbalanced dataset of 236,378 survey responses to the CDC's BRFSS2021. The target variable is binary, where 0 represents no diabetes, and 1 shows pre-diabetes or diabetes.

There are 21 feature variables in all three datasets. This project makes use of binary datasets. The datasets don't need any more preprocessing because they have been completely cleansed.

2.2 Source

The dataset used for this analysis is from the CDC's BRFSS dataset for the year 2021, and it is available on Kaggle (Kaggle Diabetes Dataset).

2.3 Data Description

2.3.1 Feature Variables

There are 21 feature variables in the dataset, as shown in Table 1 (Kaggle Diabetes Dataset . Diabetes_binary is the response variable.

Table 1: Feature variables.

Variable	**Values**
Diabetes binary	0 = no diabetes, 1 = prediabetes or diabetes
HighBP	0 = no high BP (Blood Pressure), 1 = high BP
HighChol	0 = no high cholesterol, 1 = high cholesterol
CholCheck	0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years
BMI	Body Mass Index (BMI), calculated as weight (kg) divided by the square of height(m), Metric used to assess body composition
Smoker	0 = no, 1 = yes
Stroke	0 = no, 1 = yes
HeartDiseaseorAttack	0 = no, 1 = yes
PhysActivity	0 = no, 1 = yes
Fruits	0 = no, 1 = yes, Consume Fruit 1 or more times per day
Veggies	1 = consumed 1 or more pieces of vegetable per day
HvyAlcoholConsump	0 = no, 1 = if drink Alcohol >14/week
AnyHealthcare	0 = no health insurance, 1 = health insurance
NoDocbcCost	0 = no cost barred doctor visit, 1 = cost barred doctor visit
GenHlth	1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor
MentHlth	Number of days with poor mental health in the last 30 days
PhysHlth	Number of days with physical illness and injury in the last 30 days
DiffWalk	0 = no, 1 = yes, Difficulty walking or climbing stairs
Sex	0 = Female, 1 = Male
Age	Scale 1-13: 1 = 18-24,..., 8 = 55-59,..., 13 = 80 or older
Education	Scale 1-6: 1 = Never attended school or only kindergarten, 2 = Grades 1 through 8,..
Income	Scale 1-8: 1 = less than $10,000,..., 5 = less than $35,000,..., 11 = $200,000 or more

2.3.2 Correlations of Feature Variables

Figure 2 shows how all feature variables are correlated. From this figure, it can be seen that some independent variables are seen to be somewhat correlated. This might lead to regression problems. For this time, no assumptions should be made regarding potentially removing variables from the models, even if this could cause issues with regression. In general, the characteristic(features) variables do not exhibit significant collinearity.

we observe that certain independent variables exhibit moderate correlations. For example, Income and Education show a correlation of r = 0.43, while PhysHlth and MentHlth are correlated at r = 0.54. However, most variables show weak or insignificant correlations, suggesting that multicollinearity is unlikely to be a major issue. At this stage, no assumptions should be made about excluding variables from the model.

The target variable, Diabetes_binary, also shows notable associations with several characteristics(features): GenHlth (r = 0.38) Age (r = 0.29) HighBP (r = 0.37) These correlations suggest that these characteristics(features) could serve as meaningful predictors in the model.

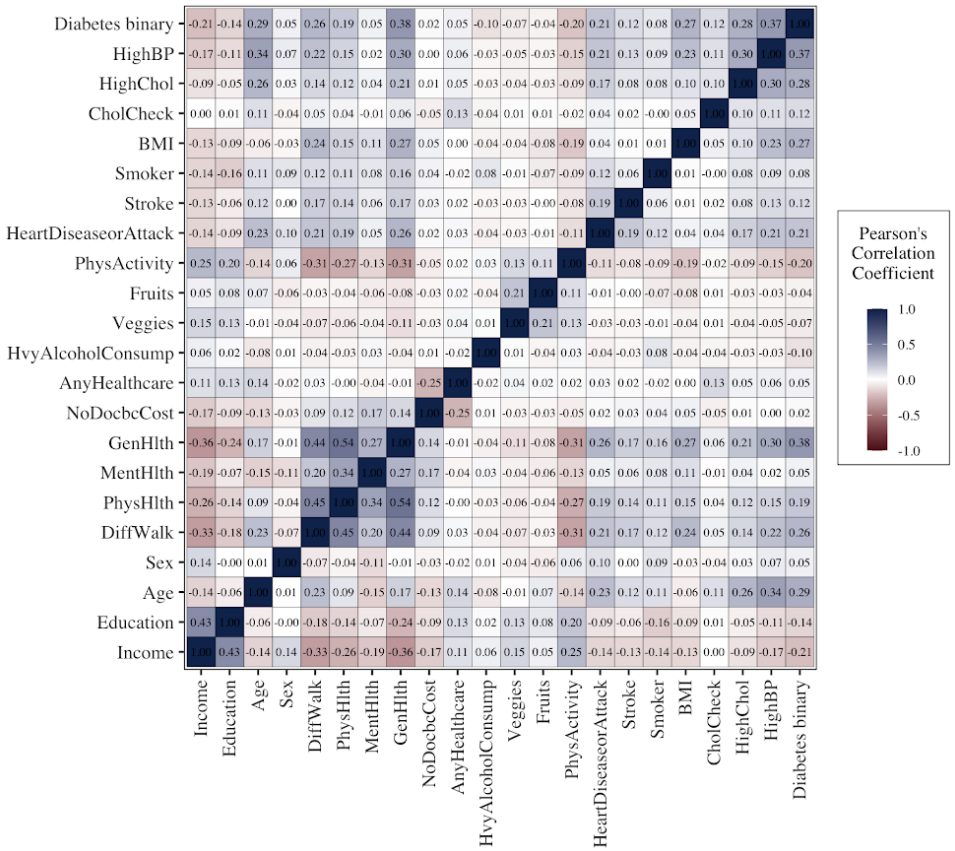

Figure 2: Correlation plot of the feature variables. [Self-made, using R]

There are three feature groups in the data set as follows:

- Categorical-Non Ordinal (e.g., HighBP, ...)
- Categorical-Ordinal (e.g., GenHlth, ...)
- Metric (e.g., BMI)

Additionally, all non-binary variables were standardized to ensure consistency in scale and improve model performance

2.3.3 Distributions of Key Feature Variables

The distribution of several important feature variables that showed strong associations with the response variable (Diabetes_binary), such as Age, GenHlth, BMI, and Income is explained as follows:

Figure 3 The density plot of BMI indicates that individuals with lower BMI are predominantly 'No Diabetes', while those with higher BMI are more likely to be 'Prediabetes/Diabetes'.

