2023

Risk Analysis of Clostridiodides Difficile Infections in a Hospital Setting and the Impact of Prior Choice on Predictive Capability

Trevor D. Blanchard

Risk Analysis of Clostridiodides Difficile Infections in a Hospital Setting and the Impact of Prior Choice on Predictive Capability

A thesis submitted to
The University of Vermont
Department of Statistics

by

Trevor Blanchard

Under the guidance of

Professor Jean Gabriel Young

THE UNIVERSITY OF VERMONT
COLLEGE OF ENGINEERING &
MATHEMATICAL SCIENCES

College of Engineering and Mathematical Sciences
Department of Mathematics and Statistics
University of Vermont
Burlington, Vermont

Risk Analysis of Clostridiodides Difficile Infections in a Hospital Setting and the Impact of Prior Choice on Predictive Capability

Trevor Blanchard[1]

[1]Department of Mathematics and Statistics, University of Vermont

Abstract

Abstract Healthcare-associated *Clostridiodides Difficile* (C. diff.) infections are one of the most common healthcare associated infections in the U.S., leading to thousands of deaths per year. Machine learning algorithms have shown some ability to predict who is most vulnerable to C. diff. infection utilizing electronic health records obtained soon after admittance, but these models have shown insufficient predictive capability. We extracted data from the electronic medical records provided in the MIMIC-III Clinical Database which contains data from the Beth Israel Deaconess Medical Center between 2001 and 2012, resulting in very large predictor matrices. We aimed to predict which patients would receive a positive test for C. diff. using a Bayesian logistic regression model. We examined the impact of three different priors, a normal, double exponential, and regularized horseshoe prior to understand how prior choice influenced predictive capability and the size of coefficients. We used cross-validation to test the predictive capability of each prior, and compared results between models using ROC and PR curves. Our results show that of the three priors, the regularized horseshoe prior achieves the highest prediction accuracy.

Keywords: *Bayesian, logistic regression, normal prior, double exponential prior, regularized horseshoe prior, FIDDLE, C. diff.*

1. Introduction

Clostridiodides Difficile (C. diff.) is a contagious spore-forming bacterium that mainly lives on and spreads via surfaces. The main symptoms of C. diff. infections (CDI) are diarrhea and colitis, and although not incredibly dangerous on an individual basis, its contagiousness and ability to live on surfaces for extended periods of time can make it a big problem in healthcare settings. In fact, C. diff. is the most contracted germ-related infection in healthcare, accounting for 12% of all germ-related infections in healthcare settings (CDC, 2014). As a result, there were roughly 235,000-313,000 annual healthcare-related infections from 2011-2017, which led to 16,000-25,000 annual deaths (Guh et al., 2010).

Hospital-associated infections tend to impact people who are already of poor health who are also sharing the same spaces. This gives C. diff. the opportunity to spread easily and quickly, which, in severe cases, can lead to the shutdown of certain sections of a hospital (Blackwell, T., 2015). Furthermore, CDI in hospitals leads to significant increases in health care costs, so preventing the spread of C. diff. will save not only lives, but also money and other hospital resources. From

2005-2015 it is estimated that U.S. hospitals spent between 1.9 and 7 billion dollars which involved roughly 2.4 million days of stay in the hospital (Zhang, 2016). Hospital resources have also become much more limited with the increased demand for health care due to COVID, so reducing CDI can stand to free up critical resources. Fortunately, preventative measures can alter the course of the infection so catching infections early or better yet, preventing infections, is critical.

An important component of limiting the spread of C. diff. is being able to predict who is most risk of contracting the infection. Thus, the goal of this thesis is to model who is most at risk for C. diff. infection by learning which health factors leave patients most vulnerable to contracting the infection. We use past electronic health records from the Beth Israel Deaconess Hospital in Boston, MA, which have recorded included numerous health outcomes including positive C. diff. cases. Using this data, we aim to use statistical machine learning to generate a model trained using a labeled dataset. The training dataset should be able to teach our model which health factors are most influential in determining if a patient is at risk for contracting CDI. From here, we will be able to test whether these factors are good predictors of C. diff. using the testing dataset. Using the testing dataset, the model will predict which patients tested positive for C. diff. which can then be compared to the actual data.

Due to the threat C. diff. can pose to hospitals, there have been multiple studies that have attempted to use machine learning to predict C. diff. cases. Li et al. studied 1,144 cases of C. diff. from October 2010 to January 2013, the data for which was acquired from the University of Michigan hospitals (Li et al., 2019). In their study, they used ℓ^2 regularization and k-best feature selection to predict C. diff. cases the day of, the day after, and two days after diagnosis (Li et al., 2019). Results were compared between a model which used all EHR and a curated model which contained a set of variables selected manually. In all instances, the model which used the EHR data outperformed the curated model and the best predictions were made two days after diagnosis, with results worsening as the time of prediction approached the day of diagnosis (Li et al., 2019). The results from this study show how in our study, using all variables present in EHR can improve our ability to predict cases of CDI.

Another study conducted by Ng et al. used data from 41 hospitals, which resulted in a study of over 15,000 C. diff. infections. The goal of this study was not to predict which patients would be infected by CDI. Rather, the goal of the study was to use machine learning to create a logistic regression model that could predict which patients were most at risk for dying from CDI and which patients were most likely to have a recurring infection soon after the initial recovery (Ng et al., 2021). Their study found that the logistic regression model performed best when attempting to predict which patients ultimately died because of C. diff. compared to predicting recurrence. In addition, it was observed that these models did not see much difference in prediction results when using all features versus the top 18 identified features (Ng et al., 2021). Since the ultimate goal of this study is different from ours, we are not concerned by the fact that the results from this study contradict the findings in the previous study regarding the benefit of using all available predictors rather than a curated set of variables. This study by Ng et al. is important to us because it demonstrates the effectiveness of using logistic regression when predicting cases of CDI.

Wiens et al. used ℓ^2-regularized logistic regression in their study which aimed to predict positive C. diff. cases. They used data from a large urban U.S. hospital where patient stays were greater than 24 hours from April 2011 to April 2013, resulting in 69,568 admissions and 727 cases of C. diff. (Wiens et al., 2014). Similar to the two studies described above, Wiens et al. wanted to see if using all features from EHR resulted in improved predictions compared to a curated set of known risk factors for CDI. The results from the study showed roughly a 10% improvement in the Area Under the Receiver Operating Characteristic (AUROC) curve when comparing the EHR model to the curated model (Wiens et al., 2014). Even though, like both studies described above, the results obtained by Wiens et al. were unable to perfectly classify cases of C. diff., using these models could significantly reduce the number of C. diff cases in hospital settings. In addition, Wiens et al. and Li et al. have shown the benefits of using all features from EHR instead of commonly known risk factors for CDI. The study by Wiens et al. shows us how using logistic regression and the entire set of predictors present in EHR can be effective in predicting cases of CDI. Overall this study combines the findings in the previous two studies and we aim to improve upon the study done by

Wiens et al. via the introduction of priors and using ℓ^1 regularization. Through the use of priors and ℓ^1 and ℓ^2 regularization, we will be able to use the same general logistic regression model but alter the way in which the model makes predictions and weights certain variables.

2. Project Goal

Although medical professionals have developed a list of health factors that leave individuals vulnerable to contracting C. diff., using all of the variables present in electronic health records could significantly improve the prediction of C. diff. infections. This project uses all available variables present in electronic health records gathered within 24 hours of admittance time from the MIMIC-III Clinical Database from the Beth Israel Deaconess Medical Center between 2001 and 2012. With this data, we plan to implement a Bayesian logistic regression model capable of predicting which patients will contract CDI. Furthermore, using Bayesian logistic regression we will also be able to observe which health factors the model found were most important in arriving at its predictions.

