Introduction

The correlation between mental health and chronic physical conditions is a critical area of research that has gained significant attention due to its implications on public health strategies and individual healthcare management. Mental health disorders such as depression and anxiety have been shown to influence the onset and progression of various chronic diseases, including diabetes. Through this project, we explore the potential link between how individuals perceive their mental health and the prevalence of diabetes, utilizing data sourced from the Behavioral Risk Factor Surveillance System (BRFSS) conducted by the Centers for Disease Control and Prevention (CDC) from 2012 to 2022.

Background and Data

Diabetes is recognized globally as a major public health issue due to its high prevalence and significant morbidity. The US faces extremely high diabetes rates, with over a third of the population having prediabetes and 11.6% being diagnosed with diabetes (CDC, 2023). The US also faces significant mental health issues involving depression, anxiety, and stress. More than one in five adults live with mental health issues, with these conditions growing even more extreme due to the COVID-19 pandemic (CDC, 2024). Mental health and physical health are equally important components of overall health. Studies have shown that the presence of chronic health conditions can increase the risk of mental illness.

This analysis will be utilizing the Behavioral Risk Factor Surveillance System dataset collected by the CDC to investigate the relationship between mental health and diabetes. This data has been collected annually since 1988 and has 390,000 to 480,000 participants per year. All information from the questionnaire is self reported from a simple random sample of the US population. This analysis will be using the variables DIABETE3 and MENTHLTH from 2012 to 2022 datasets. DIABETE3 refers to the responses for the question “Ever told you have diabetes.” This is a categorical variable with values 1 meaning “Yes” and 2 meaning “Yes, but female told only during pregnancy”. Response values 3,7, and 9 respectively means the participants were never told they had diabetes, didn’t know, or refused to answer. Response value 4 means they were diagnosed with prediabetes or borderline diabetes. The Variable MENTHLTH represents the responses to the question “Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?” The response values range from 1 to 30 to indicate the number of days the participant’s mental health was not good. If the participant didn’t experience any bad mental health days in the past 30 days, they responded with the response value 88.

Pre-Processing

After getting the data from the CDC, we selected only the columns DIABETE3 and MENTHLTH. For pre-processing the diabetes indicator, we decided that people who had diabetes during pregnancy, pre-diabetic, and borderline diabetic people would be considered diabetic for the purposes of this project. We removed the people who didn’t know whether they had diabetes or refused to answer the question, and we removed blanks/NAs. For pre-processing the mental health column, we removed people who didn’t know the number of days, who refused to answer, and blanks. After this, we categorized the number of days into 5 buckets. Anyone who answered None to the question for MENTHLTH we considered as 0 days (value 88). We then grouped the rest of the numbers:

- 1-7 Days: 1 week
- 8-14 Days: 2 weeks
- 14-21 Days: 3 weeks

- 22-30 Days: 3+ weeks

From this point on, we will refer to these groups from the labels we have given them here. Also, we use θ to denote the proportion of people with diabetes.

EDA

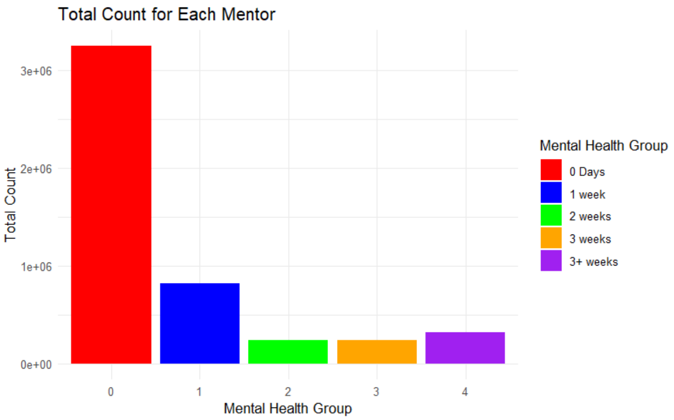


Figure 1: Number of observations for each group

In terms of observations, as we can see in Figure 1, we have much more data for 0 days than we have for 1-3+ weeks. However, the amount of data we have for the other groups is still in the hundreds of thousands, so there is enough data to conduct inference.