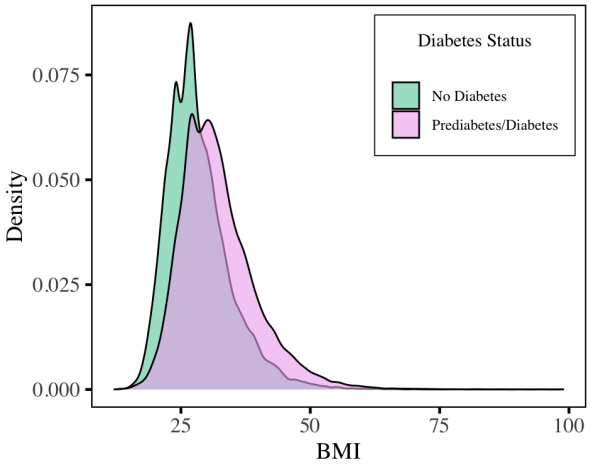

Figure 3: Density plot of BMI by Diabetes_binary. The green curve is for no diabetes and the pink curve is for Prediabetes or Diabetes. [self-made, using R]

Figure 4 According to the study, those in better general health categories are more likely to have "No Diabetes," whereas people in worse categories are more likely to have "Prediabetes/Diabetes."

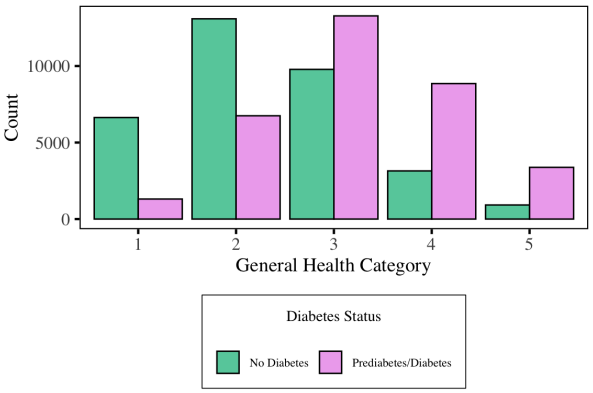

Figure 4: Histogram of GenHlth by Diabetes_binary. The green boxes are for no Diabetes and the pink boxes are for Prediabetes or Diabetes. [self-made, using R]

Figure 5 Diabetes prevalence is strongly influenced by age, with older age groups reporting more "Prediabetes/Diabetes" cases and younger age groups reporting more "No Diabetes" cases.

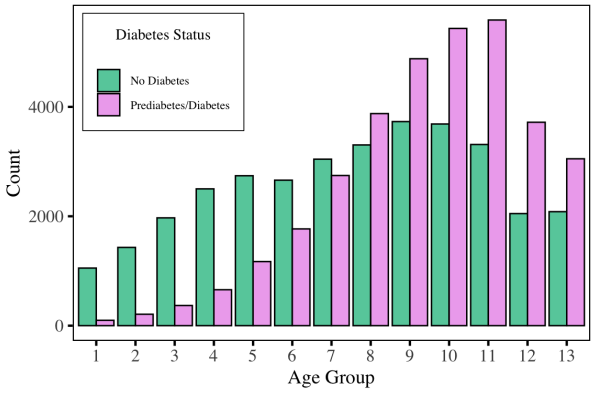

Figure 5: Histogram of Age by Diabetes_binary.The green boxes are for no Diabetes and the pink boxes are for Prediabetes or Diabetes. [self-made, using R]

The histogram in Figure 6 indicates a negative relationship between income and diabetes status, with lower income groups more likely to have prediabetes or diabetes and higher income groups having a more balanced distribution of diabetes.

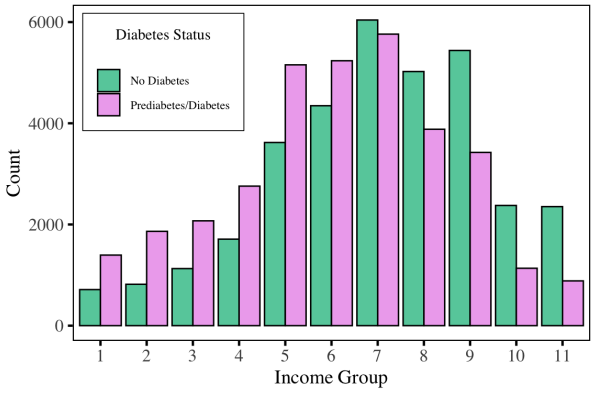

Figure 6: Histogram of Income by Diabetes_binary.The green boxes are for no Diabetes and the pink boxes are for Prediabetes or Diabetes. [self-made, using R]

3 Models

Several techniques, such as Bayesian Logistic Regression, and Multi-level (Hierarchical) Intercept models with different priors, are used in this study to model the data in order to predict diabetes. These models are explained in the following sections. All of the Bayesian models employed in this project are Generalized Linear Models (GLMs), which link the parameters to the likelihood using a Link function. This is necessary because the Bernoulli distribution is a very useful likelihood for binary classification, and because it only accepts values between 0 and 1, the possible values of the parameters must be transferred in some way to the interval $[0; 1]$.

3.1 Bayesian Logistic Regression (Logit Link Function)

The most popular GLM for binary classification issues is logistic regression. In Logit Regression, the parameters are mapped to the $[0; 1]$ interval using the logistic link function. This is the simplest and best model for binary classification in general, although it can lose its effectiveness when dealing with highly imbalanced datasets.

The following is a representation of the model:

$$y_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit}(p_i) = x_i^T \beta$$

where:

- y_i is the binary response variable for observation i.
- p_i is the probability of success.
- x_i is the vector of predictors for observation i.
- β is the vector of coefficients.
- $\text{logit}(p_i)$ is the log-odds of the probability of success, modeled using the logistic link function $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$.

The prior distribution for the coefficients β is typically assumed to be a multivariate normal distribution:

$$\beta \sim \text{Normal}(\mu_0, \Sigma_0)$$

where μ_0 is the vector of prior means for the coefficients, and Σ_0 is the prior covariance matrix for the coefficients.

3.2 Multilevel Intercept Model (Hierarchical Model)

One type of multilevel model is the multilevel intercept model, where the intercept can vary depending on one of the factors. The data will be grouped using this parameter, and each manifestation of the parameter will have a distinct intercept.