3. Data Preparation

Table 1: Datasets (Johnson et al., 2016)

Dataset	Description
DIAGNOSES_ICD	Maps subject ID and HADM ID to ICD9 code
PATIENTS	Contains subject ID, Gender, date of birth, and date of death
ADMISSIONS	Contains categorical patient information and time and type of admission
ICUSTAYS	Maps subject ID and HADM ID to ICUSTAY ID and length of stay in ICU
CHARTEVENTS	Contains item ID, measured value, and charted time
LABEVENTS	Contains laboratory test results for a patient
MICROBIOLOGYEVENTS	Contains type of specimen collected
OUTPUTEVENTS	Contains all measurements related to output for a given patient
INPUTEVENTS_MV	Contains inputs for patients monitored with the Metavision system
PROCEDUREEVENTS_MV	Contains information regarding the start and stop time for various procedures for Metavision patients
DATETIMEEVENTS	Contains charted time mapped to subject ID, HADM ID, and ICUSTAY ID

Prior to creating the model, we needed to transform the data into feature vectors. The data we used is in the MIMIC-III Clinical Database, which is a free database of cases admitted to the Beth Israel Deaconess Medical Center between 2001 and 2012 with over forty-thousand patients (Johnson et al., 2016). Of the 26 files provided, only 11 were necessary for our analysis and can be found in Table 1 (Tang et al., 2020). To combine and format these 10 datasets (excluding DIAGNOSES_ICD) to make feature vectors, we used FIDDLE. "FIDDLE (Flexible Data-Driven Pipeline) [. . .] systematically transforms structured EHR data into representations that can be used as inputs to ML algorithms" (Tang et al., 2020, pp. 1922). FIDDLE is free to use, and all the scripts can be found on GitHub. The authors of FIDDLE included an example for how to use FIDDLE specifically for the MIMIC-III dataset, which is what we used to clean and format our data. Once our data was cleaned, we used the dataset Diagnoses_ICD to discern which patients contracted CDI. From this dataset, we extracted all rows which contained the ICD9 code for C. diff., 008.45 (Dubberke et al., 2006). ICD codes are the official codes used for diagnoses and procedures in U.S. hospitals (CDC, 2021). With this subset of data, we created a column in our data that encoded our target variable, CDI, during admission. Since our goal is to predict which patients will contract C. diff in the hospital, we only used data gathered in the first 24 hours of each patient's stay. Unfortunately, using the ICD9 codes in the Diagnoses_ICD dataset, we cannot obtain an accurate time for when patients tested positive for C. diff., thus we cannot remove cases contracted prior to the patient was admitted to the hospital.

4. Methods

To model the relationship between our explanatory variables, X, and our response variable, y_i, we used logistic regression. The likelihood function for our Bayesian logistic regression model is

$$P(X, y|\beta) = \prod_{i=1}^{n} \left[\left(\frac{e^{\beta_0+\beta_1 X_{i,1}+...+\beta_k X_{i,k}}}{1+e^{\beta_0+\beta_1 X_{i,1}+...+\beta_k X_{i,k}}} \right)^{y_i} \left(1 - \frac{e^{\beta_0+\beta_1 X_{i,1}+...+\beta_k X_{i,k}}}{1+e^{\beta_0+\beta_1 X_{i,1}+...+\beta_k X_{i,k}}} \right)^{1-y_i} \right]$$

where X denotes our n x k matrix of explanatory variables for each patient, β represents a vector of coefficients for each variable, and $y_i = 1$ if a patient tests positive for C. diff. and is 0 otherwise. We denote the sum of $\beta_0 + \beta_1 X_{i,1} + ... + \beta_j X_{i,j}$ as η for simplicity. To obtain our posterior, we need a prior. With our Bayesian logistic regression model, we experimented with three different priors; a normal prior, a regularized horseshoe prior, and a double exponential prior. Our motivation for using multiple priors is to decipher which form of regularization produces the best results. There are two types of regularization that we experimented with, ℓ^1 and ℓ^2. Regularization is important because it helps prevent over-fitting from machine learning algorithms by incorporating some type of penalty into the predictions (Pykes, 2022). We can observe how choice of prior impacts which form of regularization we use by examining the log-posterior.

$$\log(P(\beta|X, y)) = \log(P(X, y|\beta)) + \log(P(\beta))$$

the latter term, $\log(P(\beta))$, serves as the penalty term and is determined by the prior we set on β. If we consider a normal prior on β, for simplicity we set $\mu = 0, \sigma = 1$,

$$\begin{aligned}
&P(\beta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\beta_i^2} \\
&\log(P(\beta)) = n\log(\frac{1}{\sqrt{2\pi}}) + -\frac{1}{2}\sum_{i=1}^{n}\beta_i^2 \\
&\propto -\sum_{i=1}^{n}\beta_i^2
\end{aligned}$$

Since the penalty term arising from a normal prior on β penalizes the squared weights, the 2-norm of the vector of weights, a normal prior produces ℓ^2 regularization.

Similar to the normal prior, the horseshoe prior also sets a normal distribution on β. Therefore it appears that the horseshoe prior produces ℓ^2 regularization (Larionov, 2019).

$$\begin{aligned}
&P(\beta) = \left(\frac{1}{\sqrt{2\pi\tau}}\right)^n \prod_{i=1}^{n}\left(\frac{1}{\lambda_i}\right) e^{-\frac{1}{2}\sum_{m=1}^{n}\beta_i^2/\lambda_i^2} \\
&n\log(\frac{1}{\sqrt{2\pi\tau}}) + \log\left(\prod_{i=1}^{n}\left(\frac{1}{\lambda_i}\right)\right) - \frac{1}{2}\sum_{m=1}^{n}\beta_i^2/\lambda_i^2 \\
&\propto -\sum_{m=1}^{n}\beta_i^2/\lambda_i^2
\end{aligned}$$

However, unlike the normal prior, where the variance is constant, we have a hyper-parameter set on λ_i which takes the form of a Half-Cauchy distribution (Larionov, 2019). More specifically, $\lambda_i \sim \text{Cauchy}^+(0,1)$ which has the PDF

$$f(\lambda_i;\mu,\sigma) = \frac{2}{\pi\sigma}\frac{1}{1+(\lambda_i-\mu)^2/\sigma^2} \rightarrow f(\lambda_i;0,1) = \frac{2}{\pi}\frac{1}{1+(\lambda_i)^2} \propto \frac{1}{1+(\lambda_i)^2} = k_i$$

(Bois, 2019). Now, $k_i \,\epsilon\, [0,1]$ and when we plot k, we get Figure 1, which is a density curve of all k_i which has a non-zero probability mass on zero (Larionov, 2019).

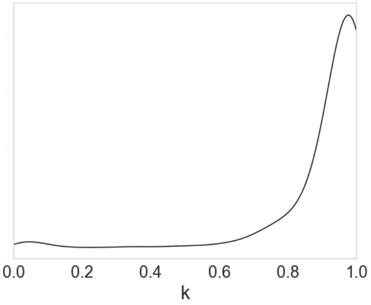

Figure 1. *distribution of k_i*

Consequently, we have situations where $k_i \approx 1$ which results in complete shrinkage of the coefficients and situations where $k_i \approx 0$ which results in no shrinkage of the coefficients (Larionov, 2019). Since some coefficients are completely shrunk towards zero while some coefficients remain unshrunk, the horseshoe prior produces regularization most similar to ℓ_1 regularization.

When we use a standard double exponential prior on β, we obtain a more explicit definition of ℓ^1 regularization.

$$P(\beta) = \left(\frac{1}{2}\right)^n e^{-\sum_{i=1}^n |\beta_i|}$$

$$n\log(1/2) - \sum_{i=1}^n \beta_i^2$$

$$\propto -\sum_{i=1}^n |\beta_i|$$

Since the penalty term arising from a double exponential prior on β penalizes the absolute value of the weights, the 1-norm of the vector of weights, a double exponential prior produces ℓ^1 regularization.