Taking an initial look at the sample proportions of the categories in the data, we can see a few trends that stand out.

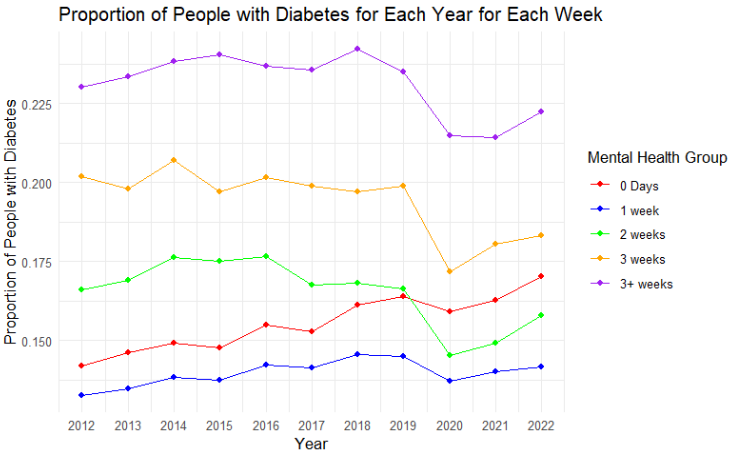
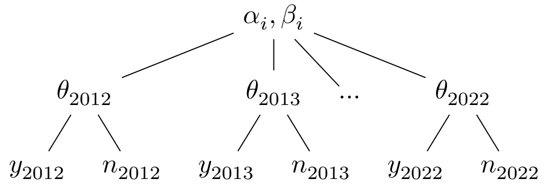


Figure 2: Sample proportions mapped over time for each group

Firstly, the sample proportion of people with diabetes in 3 weeks and 3+ weeks is always higher than the proportion of people with diabetes in 0 days, 1 week, and 2 weeks. Also, the proportion of people with diabetes in 0 days and 2 weeks is always higher than in 1 week. From 2020 to 2022, the proportion of people with diabetes in 0 days is higher than in 2 weeks. One major quantity of interest we are interested in for this project is the θ values for each of the categories for 2022. Given that we have so much data, we would expect that the true ranking of the proportion of people with diabetes in each of the categories in 2022 would closely follow the ranking of the sample proportions of these categories in 2022. We would also expect to have a higher degree of confidence in the true proportion of people with diabetes in 0 days because we have more data.

Methods/Model:

The model we chose is a beta-binomial hierarchical model. We fit a total of 5 of these models on the data subsetted by an individual's mental health status category. The main parameter of interest is $\theta_j (j \in [2012 : 2022])$ within each model, denoting the probability of diabetes. We also have hyperparameters $\alpha_i, \beta_i, j \in [0\ days, 1\ week, 2\ weeks, 3\ weeks, 3+\ \ weeks]$ within each model, which allows for comparison of posterior mean densities between the groups. We used a proper, uninformative prior for α and β, since we have numerous observations per mental health group, so the weight of the likelihood will far outweigh the prior. A visualization of the hierarchical structure is shown below:



$i \in$ [0 days, 1 week, 2 weeks, 3 weeks, 3+ weeks]

Given these structures, we ran five different STAN models to perform an MCMC algorithm. We conduct inference using the samples from these models to assess the difference in the probability of diabetes across different mental health categories. We interpret confidence intervals to assess the significance of the difference of probability of diabetes between each possible mental health group pairing. The samples will also allow us to visualize posterior distributions of θ, specifically focusing on the year 2022 across the different mental health groups. Running STAN on these models will also allow us to compare posterior means between mental health groups due to sampling of hyperparameters.