The following is a representation of the Multilevel Intercept model:

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij}$$

where:

- y_{ij} is the outcome variable for observation i in group j.
- β_{0j} is the intercept for group j.
- β_1 is the fixed effect coefficient for the predictor variable x_{ij}.
- x_{ij} is the predictor variable for observation i in group j.
- ϵ_{ij} is the error term.

The Multilevel Intercept model allows group-specific intercepts to be estimated while accounting for the hierarchical nature of the data.

This project uses logistic regression with multilevel models. In this case, the variable intercept is grouped using the GenHlth decision. After evaluating various combinations and groupings, it was determined that, of all the models that were tried, using GenHlth as the Intercept worked the best.

4 Priors

Priors, which inform the model of the range and distribution of parameters before it can read the data, are an essential component of Bayesian inference. By accounting for prior knowledge and parameter beliefs, Bayesian models can produce more accurate and detailed estimates, particularly when working with limited or inaccurate information.

We will analyse a variety of priors and their effects to demonstrate how our results are sensitive to prior data and to provide insight into the resilience of Bayesian models. Throughout this analysis, a range of priors are specifically tailored to the unique features of each model, and the priors used are intended to capture the uncertainty in the parameters that need to be estimated and are based on our understanding of the underlying phenomena.

The Bayesian logistic regression models Models 1, 2, and 3 as well as Models 4 that examined interaction effects and multilevel intercepts were all subjected to explicit priors in this investigation.

4.1 Normal prior

A normal prior distribution was employed.Usually, it is assumed that the coefficients are distributed according to a specified mean and with some variability. The proper distribution and range of regression coefficients have been assumed by selecting a normal prior distribution. The data was standardised for this purpose.

The normal prior distribution with mean μ (mean) and σ^2 (variance) can be expressed as:

$$f(\theta|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right)$$

Where:

- $f(\theta|\mu, \sigma^2)$ is the probability density function of the normal distribution with parameters μ and σ^2.
- θ represents the random variable.
- μ is the mean (or expectation) of the distribution.
- σ^2 is the variance of the distribution.
- $\sqrt{2\pi\sigma^2}$ is the normalization factor ensuring that the total area under the curve equals 1.
- $\exp(-\frac{(\theta-\mu)^2}{2\sigma^2})$ is the exponentiated term that determines the shape of the normal distribution.

4.2 Cauchy prior

Cauchy priors were taken into consideration, which have tails that extend further than typical curves. Since it's unknown whether extreme values could exist in the data, the Cauchy approach is used in this case. By adding Cauchy priors to the estimate procedure, it takes into consideration possible anomalous results.

The Cauchy prior distribution with location parameter x_0 (location) and scale parameter γ (scale) can be expressed as:

$$f(x|x_0,\gamma) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$$

Where:

- $f(x|x_0,\gamma)$ is the probability density function of the Cauchy distribution with parameters x_0 and γ.
- x represents the random variable.
- x_0 is the location parameter, representing the location of the peak of the distribution.
- γ is the scale parameter, representing the half-width at half-maximum (HWHM) of the distribution.
- π is the mathematical constant pi.

4.3 Logistic prior

Additionally, similar to linear regression, the logistic prior distribution exhibits a gradual drop towards the ends and a larger density near 0. This choice was made in order to avoid overweighting. with reliance on previous theories on the sparsity of the coefficients.

The logistic prior distribution with location parameter μ and scale parameter s can be expressed as:

$$F(x|\mu,s) = \frac{1}{1+e^{-(x-\mu)/s}}$$

where:

- $F(x|\mu,s)$ is the cumulative distribution function of the logistic distribution with parameters μ and s.
- x represents the random variable.
- μ is the location parameter, representing the location of the distribution.
- s is the scale parameter, representing the scale or spread of the distribution.

- e is the base of the natural logarithm.

To examine interaction effects and multilevel intercepts, certain relevant priors that are adapted to the data hierarchy are required. These priors improved the accuracy of parameter estimations by capturing the hierarchical character of the data and informing not just the main effects but also the interaction terms and random effects.

The priors used in this analysis often had minimal effects on our models because of the size of the dataset used in the analysis.

5 Code

In order to predict diabetes based on health markers, the brms package for Bayesian logistic regression modelling is implemented using the R programming language, a efficient tool. R offers a versatile and user-friendly framework for Bayesian data analysis when used with the brms package. For effective model fitting, Brms uses Stan to offer sophisticated sampling methods and automatic differentiation. It can be easily integrated into current workflows because to its interaction with the R environment.

5.1 Bayesian regression models using Stan (BRMS)

The brms package provides an interface for fitting Bayesian generalised (non-)linear multivariate multilevel models using Stan. For doing regression analysis, the formula syntax provides a recognisable and user-friendly interface that is quite similar to that of the software lme4.

In a multilevel context, users can fit a variety of models, such as ordinal, zero-inflated, hurdle, count data, survival, response times, robust linear, and even self-defined mixed models. This is made possible by the large number of supported distributions and connection functions. Modelling choices include smooth and non-linear variables, autocorrelation structures, censored data, meta-analytic standard errors, and many more. Furthermore, by predicting every parameter in the answer distribution, distributional regression can be carried out. Users are encouraged to use the distributions that best reflect their values because the previous criteria are flexible. Leave-one-out cross-validation and posterior predictive tests allow for the rapid evaluation and comparison of model fit. ([**BRMS**]). Using methods such as Widely Applicable Information Criterion (WAIC) and Leave-One-Out Cross-Validation (LOO), several models are constructed using the same predictors but distinct estimated prior distributions. The algorithm finds the best-fitting, performing model. It uses diagnostics like posterior predictive checks and convergence to confirm the accuracy and dependability of the chosen model. In order to deal with skewed data, this approach also assesses predictive accuracy on imbalanced datasets. The ideal model is prepared for use and additional research as soon as it is identified with exact parameters and priors.

5.2 Prior Specification

It helps to clarify expectations on the range and probabilities of coefficients before looking at data. Normal priors are typically used for regression coefficients. They have a standard deviation of one and a midpoint of zero. Prior(normal(0, 1), class="b") is equal to this basic normal prior function. One of the most important tasks in Bayesian regression is setting priors. It guides the model toward reasonable parameter magnitudes by limiting acceptable values. By doing this, Bayesian regression analysis's accuracy and interpretability are improved.