4.1. Fake Data Simulation

To get a rough idea of how different priors would impact the sizes of coefficients we ran a small simulation using fake data. We developed the plots shown in Figure 2 which track how the coefficients change based on the choice of prior in two scenarios, one in which we generated response variables with all but one coefficient being approximately zero and one with half of the coefficients approximately zero and the other half set to a large value.

From these graphs, we can see the regularized horseshoe prior favors sparsity the most compared to all priors and the double exponential and normal priors produce similar coefficients. We only used 10 covariates and 500 randomly sampled points using Normal $\sim (0,5)$ for our simulation, so we do not expect our final distribution of coefficient sizes to be identical, but it gives us a rough guide to how our choice of prior will impact each model.

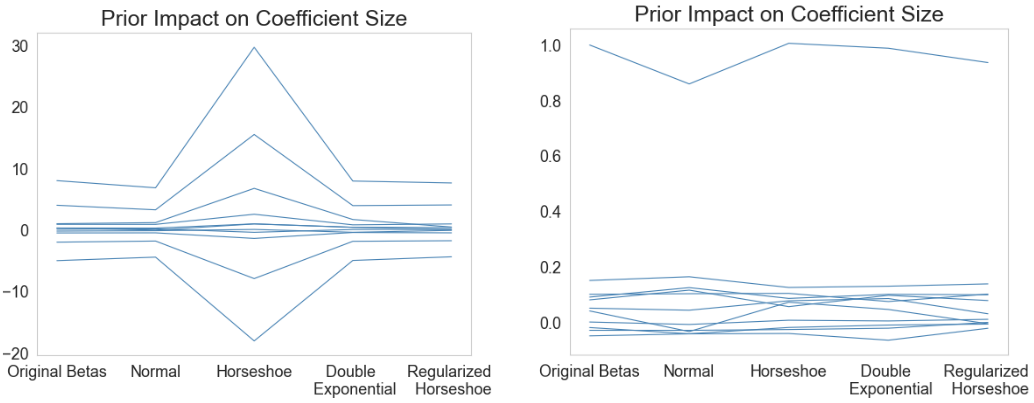

Figure 2. *Impact of prior on coefficient size*

4.2. Formulation

The first full model we used had a normal prior which has the following formulation

$$\begin{aligned} Y &\sim \text{BernoulliLogit}(\eta) \\ \beta &\sim \text{Normal}(0, 2^2) \end{aligned}$$

Since the normal prior incorporates ℓ^2 regularization into its predictions, the coefficients will be shrunk by a certain degree, but not all the way to zero (Pykes, 2022). This creates non-sparse output (many variables have small weights) (Pykes, 2022).

The regularized horseshoe prior is a modification of the horseshoe prior. As stated in Section 4 above, the horseshoe prior has the following formulation

$$\begin{aligned} \beta &\sim \text{Normal}(0, \tau^2\lambda_j^2) \\ \lambda_j &\sim \text{Cauchy}^+(0, 1) \\ \tau &\sim \text{Cauchy}^+(0, 1) \end{aligned}$$

A horseshoe prior which, as we stated before, produces ℓ^1 regularization and is a sparse prior meaning it amplifies the effects of a small number of variables and shrinks, or diminishes, the effect of the remaining variables. In the horseshoe prior, λ is known as the local shrinkage parameter and τ is referred to as the global shrinkage parameter (Carvalho et al., 2009). Setting Cauchy$^+$ priors on both of the shrinkage parameters allows the most important variables to experience no shrinkage while the least important variables are shrunk to 0 (the value of the individual $\beta_i s = 0$). Since the horseshoe prior generally only produces a few nonzero coefficients, the prior includes a near-infinite weight on 0 (Figure 3). This aspect of the prior is generally thought of as one of its strengths, but it can also be problematic when paired with logistic regression because of the flat likelihood function created by separable data (Piironen & Vehtari, 2017). The fact that some extremely large coefficients will remain unshrunk (the value of some $\beta_i s$ remains large) can lead to the vanishing of posterior means of the β_i (Piironen & Vehtari, 2017). To prevent the posterior means of the β_i from vanishing, Piironen and Vehtari developed the regularized horseshoe prior.

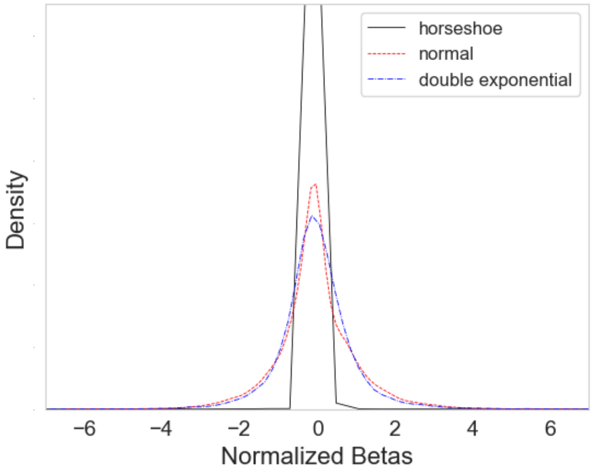

Figure 3. *Distribution of the size of $\beta_i s$, with a near infinite weight on $\beta_i \approx 0$ (Carvalho et al., 2009)*

The regularized horseshoe prior has the following formulation

$$\begin{aligned}
&\beta \sim \text{Normal}(0, \tau^2 \tilde{\lambda}_j^2)\\
&\tilde{\lambda}_j = \frac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2}\\
&\lambda_j \sim \text{Cauchy}^+(0,1)\\
&\tau \sim \text{Student T}(1, 0, \tau_0)\\
&\tau_0 = \frac{p_0 \sigma}{(D - p_0)\sqrt{n}}
\end{aligned}$$

In order to prevent the posterior means of β_i from vanishing, even the largest coefficients are shrunk slightly, or regularized, while the smallest coefficients are still shrunk toward zero (Piironen & Vehtari, 2017). This is accomplished by the introduction of $\tilde{\lambda}_j$. With $\tilde{\lambda}_j$, if $\tau * \lambda_j$ is much smaller than c, the regularized horseshoe prior will approach the horseshoe prior because $\tilde{\lambda}_j$ will approach λ_j, and thus the β_i are still shrunk to zero. However, when c is much smaller than $\lambda_j * \tau$, $\tilde{\lambda}_j$ approaches c^2/τ^2, thus $\beta_i \sim \text{Normal}(0, c^2)$. Therefore, the larger betas experience some shrinkage but are not pushed to zero (Piironen & Vehtari, 2017).

The regularized horseshoe also improves the prior for τ. Previous methods have used the prior $\tau \sim \text{Cauchy}^+(0,1)$. Piironen & Vehtari demonstrate, however, that doing so puts too much emphasis on larger and unshrunk coefficients. Instead, Piironen & Vehtari recommend using a hyper-parameter for τ, $\tau_0 = \frac{p_0 \sigma}{(D-p_0)\sqrt{n}}$ which is determined by the data. p_0 is a prior guess for the number of non-zero coefficients. In our study, we set $p_0 = 10$ due to the assumption that we would obtain roughly 10 influential coefficients while utilizing the horseshoe prior. D is the number of covariates, n is the number of samples, and $\sigma = \frac{1}{\sqrt{E(Y)*(1-E(Y)}}$ (Piironen & Vehtari, 2017). Doing so will result in a value much less than 1 for τ_0, promoting the shrinking of more coefficients while still allowing some larger values to escape while being loosely regularized.

Finally, the next prior we used was the double exponential prior.

$$\beta \sim \text{DoubleExponential}(0, 8^2)$$

The double exponential prior is an explicit use of ℓ^1 regularization, but the values of the resulting coefficients will be quite different than the regularized horseshoe prior. Like the regularized horseshoe prior, the double exponential prior will also shrink the majority of the coefficients to zero. However, instead of having only a few larger coefficients, the double exponential prior will have numerous larger coefficients as well. In return, more variables are important, but it is more difficult to discern which are truly the most impactful.