Diagnostics

Given our methodology relies on MCMC algorithms, it's crucial to perform diagnostic tests prior to interpreting any results. To be certain the algorithm provides reliable analysis, trace plots will provide insight on whether the models converged. The following 5 trace plots are from all five of our models.

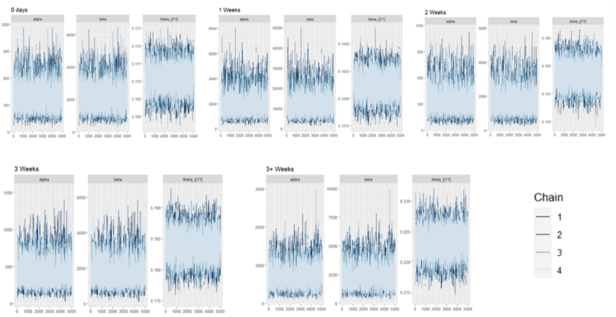
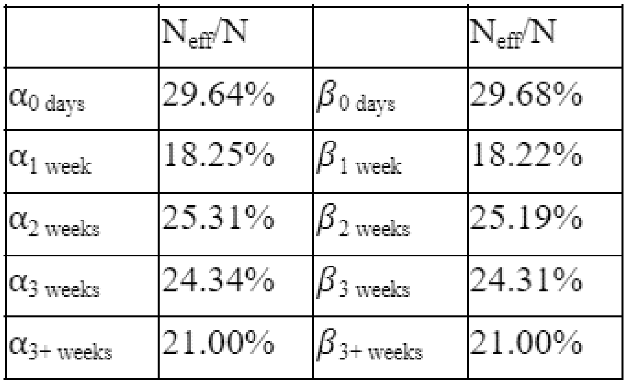


Figure 3: Trace Plots for each STAN model with 4 chains and 10,000 iterations

All trace plots indicate that our chains converged and exhibit good mixing.



	N_{eff}/N		N_{eff}/N
$\alpha_{0\ days}$	29.64%	$\beta_{0\ days}$	29.68%
$\alpha_{1\ week}$	18.25%	$\beta_{1\ week}$	18.22%
$\alpha_{2\ weeks}$	25.31%	$\beta_{2\ weeks}$	25.19%
$\alpha_{3\ weeks}$	24.34%	$\beta_{3\ weeks}$	24.31%
$\alpha_{3+\ weeks}$	21.00%	$\beta_{3+\ weeks}$	21.00%

Figure 4: Effective sample size over N for our alpha and beta parameters for each group

Visually inspecting Figure 4 above, the acceptance rate is also around 25%, which is supporting evidence that the chain converged as well. Some of them are lower than 25%, especially $\alpha_{1week}, \beta_{1week}$ suggesting that the hyperparameters for 1 week have lower ESS than the other hyperparameters for the other models. This could be due to higher autocorrelation at these points or from temporal dependence, which leads to less effective exploration of the sample space. As viewed in the Figure 4, the ESS for $\alpha_i, \beta_i, i \in [0$ days,2 weeks,3 weeks] is large enough as $\frac{N_{eff}}{N}$ is close to or above 25%, but the concern of the $\frac{N_{eff}}{N}$ ratio of both $\alpha_j, \beta_j, , j \in [1$ week,3+ weeks] being >4% less than 25% brings up concern of higher autocorrelation between the samples or a temporal dependence structure within the model for mental health groups 1 week and 3+ weeks.