5.3 Chains

Bayesian analysis requires multiple chains. Here, either four or eight chains were utilized. The figure affects the model's ability to converge and provide accurate estimations. Adding more chains can assist ensure that the model fully investigates every possibility. The likelihood of it getting stuck in local places and missing the actual dispersion decreases. Diagnostics such as the Gelman-Rubin statistic are more accurate in assessing convergence when there are more chains. Variation both inside and between chains is examined by this statistic. Values close to 1 indicate proper convergence. However, utilizing an excessive number of chains also requires additional processing resources. Therefore, the number strikes a compromise between maintaining efficiency and obtaining precise outcomes. In this manner, the Bayesian model produces reliable conclusions.

5.4 Model Complexity

Links between predictions and actual events are displayed by the brm function. When you include more predictor variables and areas where things interact, it becomes more complex. In a less straightforward manner, special words describe the way predictors collaborate. This improves the model's comprehension of the specifics. Each of these components provides a clear image of the key relationships in the diabetes prediction Bayesian logistic regression models.

5.5 An Overview of the Multilevel Intercept Model (Hierarchical Model), the best model utilized in this analysis

The code output that follows shows an analysis of the Multilevel Intercept Model conducted in R using the brms package. It encapsulates this analysis. By using a variety of predictor factors, the model seeks to predict diabetes binary outcomes.Effects at the population level are indicated by estimates. The mean contribution of each predictor variable to the log-odds of diabetes is shown here.Higher values are associated with higher log-odds of diabetes, according to positive coefficient estimates for variables such as CholCheck, HighBP, and HighChol.

There are also group-level effects. The baseline log-odds of diabetes vary among different levels of perceived general health, as seen by the standard deviation estimate for the intercept across levels of the grouping variable GenHlth. The standard deviation estimate for the intercept at different levels of the grouping variable GenHlth shows the Group Level Effects. This illustrates how varying baseline log-odds of diabetes correlate with varying degrees of perceived overall health. Rhat evaluates the convergence of Markov Chain-Monte Carlo (MCMC) chains. Convergence of the possible scale reduction factor is indicated by values close to 1. When taken as a whole, the model shows that the likelihood of developing diabetes is significantly influenced by age, BMI, blood pressure, cholesterol test results, and other health indicators. Future studies and intervention initiatives are informed by these findings.

```
  Family: bernoulli
  Links: mu = logit
Formula: Diabetes_binary ~ CholCheck + HighBP + HighChol + Stroke + HeartDiseaseorAttack + HvyAlcoho
   Data: scaledData[sampleIDs, ] (Number of observations: 3000)
  Draws: 8 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup draws = 6000

Multilevel Hyperparameters:
~GenHlth (Number of levels: 5)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.21      0.54     0.54     2.62 1.00     1874     2860

Regression Coefficients:
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept               -1.97      0.72    -3.44    -0.55 1.00     2291     2413
CholCheck1               1.07      0.35     0.39     1.80 1.00     7622     4020
HighBP1                  0.67      0.10     0.48     0.85 1.00     7876     4005
HighChol1                0.56      0.09     0.38     0.73 1.00     9262     4448
Stroke1                  0.12      0.19    -0.24     0.49 1.00     8571     4177
HeartDiseaseorAttack1    0.29      0.14     0.02     0.56 1.00     8361     4785
HvyAlcoholConsump1      -0.86      0.21    -1.28    -0.44 1.00     8087     3671
AnyHealthcare1           0.22      0.24    -0.26     0.69 1.00     8350     4070
BMI                      0.57      0.05     0.47     0.68 1.00     7630     4836
Age                      0.49      0.05     0.39     0.59 1.00     7409     4637
Sex1                     0.24      0.09     0.07     0.41 1.00     8138     4677
Education               -0.08      0.05    -0.17     0.01 1.00     8551     3813
BMI:Age                  0.03      0.05    -0.08     0.12 1.00     9338     4141
```

The Appendix of this report contains a summary of every model made with the BRMS software.

A more visual representation of these parameters can be seen in Figure 7. Here, the Credibility Intervals are plotted using a Boxplot diagram, which shows both the estimated value as well as the Credibility Bounds. CholCHeck for example has a very positive influence on the likelihood of a patient having diabetes while having healthcare seems to reduce the risk of Diabetes.

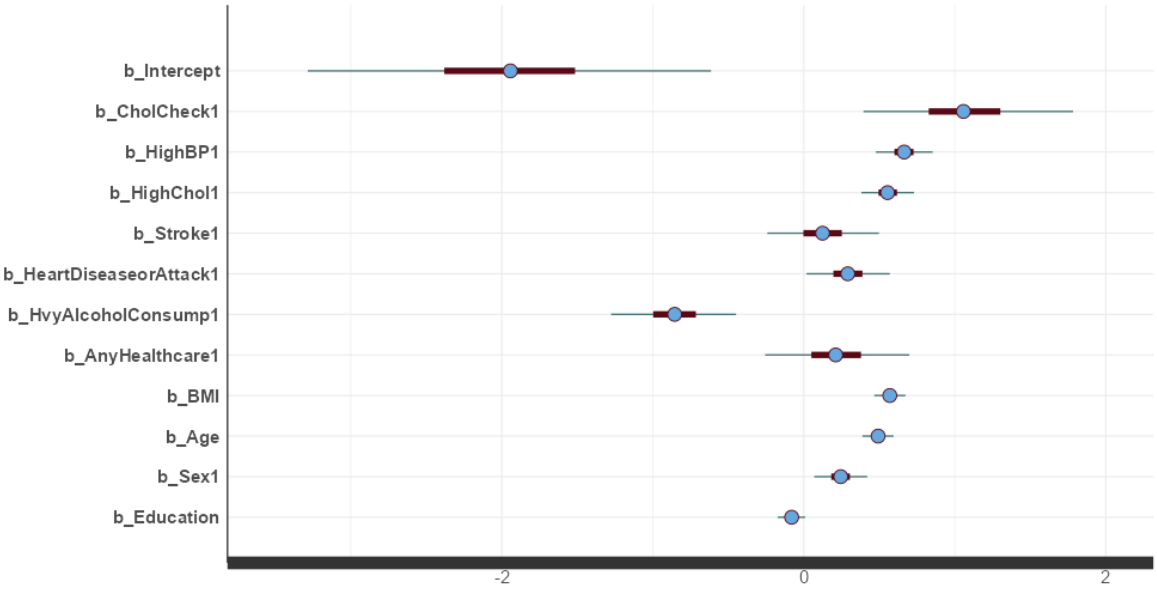

Figure 7: Boxplot of the Coefficient Values for the Multilevel Intercept Model (Hierarchical Model). [self-made, using R]

6 Divergence Diagnostics

6.1 Trace Plots

The Multilevel Intercept Model described in Sections 3 and 5 are tested using various methods to see if they converge. First, the parameters' trace plots can be used to identify any significant issues that may have arisen throughout the MCMC process and to provide a broad sense of the overall convergence. Figure 8 trace plots demonstrate that, for the model in use, there are no discernible convergence problems. Due to the report's brevity, all other developed models are excluded because their trace plots are extremely comparable.