Through our process of predicting C. diff. infections with the first twenty-four hours of electronic health records, we aim to find whether ℓ^1 or ℓ^2 regularization, or sparse or non-sparse priors produce the best results for sparse data.

4.3. Model Implementation into STAN

We used STAN to create each of our models, all of which can be found in Appendices A-C. STAN uses a Hamiltonian Monte Carlo (HMC) algorithm, which is a Markov chain Monte Carlo (MCMC) method, to generate samples encompassing the posterior (Stan Development Team, n.d.). For the regularized horseshoe prior, we set the algorithm to run 4 chains with 1000 warm-up samples and 1000 samples. For the normal and double exponential priors, we set the algorithm to run 4 chains with 800 warm-up samples and 800 samples. The reason for the differences in the number of samples is because the models using a normal and double exponential prior took much longer to run. Since we were allotted a limited amount of time to run our models we had to reduce the number of samples so the models using the normal and double exponential priors could complete at least one chain. Each chain is independent, so having all 4 chains finish running is not essential for obtaining results. After our models finished running, we checked for issues regarding divergences and large R hats for the coefficients, and across all models, 96% of R hats were equal to 1 and there were no divergences (McElreath, 2016).

5. Results

After applying FIDDLE, our dataset contained 19,596 patients, 693 cases of CDI, and 11,489 explanatory variables. However, we wanted two datasets, one that contained all predictors and one with specific variables removed to reduce the effect of reverse causation. This is because certain variables could be present in our data because doctors had already diagnosed a patient with C. diff., or had a strong suspicion that the patient had C. diff., and were taking actions to combat the infection. After talking with clinical experts, we removed the antibiotics vancomycin, metronidazole, and clindamycin. In addition, we removed variables associated with dextrose and stool samples that were found to be influential in the testing of our model. Our resulting dataset contained 11,416 predictors, allowing us to examine the effectiveness of our model without these variables.

For each of our three priors, we used cross-validation in order to create a platform where we could both evaluate the individual performances of each of our models and compare their relative performances. We chose to divide the original datasets into training and testing datasets with an 80-20 split. For our model with removed features, we had a training dataset with 15,676 patients, which included 553 cases of C. diff, and a testing dataset with 3,920 patients and 140 cases of C. diff. For our model with all features, we had a training dataset with 15,676 patients, which included 556 cases of C. diff, and a testing dataset with 3,920 patients and 137 cases of C. diff.

Using the information learned in the training portion, each model calculates a probability that the patient contracted CDI. By varying the classification probability, which can range anywhere from 0 to 1, where we denote whether or not a patient will contract C. diff., we are able to compare performance across each model using receiver operating characteristic (ROC) curves and precision-recall (PR) curves. In addition to comparing model performance, we also compare the most important health factors related to CDI identified by each of our models.

5.1. Priors and Impact on Results

For our model using all available predictors present in the EHR, we find the ten most important variables for indicating positive and negative C. diff. cases for each prior in Table 2 and the resulting area under ROC curves and area under the precision-recall (AUPR) curves for each prior in Table 3.

For our model with removed features, we find the ten most important variables for indicating positive and negative C. diff. cases for each prior in Table 4 and the resulting AUROC and AUPR curves for each prior in Table 5.

We use ROC curves as a means to compare our results with the results from other studies, but precision-recall curves to evaluate our model performance. Precision-recall curves have precision on the y-axis and recall on the x-axis and ideally, the area under this curve is 1 meaning the model was able to recall all "positive" points with perfect precision. Precision-recall curves better suit our data because they are meant for studies attempting to detect rare events (Tran-The, 2021). This is because precision-recall curves are not affected by data imbalances such as a comparatively large amount of negative C. diff. cases compared to positive C. diff. cases. They accomplish this by ignoring the true negative rate and varying the classification threshold until all positive cases are identified (Tran-The, 2021). In each figure you will also see black dotted lines. These lines indicate random baselines, or the area under the curve we would expect if our model made random predictions.

Table 2: Positive and Negative Betas for Each Prior (all predictors)

	Normal Prior Positive Betas	Value (95% C.I.)	**Normal Prior Negative Betas**	Value (95% C.I.)
1	Vancomycin Non-IV	14.8 (14.68,14.86)	Vancomycin Drug Push	-4.75 (-4.84,-4.66)
2	Multi Lumen Dressing Change	5.01 (4.92,5.10)	Pain Cause Incision	-4.19 (-4.28,-4.10)
3	Vancomycin dose (2.0, 1250.0]	5.01 (4.91,5.11)	Gastric Meds	-3.89 (-3.99,-3.79)
4	Osmolality, Urine (312.0, 360.0]	4.80 (4.70,4.90)	Dextrose 5%	-3.78 (-3.86,-3.69)
5	Safety Measures Lines and tubes concealed	4.39 (4.30,4.49)	Daily Wake up Deferred	-3.53 (-3.62,-3.43)
6	STOOL	4.39 (4.32,4.47)	Blood value LG	-3.50 (-3.48,-3.28)
7	Stool Estimate Small	3.83 (4.03,4.17)	Pain Location Neck	-3.38 (-3.45,-3.25)
8	Fibrinogen, Functional (44.999,165.0]	3.78 (3.73,3.93)	Mean Airway Pressure min(8.0,9.0]	-3.35 (-3.36,-3.18)
9	Abdominal Assessment Distented	3.68 (3.71,3.86)	CV - past medical history PVD	-3.27 (-3.35,-3.16)
10	Incision Closure Staples	3.63 (3.59,3.78)	Suptum Color Blood Tinged	-3.26 (-3.29,-3.15)
	Double Exponential Positive Betas	Value (95% C.I.)	**Double Exponential Negative Betas**	Value (95% C.I.)
1	Vancomycin Non-IV	155.6 (154.7,156.5)	Gastric Meds	-39.6 (-40.9,-38.3)
2	Multi Lumen Dressing Change (0.728, 3652.85]	47.7 (47.1,48.4)	Pain Cause Incision	-33.0 (-33.9,-32.2)
3	Safety Measures lines and tubes concealed	37.9 (36.9,38.9)	Cortisol	-28.7 (-30.0,-27.3)
4	Osmolality, Urine (312.0, 360.0]	37.8 (37.0,38.6)	Mean Airway Pressure min(8.0,9.0]	-28.5 (-29.5,-27.5)
5	Fibrinogen, Functional (44.999,165.0]	35.8 (35.0,36.7)	Pain Location Neck	-27.8 (-28.7,-26.8)
6	Oral Cavity not Assessed	35.7 (34.5,36.9)	Daily Wake up Deferred	-27.5 (-28.3,-26.8)
7	Dressing Status Reinforced	33.0 (31.9,34.1)	Impaired Skin Dressing Change	-26.4 (-27.3,-25.4)
8	Incision Closure Staples	30.9 (30.0,31.8)	Emesis Appearance	-26.0 (-27.1,-24.9)
9	Acyclovir	29.7 (28.2,31.3)	NaCl 0.9%	-25.2 (-25.8,-24.6)
10	Impaired Skin - Dressing Removed	27.9 (26.3,29.5)	Pulmonary Artery Pressure Alarm - High	-25.0 (-26.3,-23.6)
	Regularized Horseshoe Positive Betas	Value (95% C.I.)	**Regularized Horseshoe Negative Betas**	Value (95% C.I.)
1	Vancomycin Non-IV	2.54 (2.50,2.58)	Braden Friction/Shear No Problem	-0.76 (-0.77,-0.75)
5	STOOL	0.83 (0.28,0.85)	Skin Integrity: Intact	-0.25 (-0.26,-0.24)
3	Metronidazole	0.63 (0.60,0.66)	16 Gauge	-0.19 (-0.21,-0.18)
4	Metronidazole Drug Push	0.54 (0.52,0.56)	Chest Tube Site Mediastinal	-0.14 (-0.15,-0.12)
5	Stool Estimate Small	0.37 (0.35,0.39)	Ciprofloaxin	-0.10 (-0.12,-0.09)
6	CV - past medical history CHF	0.36 (0.34,0.37)	Temporary Pacemaker Wires Venticular	-0.10 (-0.11,-0.09)
7	RDW (16.8,29.4]	0.29 (0.28,0.30)	RL Strength/Movement Normal	-0.07 (-0.07,-0.06)
8	Vancomycin Dose Value (2.0, 1250.0]	0.18 (0.16,0.20)	16 Gauge Site Appear	-0.06 (-0.07,-0.05)
9	White Blood Cells (14.8, 462.6]	0.14 (0.13,0.15)	Heart rate Alarm - High	-0.05 (-0.06,-0.05)
10	Creatine Kinase (CK)	0.11 (0.10,0.12)	OR Crystalloid Intake	-0.05 (-0.06,-0.04)