Results

A major parameter of interest from our results is the posterior mean. We generated histograms for the posterior mean from our samples in the figure below:

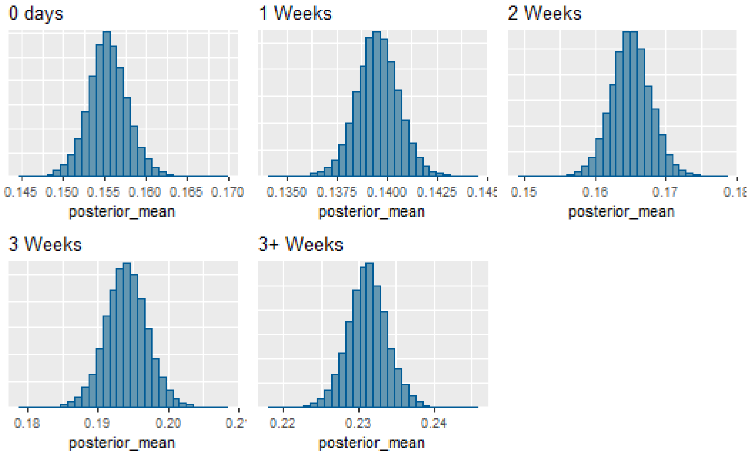


Figure 5: Histograms of the posterior mean samples from each of the hierarchical models.

We see from Figure 5 that these posterior means follow a normal distribution, which we would expect because the chain converged.

To more easily see the difference between these posterior means, we can see the trend of the expected value of the posterior means in Figure 6 below:

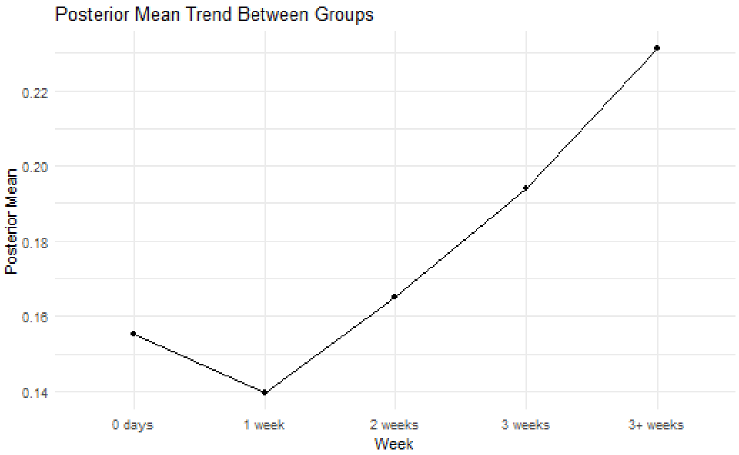


Figure 6: Comparison of the expectations of the posterior means for each of the mental health groups.

The posterior mean decreases from 0 days to 1 week, then gradually increases with the largest posterior mean being 3+ weeks. The graph is a representation of the increasing trend between worsening mental health and probability of diabetes, taking into account data from all years in our analysis.

Our second major parameter of interest is the year 2022 because it is the most recent and the most representative of what we would see in 2023 and 2024. We generated five posterior density plots for each mental health group in 2022 to visualize the estimated distributions of our θ's. We generated this by analytically deriving the beta function for each group in 2022. As an example, here is the posterior density function of theta for 0 days for 2022:

$$P(\theta_{2022}|\alpha_{0\ days}, \beta_{0\ days}, y_{0\ days,2022}, n_{0\ days,2022}) \propto (\theta)^{45084+239.8022-1}(1-\theta)^{264695-45084+1302.532-1}$$

After deriving this for each of the categories for 2022, we plotted them below:

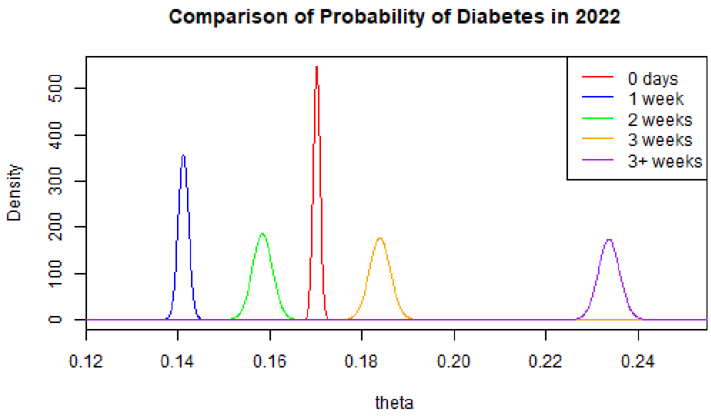


Figure 7: Density plot of the Probability of Diabetes in 2022 for each group

We can see see a rank order similar to the sample proportions for 2022 from Figure 7 above. This is in contradiction to the overall trend of the data where the order is 1 week, 0 days, 2 weeks, 3 weeks, 3+ weeks. This is expected because in the beta-binomial hierarchical model we would expect to see the information from 2022 show through more than the information we gain from 2012-2021. We also see a higher degree of confidence in the groups that have more data, so 0 days and 1 week are significantly more dense than the distributions for 2, 3, and 3+ weeks. We also see almost no overlap, which means we are confident in our findings that these groups are different. We developed confidence intervals for the differences in the groups, which we see in Figure 8 below:

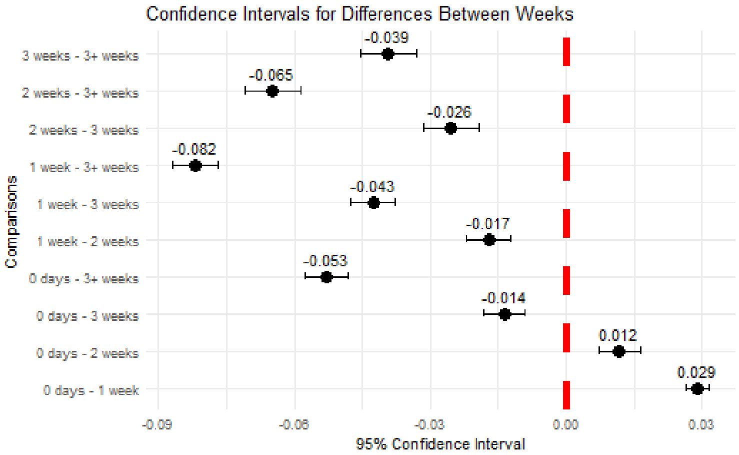


Figure 8: 95% confidence intervals for the difference in the thetas for 2022 for each group

Figure 8 visualizes the confidence intervals for the difference of θ's between groups in 2022, on the y-axis having the 10 possible pairings of the mental health groups. The x-axis denotes a 95% confidence interval for the difference of θ, where the confidence interval is constructed with:

$$E(\theta_i) - E(\theta_j) \pm 1.96 \cdot \sqrt{Var(\theta_i - \theta_j)} =$$
$$\mu_i - \mu_j \pm 1.96\sqrt{Var(\theta_i) + Var(\theta_j) - 2Cov(\theta_i, \theta_j)}$$
$$i, j \in [0 \text{ days, } 1 \text{ week, } 2 \text{ weeks, } 3 \text{ weeks, } 3+ \text{ weeks}]$$

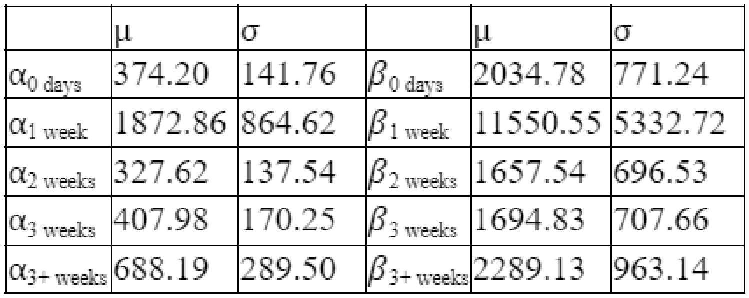
1.96 is a reasonable critical value to use for these 95% confidence intervals as the sample distribution of each θ in its respective mental health group is approximately normally distributed due to the fact that the chains converged and the trace plot diagnostics show the chains mixing well.