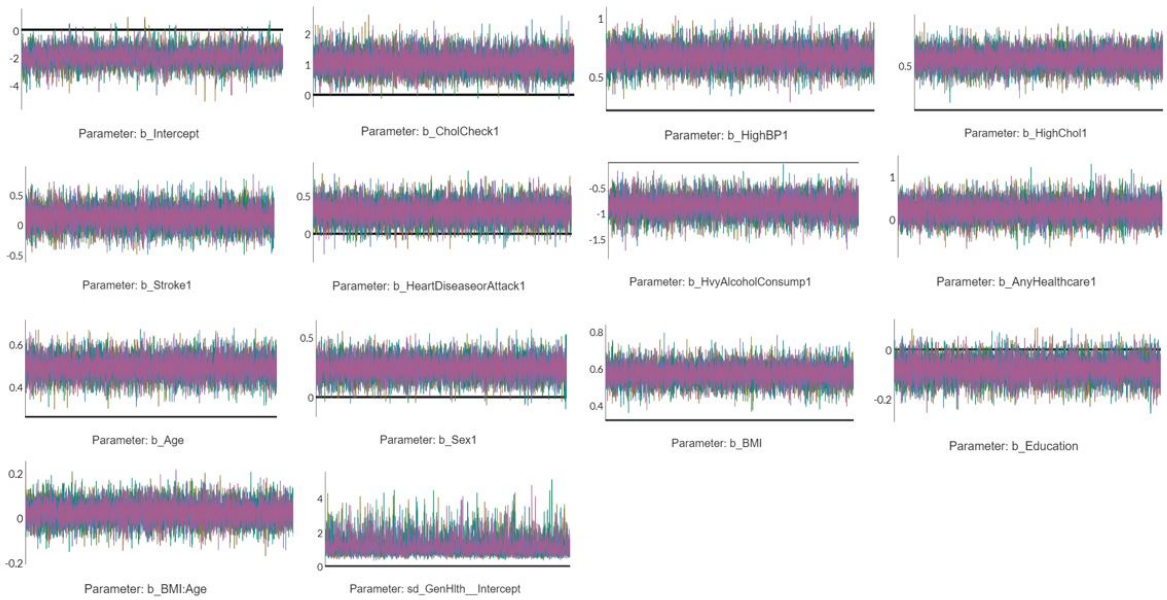

Figure 8: Trace Plots of the Parameters for Multilevel Logistic Model. [self-made, using R]

All of the created Models properly converge, this can be shown using different summary Statistics.

6.2 Dens Plots

The distribution of variables is ascertained by analyzing the developed Models that are detailed in Seections 3 and 5. Figure 9 demonstrates how the coefficients have a normal distribution.

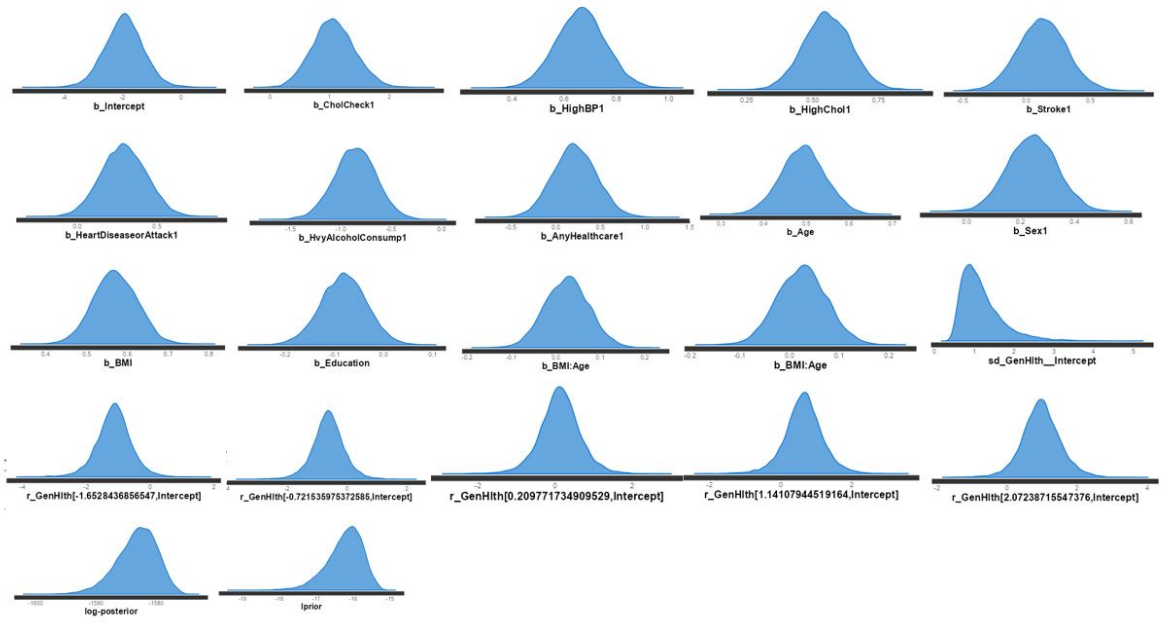

Figure 9: Dens plots of the Parameters for Multilevel Logistic Model. [self-made, using R]

7 Prior Predictive Check

Model <chr>	Mean_Prediction <dbl>	SD_Prediction <dbl>
Normal Prior	0.5011060	0.006471435
Logistic Prior	0.5013053	0.006366352
Cauchy Prior	0.5011177	0.006462410

Figure 10: Prior Predictive Distribution

Mean Prediction (0.501 for all models)
Since the mean of prior predictions is very close to 0.5 for all three priors, this suggests that the prior distributions are not strongly skewing predictions before the data is introduced.

SD Prediction (0.0064 for all models)
The standard deviations are very small (0.0064), indicating that the prior predictive distributions are quite narrow. This suggests that the prior distributions do not allow for much variation in predictions before observing data.

Overall Conclusion
All three priors produce very similar results, meaning the model is not very sensitive to the choice of prior. This suggests that the priors are reasonable. Low SD (0.0064) suggests that the model assigns most prior predictive probability around 0.5, indicating low variability in predictions.