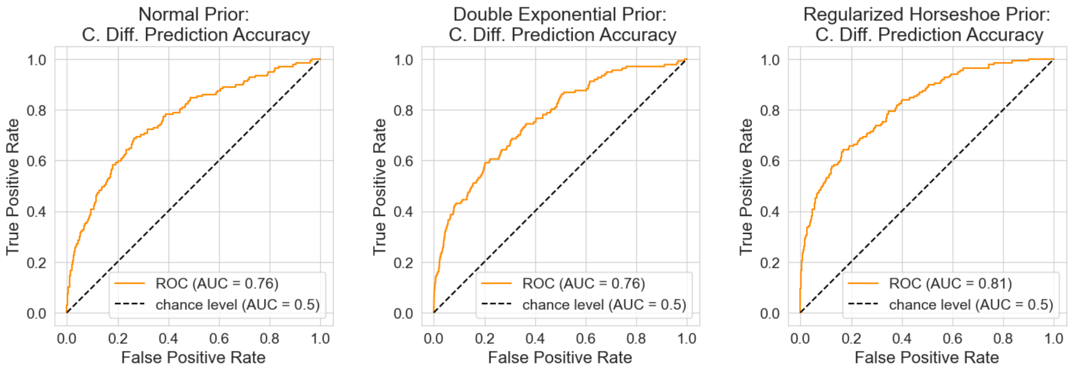

Figure 4. *ROC curves for all three priors (all predictors)*

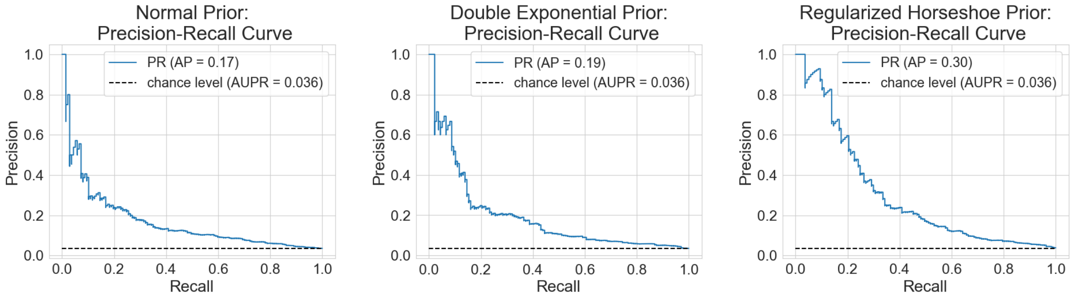

Figure 5. *Precision-recall curves for all three priors (all predictors)*

Table 3: AUROC and AUPR for each prior (all predictors)

Model	AUROC (95% C.I.)	AUPR (95% C.I.)
Normal Prior	0.76 (0.71, 0.81)	0.17 (0.14, 0.20)
Double Exponential Prior	0.76 (0.72, 0.81)	0.19 (0.16, 0.22)
Horseshoe Prior	0.81 (0.77, 0.86)	0.30 (0.26, 0.34)

Table 4: Positive and Negative Betas for Each Prior (subset of predictors)

	Normal Prior Positive Betas	Value (95% C.I.)	**Normal Prior Negative Betas**	Value (95% C.I.)
1	RSBI Deferred	4.73 (4.61,4.84)	Pain Cause Incision	-21.6 (-4.02,-3.84)
2	Less Restrictive Measures	4.13 (4.03,4.23)	Endoscopy	-3.93 (-3.95,-3.71)
3	Ventilator Tank #2	4.12 (4.03,4.21)	LLE Color Pale	-3.83 (-3.94,-3.72)
4	Pain Type Cramping	4.02 (3.92,4.11)	Pain Location Right Lower Quadrant	-3.79 (-3.92,-3.67)
5	Pain Management Local/regional Anaesthetic	3.98 (3.86,4.10)	Pressure Reducing Device Multipodis Boots	-3.72 (-3.83,-3.60)
6	Furosemide (Lasix)	3.92 (3.81,4.04)	Minute Volume Alarm - Low	-3.49 (-3.59,-3.39)
7	Osmolality, Urine	3.90 (3.80,4.00)	GU Catheter Insertion Date	-3.44 (-3.55,-3.32)
8	Pain Location Lower Left Quadrant	3.77 (3.65,3.90)	Multi Lumen Insertion Date	-3.33 (-3.45,-3.21)
9	Stool Guaiac QC	3.65 (3.56,3.73)	Cough Type Congested	-3.29 (-3.39,-3.20)
10	Riker-SAS Scale Sedated	3.61 (3.49,3.72)	NaCl 0.9%	-3.29 (-3.36,-3.21)
	Double Exponential Positive Betas	Value (95% C.I.)	**Double Exponential Negative Betas**	Value (95% C.I.)
1	Pain Location Lower Left Quadrant	42.2 (41.2,43.3)	Pain Location Lower Right Quadrant	-50.7 (-52.4,-49.0)
2	RSBI Deferred	40.6 (39.5,41.7)	Pressure Reducing Device Multipodis Boots	-37.3 (-38.3,-36.3)
3	Pain Type Cramping	33.0 (32.3,33.6)	Multi Lumen Insertion Date	-34.7 (-36.1,-33.3)
4	Pain Management Local/regional Anaesthetic	32.8 (31.8,33.7)	GU Catheter Insertion Date	-32.9 (-34.1,-31.7)
5	Furosemide (Lasix)	31.0 (30.1,31.8)	Edema Amount 8mm	-29.5 (-30.9,-28.2)
6	Riker-SAS Scale Sedated	29.5 (28.7,30.4)	Dextrose 10%	-28.8 (-30.5,-27.1)
7	BiPap O2 Flow (3.0,4.0]	29.1 (27.6,30.7)	LLE Color Pale	-28.6 (-29.7,-27.6)
8	Ceftazidime Drug Push	29.1 (27.4,30.7)	Endoscopy	-28.4 (-29.3,-27.5)
9	Less Restrictive Measures	28.5 (27.8,29.2)	Urobilinogen	-27.5 (-28.8,-26.2)
10	Cardiac Index (CI NICOM)	28.4 (26.6,30.1)	Pain Cause Incision	-26.9 (-27.6,-26.3)
	Regularized Horseshoe Positive Betas	Value (95% C.I.)	**Regularized Horseshoe Negative Betas**	Value (95% C.I.)
1	RDW (16.8, 29.4]	0.61 (0.61,0.62)	Chest Tube Site Mediastinal	-0.51 (-0.55,-0.47)
2	White Blood Cells (14.8, 462.6]	0.57 (0.56,0.57)	Temporary Ventricular Capture	-0.45 (-0.48,-0.42)
3	Atypical Lymphocytes	0.55 (0.55,0.56)	Self ADL	-0.19 (-0.20,-0.18)
4	Cefepime	0.50 (0.49,0.51)	16 Gauge placed in outside facility value 0	-0.18 (-0.20,-0.17)
5	Stool Management	0.47 (0.45,0.48)	LL Strength/Movement Normal	-0.16 (-0.17,-0.15)
6	Calcium, Total max(4.6,7.8]	0.33 (0.32,0.34)	18 Gauge placed in the field mean (-.001,1.0]	-0.14 (-0.15,-0.13)
7	Stool Culture	0.28 (0.27,0.29)	Incision Dressing Ace Wrap	-0.10 (-0.12,-0.08)
8	Self ADL	0.26 (0.24,0.28)	16 Gauge Dressing Occlusive value 1.0	-0.06 (-0.07,-0.05)
9	Skin Integrity Impaired	0.19 (0.18,0.20)	Braden Friction/Shear No Problem	-0.05 (-0.06,-0.04)
10	Heart Rhythm AF	0.13 (0.12,0.14)	18 Gauge placed in the field max 0.0	-0.04 (-0.05,-0.04)