The method of interpreting these confidence intervals is considering $\theta_{i,2022} - \theta_{j,2022}$, where θ_i is the θ for the mental health group on the left side of the comparison y-axis and θ_j is the θ for the mental health group on the right side of the comparison y-axis. Most of the confidence intervals cover areas that are strictly less than 0, meaning that group i's θ is less than group j's θ with high confidence. Similarly, if the confidence interval is to the right of zero, group i's θ is greater than group j's θ with higher confidence. None of the confidence intervals cover 0 likely due to our large ESS values from each model, meaning we can conclude that there is a significant disparity between the sampled θ's in each mental health group pairing for the year 2022.
The most extreme negative difference is between 1 week and 3+ weeks, with the mean difference being -0.082 and the 95% confidence interval being [-0.087, -0.077]. This result suggests that worse mental health has a significant impact on the proportion of diabetes within the population in the year of 2022.

Discussion

The results from our model can be stated with a lot of certainty due to the large sample size from the BRFSS dataset. However, several limitations must be noted about our methodology and data. The data collected has inherent bias due to the nature of the survey. Asking participants to respond about their health conditions and mental illness can inherently lead to conformity bias. Additionally, the dataset suffers from selection bias due to the self reported nature of the data collection. The data relies on respondent reliability and their willingness to report sensitive information. The dataset will inevitably lack information from those who are reluctant to answer the questionnaire in its entirety, which causes an incomplete or biased representation of the population.

Another limitation we considered was the possibility that the prevalence of diabetes in these populations might be temporally dependent.



	μ	σ		μ	σ
$\alpha_{0\text{ days}}$	374.20	141.76	$\beta_{0\text{ days}}$	2034.78	771.24
$\alpha_{1\text{ week}}$	1872.86	864.62	$\beta_{1\text{ week}}$	11550.55	5332.72
$\alpha_{2\text{ weeks}}$	327.62	137.54	$\beta_{2\text{ weeks}}$	1657.54	696.53
$\alpha_{3\text{ weeks}}$	407.98	170.25	$\beta_{3\text{ weeks}}$	1694.83	707.66
$\alpha_{3+\text{ weeks}}$	688.19	289.50	$\beta_{3+\text{ weeks}}$	2289.13	963.14

Figure 9: Expectations and standard deviations of our alpha and beta samples from the MCMC sampling

The table above in conjunction with figure 4 from the diagnostics section reemphasizes the concern of high autocorrelation between the samples in two of the models, model for week 1 and week 3. There is likely temporal dependence within these models as their ESS for the hyperparameters is smaller in comparison to the other models, as well as the standard deviation of both hyperparameters being much greater than the other models. This indicates that there is much overlap in information between the years and the addition of information from 2012-2021 significantly impacts the estimation of θ for 2022, implying that $\text{P}(\theta_{2022}|\theta_{2012-2021})$ is not equal to $\text{P}(\theta_{2022})$.

Conclusion

We explored the relationship between self-perceived mental health and the prevalence of diabetes using data from the BRFSS collected between 2012 and 2022. By applying a beta-binomial hierarchical model, we were able to effectively draw samples from the posterior distribution and conduct inference. This inference concluded that there is a significant difference of the probability of diabetes between the mental health groups in 2022, especially with higher, positive differences between the groups with severe mental health issues and groups with moderate to no mental health issues.

The findings from our analysis provide insight and direction on what can be done in the healthcare industry in the future. Physicians and other healthcare professionals can leverage the statistical findings from our analysis to improve patient care. These findings can lead to integrating mental health treatment into diabetes treatment and vice versa. Understanding the link between these health concerns can create better treatment plans for future patients affected by diabetes and mental health issues. This further promotes the need for holistic medicine and the importance of treating the whole person, rather than focusing on specific symptoms.