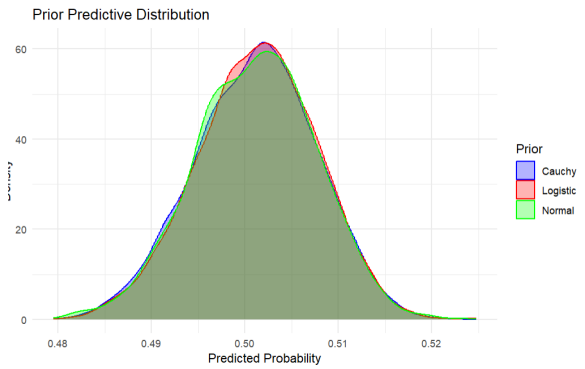

Figure 11: Prior Predictive Distribution

8 Posterior Predictive Checks

In Posterior Predictive Checks, the created Model is used to generate simulated Data. This Data can be compared to actual observed Data to asses general goodness of fit and as a Sanity Check to see if the created Models make sense or have significant problems. In Figure 12 it can be seen that the Model(s) follow the Mean of the observed Data relatively well. While this isn't too difficult to achieve in Logistic Regression, it can still make sense to look at since if it was no closely following the Observed Mean, this would indicate some serious Issues with the Model.

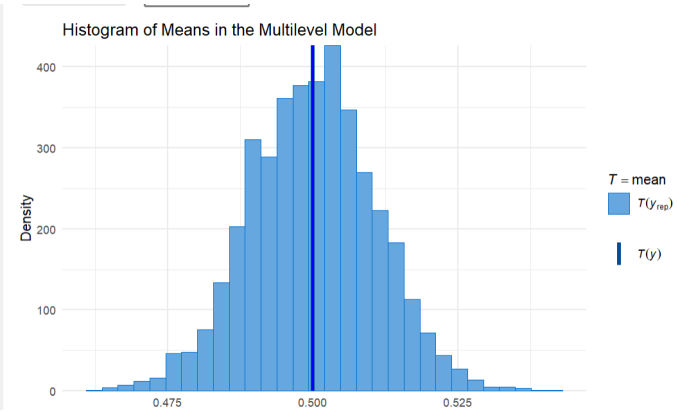

Figure 12: Histogram of Means in the Multilevel Model compared to the Observed Data. The Light Blue bars are the simulated Data and the Dark Blue Line is the True Observed Value. [self-made, using R]

Similarly, Figure 13 shows that the Multilevel Model's simulated Data closely approaches the standard Deviation of the Observed Data. This is the Main Model that is examined in this Report, but these Plots look very Similar for all other Created Models as well.

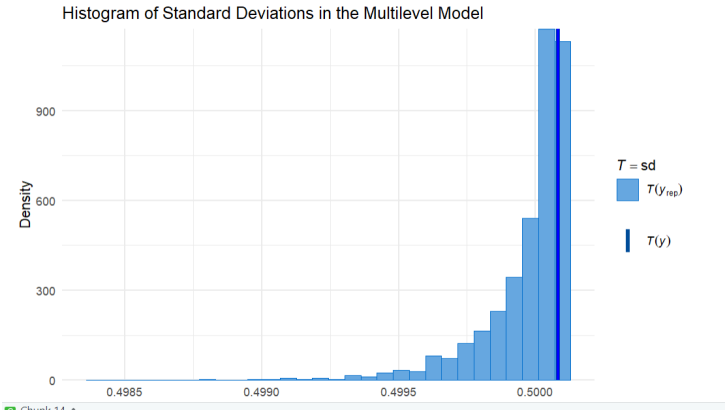

Figure 13: Histogram of Standard Deviations in the Multilevel Model compared to the Observed Data. The Light Blue bars are the simulated Data and the Dark Blue Line is the True Observed Value. [self-made, using R]

9 Model Comparison

To compare the Models, this report uses the LOO-CV (Leave One Out Cross Validation) Approximation of ELPD (Expected log Point-wise Predictive Density). This Can be Calculated for a Model using the Formula:

$$\widehat{elpd}_{loo}(M_k|y) = \sum_{i=1}^{n} \log p_{M_k}(y_i|y_{-i})$$

Where M_k is the Model used, y is the Dataset, y_i is the ith Observation of the Data and y_{-i} is the Dataset without the ith Observation. This kind of Model re-evaluation with LOO-CV is very efficient to Compute for Bayesian Models, which makes it a good option for Model Comparison.

In Table 2 The table compares different Bayesian logistic regression models trained on balanced data, ranked by their Expected Log Predictive Density (ELPD). The Logit Multi. Bal. model achieves the highest ELPD (-1568.8), indicating the best predictive performance, and serves as the reference model with an $ELPD_{diff}$ of 0.0. Other models, including Logit Normal Bal., Logit Cauchy Bal., and Logit Logistic Bal., have slightly lower ELPD values, but their differences (ranging from -1.0 to -1.5) are within the range of uncertainty as indicated by their SE_{diff} values (7.0). Since these differences are small and not statistically significant, the alternative models remain competitive choices. This suggests that while the Logit Multi. Bal. model is the best performer, the other models provide similar predictive power and could still be viable depending on factors such as interpretability and computational efficiency.

Table 2: Models trained on Balanced Data From Highest to Lowest ELPD

Model	$ELPD$	SE	$ELPD_{diff}$	SE_{diff}
Logit Multi.	-1568.8	27.9	0.0	0.0
Logit Normal	-1569.8	28.2	-1.0	6.9
Logit Cauchy	-1570.2	28.3	-1.4	7.0
Logit Logistic	-1570.3	28.3	-1.5	7.0

10 Predictive Performance

The predictive Performance of the created Model can be estimated using test Data. For this purpose, 3000 observations that are left out in the model Training are used and the model tries to predict their dependent Variable. Here, the balanced Dataset and the model trained on it are used, since these more closely represent real world Data. As can be seen in Table 3,The model demonstrates moderate predictive performance, correctly classifying 75.3% of cases, with a balanced ability to identify both diabetic (sensitivity: 74.07%) and non-diabetic individuals (specificity: 76.33%). While it slightly outperforms a baseline model that always predicts the majority class (75.7% vs. 75.3% precision), its predictive power remains limited. The F1-score (74.91%) indicates a reasonable balance between precision and recall, but the McNemar's test suggests no significant difference in misclassification rates. Overall, while the model provides some value in predicting diabetes, its reliability is not strong enough for standalone clinical use and should be supplemented with additional diagnostic tools.