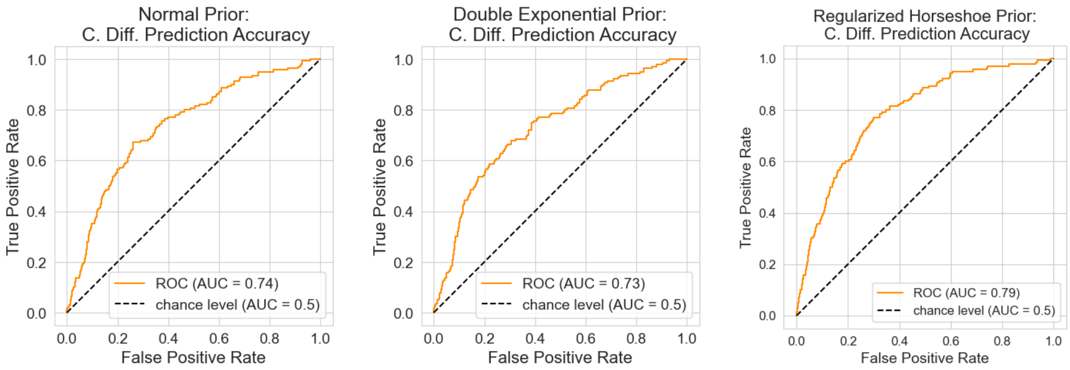

Figure 6. *ROC curves for all three priors (subset of predictors)*

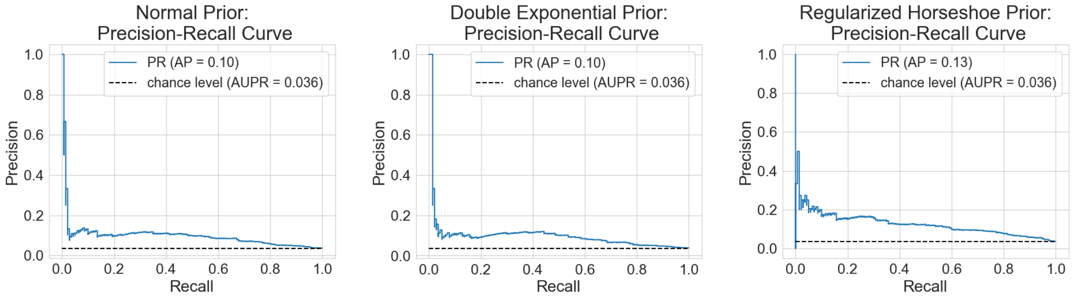

Figure 7. *Precision-recall curves for all three priors (subset of predictors)*

Table 5: AUROC and AUPR for each prior (subset of predictors)

Model	AUROC (95% CI)	AUPR (95% C.I.)
Normal Prior	0.72 (0.68, 0.77)	0.10 (0.08, 0.12)
Double Exponential Prior	0.73 (0.67, 0.77)	0.10 (0.08, 0.12)
Horseshoe Prior	0.79 (0.74, 0.83)	0.13 (0.11, 0.15)

6. Discussion

6.1. Impact of Data Used

Comparing the results between the models which use all of the available predictors versus the models using the subset of predictors, we can see drastic differences in predictive capability favoring the model using all predictors. With the normal prior, there is a 4% reduction in the AUROC curve and a 7% reduction in the AUPR curve. For the double exponential prior there is also a 4% reduction in the AUROC curve in addition to a 9% reduction in the AUPR curve. With the regularized horseshoe prior, there is a 2% reduction in the AUROC curve and a 17% reduction in the AUPR curve. Here we can clearly see the over-optimistic characteristic of ROC curves since the difference in the AUROC curves is extremely small compared to the difference seen between the AUPR curves.

6.2. Impact of Regularization

Disregarding the data given to the model, we can clearly see the impact our choice of prior had on predictive capability. Originally we were curious to see if ℓ^1 or ℓ^2 regularization had an impact on prediction capability, and there is some evidence pointing towards using ℓ^1 regularization. However, the aspect of the priors that appears to have the largest impact on predictive capability is the sparsity of predictions. The regularized horseshoe prior is a sparse prior, meaning most of the coefficients will be zero and there will be a few non-zero coefficients. The normal prior and double exponential priors are not sparse priors and thus they will have many non-zero coefficients. The differences in coefficient size can be seen in Figures 8 and 9.

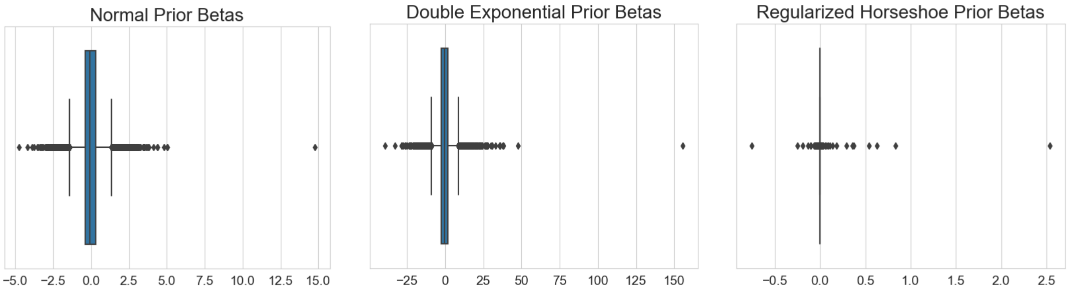

Figure 8. *Plots for the distribution of betas for all three priors with all predictors*

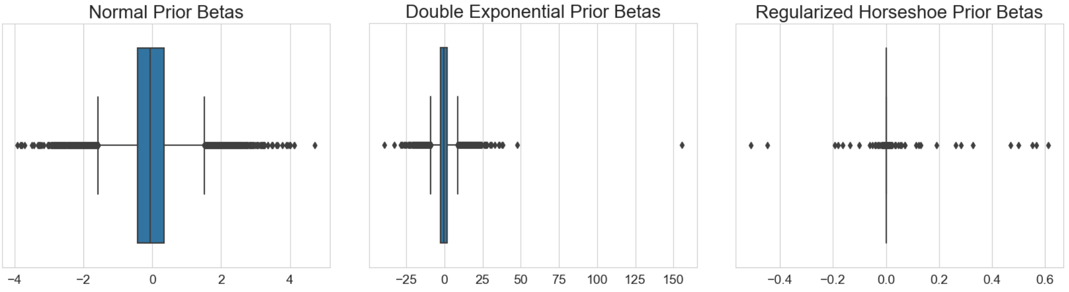

Figure 9. *Plots for the distribution of betas for all three priors with subset of predictors*

6.3. Impact of Sparsity

For the model with all predictors included, we see a 5% improvement in the AUROC curve when using the regularized horseshoe prior compared to both the normal and double exponential priors. In addition, with the regularized horseshoe prior we see a 13% improvement in the AUPR curve over the normal prior and a 11% improvement over the double exponential prior. For the model with the subset of predictors, we see a 7% improvement in the AUROC curve when using the regularized horseshoe prior compared to both the normal and double exponential priors. With the regularized horseshoe prior we also see a 3% improvement in the AUPR curve over both the normal and double exponential priors.