Table 3: Predicted Vs. True Value on 3000 new Observations for Multi. Logit Model

Prediction / Actual Value	Diabetes	No Diabetes
Diabetes	1145	389
No Diabetes	355	1111

11 Limitations and Potential Improvements

Computational power was one of the analysis's main limitations. Just 3,000 of the approximately 70,000 observations were used to train the models. Better models with improved prediction performance might be produced by using more data and/or increasing the number of iterations or chains in the models. Better predictions for upcoming patients may also result from the utilization of more recent data.

For future Work, other types of Models could be considered. These include Models like Bayesian Decision Trees, Bayesian Neural Networks and Bayesian Random Forests. Non-Bayesian Models like Neural Networks and gradient-Boosted Decision Trees could also be used to obtain information and predictive Models from the Data.

12 Conclusion

This study explored the application of Bayesian data analysis techniques for the prediction of diabetes risk using the BRFSS dataset. Using Bayesian logistic regression and hierarchical modeling, we demonstrated the potential of probabilistic approaches in handling binary classification problems with real-world health data. The multilevel intercept model emerged as somewhat effective in being able to predict Diabetes correctly 75.3% of the time, and having a true positive rate of about 74. 3%, capturing the variability across different health categories while providing interpretable insights into diabetes risk factors.

Our findings indicate that key predictors such as general health status, age, BMI, and cholesterol levels significantly influence diabetes risk. Model comparisons using LOO-CV and posterior predictive checks confirmed the robustness of our Bayesian framework, with well-calibrated priors ensuring stable inferences.

In conclusion, Bayesian modeling offers a powerful approach for risk prediction in healthcare, combining flexibility with uncertainty quantification. Future research should focus on integrating larger datasets and refining model structures to enhance predictive capabilities and clinical applicability.

13 Self Reflection

Our team has learned a great deal about the field of Bayesian data analysis during the research and production of this study. This entails becoming knowledgeable about the various Bayesian Regression Models' theoretical and applied components. In particular, our understanding of Bayesian Multilevel Models and Bayesian Binary classification have increased significantly and we learned about a lot of different ways to structure Models. These include different Link functions and priors that can be used in Models, as well as ways of comparing Models to each other and performing Model diagnostics.

We also learned valuable practical skills, like working with Stan and BRMS. These skills will be helpful in the future and open up new ways to explore and analyse Data. The different ways of creating plots for Bayesian Models and interpreting them will be a great addition to our overall skill sets and could prove very beneficial in our future academic and professional work.

References

1.Robert Kissell, Jim Poserina. (2017). Optimal Sports Math, Statistics, and Fantasy.

2.Jean-Michel Marin, C. P. R. (2007). Generalized linear models. In Bayesian core: A practical approach to computational bayesian statistics (pp. 85–118). Springer New York..

3.Alex Teboul. (2022). Diabetes Health Indicators Dataset.

4.Sergio Verd´u. (2023). The Cauchy Distribution in Information Theory. .

5.Sivula, T., Magnusson, M., Matamoros, A. A., Vehtari, A. (2023). Uncertainty in bayesian leave-one-out cross-validation based model comparison..

6.Encylopedia of Mathematics. (2024). Normal Distribution..

7.Paul Buerkner. (2024). Bayesian regression models using Stan..

8.Posterior and Prior Predictive Checks, 2024.

9.LOO package glossary, ELPD and elpd$_{loo}$

14 Appendix

14.1 Appendix A: Output of Logit Model with Normal(0,1) Prior

```
Family: bernoulli
Links: mu = logit
Formula: Diabetes_binary ~ HighBP + HighChol + CholCheck + BMI + Smoker + Stroke
+ HeartDiseaseorAttack+ PhysActivity + Fruits + Veggies + HvyAlcoholConsump + AnyHealthcare
+ NoDocbcCost + GenHlth + MentHlth + PhysHlth + DiffWalk + Sex + Age + Education + Income

Data: scaledData[sampleIDs, ] (Number of observations: 3000)
Draws: 4 chains, each with iter = 3000; warmup = 1500; thin = 1;
    total post-warmup draws = 6000

Regression Coefficients:
                      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                -1.91      0.44    -2.77    -1.08 1.00     9703     5425
HighBP1                   0.67      0.10     0.47     0.85 1.00     7537     4471
HighChol1                 0.56      0.09     0.38     0.74 1.00     8082     4401
CholCheck1                1.07      0.35     0.42     1.77 1.00     9324     4728
BMI                       0.56      0.05     0.45     0.66 1.00     7062     4794
Smoker1                   0.01      0.09    -0.17     0.19 1.00     7527     4420
Stroke1                   0.08      0.19    -0.30     0.46 1.00     8016     4723
HeartDiseaseorAttack1     0.28      0.14     0.00     0.56 1.00     7978     4397
PhysActivity1            -0.20      0.11    -0.40     0.01 1.00     8089     4741
Fruits1                   0.01      0.09    -0.17     0.19 1.00     8587     4829
Veggies1                 -0.11      0.12    -0.33     0.12 1.00     8318     4871
HvyAlcoholConsump1       -0.82      0.21    -1.25    -0.41 1.00     9496     4361
AnyHealthcare1            0.26      0.25    -0.21     0.75 1.00     7743     5055
NoDocbcCost1             -0.09      0.19    -0.46     0.28 1.00     7217     4891
GenHlth                   0.61      0.06     0.50     0.73 1.00     6744     4663
MentHlth                 -0.06      0.05    -0.16     0.04 1.00     8471     4860
PhysHlth                 -0.01      0.06    -0.12     0.10 1.00     6543     4901
DiffWalk1                 0.09      0.13    -0.16     0.35 1.00     7208     4352
Sex1                      0.28      0.09     0.10     0.46 1.00     7332     4977
Age                       0.45      0.05     0.34     0.56 1.00     6283     5080
Education                -0.02      0.05    -0.12     0.08 1.00     6078     4488
Income                   -0.11      0.05    -0.21    -0.00 1.00     6601     4653

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

14.2 Appendix B: Output of Logit Model with Logistic(0,3) Prior

```
Family: bernoulli
Links: mu = logit
Formula: Diabetes_binary ~ HighBP + HighChol + CholCheck + BMI + Smoker + Stroke
+ HeartDiseaseorAttack + PhysActivity + Fruits + Veggies + HvyAlcoholConsump + AnyHealthcare
+ NoDocbcCost + GenHlth + MentHlth + PhysHlth + DiffWalk + Sex + Age + Education + Income
Data: scaledData[sampleIDs, ] (Number of observations: 3000)
Draws: 4 chains, each with iter = 3000; warmup = 1500; thin = 1;
    total post-warmup draws = 6000