The reasoning behind why sparse priors such as the regularized horseshoe prior perform better in this study is due to the nature of our data. Our dataset with all predictors contains 19,596 patients and 11,489 predictors and our dataset without all predictors contains 19,596 patients and 11,416 predictors. Therefore, our data is very wide (large amount of predictors for the number of patients) which can lead overfitting (Zhang, 2014). Since we have a large number of predictors and both the double exponential and normal priors produce numerous comparatively large coefficients, these two models can be negatively impacted by the overfitting of the training dataset. The regularized horseshoe prior, on the other hand, combats the issue of overfitting by only allowing a small number of coefficients to be comparatively large. We can see from the box and whisker plots in Figures 8 and 9 that nearly all of the mass is on zero for this prior. To assess overfitting in our models, we can run a posterior predictive check to see how well our models fit our training data (Gelman et al., 2020). To perform the posterior predictive check, we used all η_i from the training dataset and took the expit(η_i), giving us a probability, described in Section 6.5, that a patient tested positive for CDI. With this probability, p, we randomly selected a 1 with probability p or 0 with probability (1-p). We then compared our new vector of 1s and 0s to the training vector of C. diff. cases using confusion matrices shown in Tables 6 and 7. In Tables 6 and 7, we can see that the models using the normal prior classify all patients nearly perfectly and the models using the double exponential prior classify all patients perfectly. Due to the number of predictors we have and the fact that these models performed quite poorly when evaluated on the testing set we have reason to believe that both of these models suffer from overfitting. On the contrary, the models using the regularized horseshoe prior are much less effective in classifying the patients in the training dataset. Although this does provide evidence that performance on the testing dataset will not be spectacular, it does show how the regularized horseshoe prior helps prevent overfitting when there are a relatively large number of predictors and a relatively small number of patients.

Table 6: Posterior Predictive Checks (subset of predictors)

Normal				Double Exponential				Regularized Horseshoe			
		Predicted				Predicted				Predicted	
		Pos	Neg			Pos	Neg			Pos	Neg
Actual	Pos	548	8	Actual	Pos	556	0	Actual	Pos	112	444
	Neg	0	15,120		Neg	0	15,120		Neg	383	14,737

Table 7: Posterior Predictive Checks (all predictors)

Normal				Double Exponential				Regularized Horseshoe			
		Predicted				Predicted				Predicted	
		Pos	Neg			Pos	Neg			Pos	Neg
Actual	Pos	539	14	Actual	Pos	553	0	Actual	Pos	40	513
	Neg	0	15,123		Neg	0	15,123		Neg	467	14,656

6.4. Interpolation of Coefficients

In addition to having a smaller number of large coefficients, the regularized horseshoe prior also produces coefficient values closer to zero which is beneficial in improving prediction accuracy. If you recall from earlier, η was the resulting sum from the product of $X\beta$. Therefore, larger coefficients (in absolute value) seen in the vector β will result in more extreme values for η. Our model then takes the expit(η), or the inverse-logit(η), in order to obtain a probability that the patient will contract CDI. The problem with extreme coefficients and thus extreme values of η can be seen in Figure 10. With large coefficients, it can be extremely easy to obtain an large value for η, and this large value may only be influenced by one variable. For example, in the model with all predictors which uses a double exponential prior, the β_i for vancomycin non-IV is 155.6. Given that β_0 for this model is -0.102, a value of 155.6 increases the probability of a positive C. diff. case by $\text{expit}(\beta_0 + 155.6) - \text{expit}(\beta_0) = \overline{.99} - .47 = .53$ meaning the variable vancomycin non-IV increases the probability of the model predicting C. diff. by 53% over the baseline (expit(β_0)). Furthermore, to decrease this value to 52% over the baseline, we would need to decrease η to 4.60 ($\text{expit}(4.60) = .99$). Since the largest negative β_i for the same model is -39.6 and $\text{expit}(\beta_0 + 155.6 - 39.6) = \overline{.99}$ as well, it seems reasonable to assume that if a patient takes vancomycin not through an IV then this model will classify them as contracting CDI. With the regularized horseshoe prior, $\beta_0 = 3.49$ and the largest $\beta_i = 2.54$ for the model using all predictors and for the model using the modified set of predictors, $\beta_0 = -3.93$ and the largest $\beta_i = 0.61$. With these coefficient sizes, our model is mush less likely to make predictions based on only one variable since small changes in η will have a noticeable impact on expit(η).

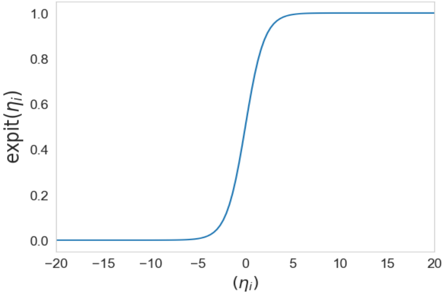

Figure 10. *Expit graph showing how extreme etas result in a probability of 1 or 0*

6.5. All Predictors versus Censored Model

Of the two sets of data we used, fitting a model to the modified set of predictors provides us with the most valuable information about which variables are most important. However, the benefit of also fitting a model which uses all EHR data is that our results from this model provide an upper bound for predictive capability. Being aware of the upper bound on predictive capability shows how much of an impact a set of variables has on each model. It also helps visualize how difficult the problem is that we are trying to predict.

The reason behind why the model with all predictors was substantially more accurate can be attributed to reverse causation. Given a dataset with a large number of predictors, we are unaware of what all of the predictors are in our data. Therefore, when we use all of the predictors we will be including variables which are present because doctors have either begun treating a patient for C. diff. or have a strong suspicion that the patient has C. diff. and are taking actions accordingly. For example, all three priors identified vancomycin non-IV as being the most important predictor for identifying a positive C. diff. case. This, however, is due to the fact that vancomycin, which is an antibiotic, taken orally is one of the most common treatments for C. diff. (Shen et al., 2008). The fact that vancomycin is generally taken orally could explain why vancomycin administered through an IV (vancomycin drug push) was the most important variable in determining negative C. diff. cases for the normal prior. In addition to vancomycin, the regularized horseshoe prior found metronidazole, which is also one of the main antibiotics used to treat CDI, to be positively associated with CDI (Shen et al, 2008). The other variable present in the top ten positive coefficients for each model which also implies reverse causation is STOOL. A stool sample involves the collection of fecal matter which can then be sent to the lab and analyzed for different medical conditions (Stool Tests, n.d.). Stool tests can be used for the detection of many different conditions, but it is also one of the main methods for diagnosing patients with C. diff. (MedlinePlus Medical Encyclopedia, n.d.).

Apart from these variables, when using the dataset with all predictors the normal and double exponential priors picked up on similar variables, but very few seem to be related to CDI. The normal prior found that the variable abdominal assessment distented, which signifies a swollen abdomen, to be positively associated with CDI and research has shown that abdominal distention can be caused by CDI (Cowan & Kutty, 2018). Apart from this variable, the remaining positive coefficients do not have apparent commonalities with CDI. This trend can also be observed with the negative coefficients for these two priors. The negative coefficients should theoretically have nothing in common with C. diff., but they are rather uninformative. For instance, if a patient comes in with neck pain or they have pain due to an incision, which could imply a surgical procedure, it is clear the patient's issue is not that they have C. diff., but these variables do not inform us about the patient's susceptibility to CDI.