Regression Coefficients:
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept               -2.06      0.47    -3.02    -1.17 1.00    10195     4964
HighBP1                  0.67      0.09     0.49     0.86 1.00     8236     4781
HighChol1                0.56      0.09     0.38     0.74 1.00    10065     4563
CholCheck1               1.22      0.39     0.48     2.04 1.00     9781     4707
BMI                      0.56      0.05     0.45     0.67 1.00     8043     4975
Smoker1                  0.01      0.09    -0.17     0.19 1.00     8387     4885
Stroke1                  0.08      0.20    -0.30     0.47 1.00     8881     4564
HeartDiseaseorAttack1    0.28      0.15    -0.01     0.56 1.00     8969     4608
PhysActivity1           -0.20      0.10    -0.40     0.01 1.00     9523     4531
Fruits1                  0.01      0.09    -0.17     0.19 1.00     9807     4487
Veggies1                -0.11      0.12    -0.34     0.11 1.00     8404     4339
HvyAlcoholConsump1      -0.86      0.22    -1.29    -0.43 1.00     9856     4564
AnyHealthcare1           0.27      0.26    -0.24     0.78 1.00     8887     4452
NoDocbcCost1            -0.10      0.19    -0.47     0.29 1.00     8584     4977
GenHlth                  0.61      0.06     0.50     0.73 1.00     8187     5286
MentHlth                -0.06      0.05    -0.16     0.04 1.00     8742     4837
PhysHlth                -0.01      0.06    -0.12     0.10 1.00     8338     4708
DiffWalk1                0.09      0.13    -0.15     0.34 1.00     8227     5148
Sex1                     0.28      0.09     0.10     0.46 1.00     9153     4843
Age                      0.45      0.06     0.34     0.56 1.00     6934     5194
Education               -0.02      0.05    -0.12     0.08 1.00     9300     4984
Income                  -0.11      0.05    -0.22    -0.00 1.00     7936     5075

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

14.3 Appendix C: Output of Logit Model with Cauchy(0,3) Prior

```
Family: bernoulli
Links: mu = logit
Formula: Diabetes_binary ~ HighBP + HighChol + CholCheck + BMI + Smoker + Stroke
+ HeartDiseaseorAttack + PhysActivity + Fruits + Veggies + HvyAlcoholConsump + AnyHealthcare
+ NoDocbcCost + GenHlth + MentHlth + PhysHlth
+ DiffWalk + Sex + Age + Education + Income
Data: scaledData[sampleIDs, ] (Number of observations: 3000)
Draws: 4 chains, each with iter = 3000; warmup = 1500; thin = 1;
    total post-warmup draws = 6000

Regression Coefficients:
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept               -2.04      0.47    -3.00    -1.14 1.00     7403     4566
HighBP1                  0.67      0.10     0.49     0.85 1.00     8182     4754
HighChol1                0.56      0.09     0.38     0.74 1.00     8690     4501
CholCheck1               1.20      0.39     0.49     1.99 1.00     7669     4103
BMI                      0.56      0.05     0.45     0.66 1.00     6990     4601
Smoker1                  0.01      0.09    -0.18     0.19 1.00     8857     4293
Stroke1                  0.08      0.19    -0.29     0.46 1.00     7949     4588
HeartDiseaseorAttack1    0.28      0.14    -0.00     0.57 1.00     8038     5013
PhysActivity1           -0.19      0.11    -0.41     0.02 1.00     8177     4501
Fruits1                  0.01      0.09    -0.18     0.18 1.00     8965     5092
Veggies1                -0.11      0.12    -0.34     0.12 1.00     8856     4310
HvyAlcoholConsump1      -0.85      0.21    -1.28    -0.44 1.00     9074     4894
AnyHealthcare1           0.27      0.27    -0.24     0.79 1.00     7225     4755
NoDocbcCost1            -0.09      0.20    -0.48     0.28 1.00     7636     4780
GenHlth                  0.62      0.06     0.50     0.73 1.00     6603     4901
MentHlth                -0.06      0.05    -0.16     0.04 1.00     6839     4266
PhysHlth                -0.01      0.06    -0.12     0.10 1.00     6581     4586
DiffWalk1                0.10      0.13    -0.16     0.35 1.00     7330     4879
Sex1                     0.28      0.09     0.10     0.47 1.00     8229     4918
Age                      0.45      0.06     0.34     0.56 1.00     6728     5265
Education               -0.02      0.05    -0.12     0.08 1.00     6959     4715
Income                  -0.11      0.05    -0.22    -0.00 1.00     6438     4838

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1)
```

14.4 Appendix D: Output of Multilevel Logit Model with Normal(0,1) Prior

```
Family: bernoulli
Links: mu = logit
Formula: Diabetes_binary ~ CholCheck + HighBP + HighChol + Stroke + HeartDiseaseorAttack
+ HvyAlcoholConsump + AnyHealthcare + BMI + Age + Sex
+ Education + BMI * Age + (1 | GenHlth)
Data: scaledData[sampleIDs, ] (Number of observations: 3000)
Draws: 8 chains, each with iter = 3000; warmup = 1500; thin = 1;
    total post-warmup draws = 6000

Group-Level Effects:
~GenHlth (Number of levels: 5)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.21      0.54     0.54     2.62 1.00     1874     2860

Regression Coefficients:
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept               -1.97      0.72    -3.44    -0.55 1.00     2291     2413
CholCheck1               1.07      0.35     0.39     1.80 1.00     7622     4020
HighBP1                  0.67      0.10     0.48     0.85 1.00     7876     4005
HighChol1                0.56      0.09     0.38     0.73 1.00     9262     4448
Stroke1                  0.12      0.19    -0.24     0.49 1.00     8571     4177
HeartDiseaseorAttack1    0.29      0.14     0.02     0.56 1.00     8361     4785
HvyAlcoholConsump1      -0.86      0.21    -1.28    -0.44 1.00     8087     3671
AnyHealthcare1           0.22      0.24    -0.26     0.69 1.00     8350     4070
BMI                      0.57      0.05     0.47     0.68 1.00     7630     4836
Age                      0.49      0.05     0.39     0.59 1.00     7409     4637
Sex1                     0.24      0.09     0.07     0.41 1.00     8138     4677
Education               -0.08      0.05    -0.17     0.01 1.00     8551     3813
BMI:Age                  0.03      0.05    -0.08     0.12 1.00     9338     4141

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
Warning message:
There were 3 divergent transitions after warmup. Increasing adapt_delta above 0.8 may help.
```