Again, disregarding variables related to stool, vancomycin, and metronidazole, the regularized horseshoe prior appears better able to find positive predictors that are associated with CDI which do not result from reverse causation. RDW, or red blood cell distribution width, measures the size distribution amongst red blood cells (Fava et al., 2019). RDW generally ranges from 12%-15%, but our model picked up on a RDW of 16.9%-29.4% which is abnormally high (Fava et al., 2019). High RDW can be a sign of anemias and a study conducted by Othman et al. found that patients with pernicious anaemia have an increased risk of CDI (Fava et al., 2019)(Othman et al., 2017). The regularized horseshoe prior also found that patients with low white blood cell (WBC) counts (148 to 4,626 WBCs per microliterare) were more likely to contract CDI. For reference, a normal WBC count is 4,500 to 11,000 WBCs per microliter (WBC Count, n.d.). Low WBC counts, called leukopenia, reduce the body's ability to fight infections, thus leaving patients more predispositioned to C. diff. (Mayo Clinic Staff, n.d.). Furthermore, the regularized horseshoe prior found Creatine Kinase (CK) to be important in identifying positive C. diff cases. CK is a muscle enzyme which can be produced after Rhabdomyolysis (RBD), and RBD has been reported to have been caused by C. diff. (Dungan et al., 2022). Therefore, the presence of CK does not leave patients susceptible to C. diff., but its presence could potentially help doctors detect CDI.

Examining the models which do not use the entire set of available predictors, we notice reverse causation is still an issue, but it has much less of an impact on our results. We ran our model twice before, and after both runs we removed all of the important variables which clearly produced reverse causation. However, looking at the positive betas for the regularized horseshoe and normal prior there are three variables associated with stool which most likely result in some reverse causation.

These variables, however, are not as influential in this model. Aside from the variables associated with stool, the regularized horseshoe model using the subset of predictors also found RDW and low WBC count important in determining positive C. diff. cases. In addition, the regularized horseshoe model found positive associations between C. diff. and atypical lymphocytes, cefepime, and calcium levels. Atypical lyphocytes can be found in a patient's blood, and there presence signifies an immune response to numerous medical conditions including some bacterial infections (Shiftan & Mendelsohn, 1978). Since CDI is a bacterial infection, the presence of atypical lyphocytes could be a sign of CDI. Cefepime is an injection which has been used to treat bacterial infections (Cefepime (Injection Route) Side Effects—Mayo Clinic, n.d.). However, a study done by Muldoon et al. found that increased use of cefepime led to significantly increased rates of C. diff. (Muldoon et al., 2013). Lastly, our regularized horseshoe model found that a calcium level maximum of 4.6-7.8 mg/dL was important in determining positive C. diff. cases. The normal range for ionized calcium is 4.4-5.2 mg/dL, therefore a calcium level maximum of 4.6-7.8 mg/dL is implying a relatively high level of ionized calcium (Normal Calcium Levels, n.d.). The significance of this is that Kochan et al. found that increased levels of ionized calcium, which implies increased concentrations of intestinal calcium, can promote the growth of C. diff. spores (Kochan et al., 2017).

Unlike the positive coefficients identified by the regularized horseshoe prior, the normal and double exponential priors most influential positive coefficients, seven of which are shared between the two models, do not provide much valuable information about patient characteristics which would leave them vulnerable to CDI. This could be the product of the overfitting we saw with these two priors. Both priors found pain in the lower left quadrant, which corresponds to pain in the lower left corner of the abdomen, and cramping to be important predictors of C. diff. These variables could be a signal to doctors that a patient has C. diff. since symptoms of C. diff are abdominal pain and cramping, but they would not provide doctors with information about which patients are more likely to test positive for C. diff. (C. difficile infection—Symptoms and causes, n.d.).

The negative coefficients for all three priors also do not provide us with valuable information about patient characteristics that provide "defense" against CDI. Similar to what we observed from the results gathered using all predictors, the negative coefficients tell us the patient's issue is not C. diff., but they do not provide us with any information regarding a patients susceptibility to CDI.

7. Model Limitations

The main criticisms we have about our model can be attributed to the data we are using, the first of which is reverse causation. Given we have a very large number of predictors, we do not know the nature of all of the variables in our model. Therefore the best way for us to reduce the effect of reverse causation is by repeating the process of running our model, looking at the most important predictors our model found, and then removing any variables that cause reverse causation. This process can reduce the impact of reverse causation, but it cannot remove the entire effect.

Another issue with having a large number of predictors is overfitting. Looking at Tables 6 and 7, we can see the effect of overfitting when using the normal and double exponential priors. Since the models are so well fit to our training datasets, their ability to make accurate predictions on our testing dataset is hindered.

The next issue we have with the data is our inability to identify the timing of ICD9 codes. Due to our inability to discern the timing of ICD9 codes, we cannot tell when in their stay patients tested positive for C. diff., which prevents us from removing any patients who were admitted to the hospital with CDI. This is an issue because our goal is to predict which patients will contract C. diff. during their stay, so individuals who are admitted with C. diff. may negatively impact our results.

An additional area where we would like to improve our data is we would like to be able to use data that has hourly entries per patient from 0-24 hours rather than one entry per patient from the time of admission to 24 hours after admission. We predict that having this data will improve predictions due to repeated instances of certain variables that would be present with hourly entries.

However, we lacked the computational ability and the time to run a model with this much data.

8. Conclusion

We showed that Bayesian logistic regression is able to make predictions in regard to whether or not a patient will contract C. diff. during their stay. Using a regularized horseshoe prior, we were able to obtain an AUROC curve of 0.81 and an AUPR curve of 0.30 when using all predictors and an AUROC curve of 0.79 and an AUPR curve of 0.13 when using a modified set of predictors. Our results showed our model had a similar predictive capability compared to previous studies such as the study by Wiens et al. in "Learning Data-Driven Patient Risk Stratification Models for *Clostridium difficile*", thus we did not see any clear improvements using a Bayesian approach with the data we used (Wiens et al., 2014). It is important to note that although we do not see an improvement in predictive capability, we cannot compare our results directly to those obtained by Wiens et al. since we used different datasets. Furthermore, our main tool for evaluating our results was using precision-recall curves, which have not been used in prior studies. Precision-recall curves are much better suited for evaluating the predictive capability of our models since precision-recall curves are much better suited for unbalanced data. Unbalanced data is similar to the idea of a sparse matrix in that we had an extremely small percentage of C. diff. cases. ROC curves are generally over-optimistic when used to evaluate predictions on unbalanced data since an additional true positive case will increase the true positive rate by a significantly larger margin than one additional false positive case will increase the false positive rate. Precision-recall curves avoid this issue by decreasing the classification threshold and evaluating the precision of our model's predictions at each classification level until all positive cases have been recalled. Since previous studies have not used precision-recall curves in their analysis we cannot compare the performance of our model to other studies using this metric but were able to show the benefit of using a sparse prior with a Bayesian model to predict C. diff. infection.

Our goal in developing a model which predicts C. diff. infection is to create a model that can be used in conjunction with healthcare settings. Implementing such a model which can detect patients who are susceptible to CDI could significantly decrease the rates of CDI in a healthcare setting. For instance, if there was an interface where doctors could enter all variables associated with a patient which was connected to the model, as doctors enter patient information the model could instantaneously identify patients most at risk for C. diff. and alert medical professionals prior to infection. As a result, doctors would be able to take the necessary precautions to limit said patient's exposure to C. diff. infection. There are numerous challenges associated with this task, such as a lack of collaboration between medical centers and statisticians and how we should label and structure medical data, which prevent models from being easily implemented in healthcare settings. In addition to improving the data collection process, we also need to improve our model's predictive capability. More specifically, we need to improve upon our Bayesian approach to logistic regression with a regularized horseshoe prior. We have shown the regularized horseshoe prior outperforms the normal and double exponential priors, so in the creation of a new model, we should strive to improve upon our model using a regularized horseshoe prior. We also do not know how this model performs using different datasets, so it would be wise to compare all new models with our current regularized horseshoe model using the same data. Once we create a model which demonstrates sufficient predictive capability to be effectively used in a healthcare setting, we can begin the process of integrating our model into a healthcare setting. It is also important to note that we can use the same model to predict numerous conditions in addition to C. diff. infection. However, we must be aware of the issue of reverse causation and collaborate with medical professionals to create unique training datasets for each condition we are aiming to predict. During the formation of predictor datasets, it may also be useful to include admission diagnosis (why the patient is at the hospital). Our model only used the ICD9 codes to identify which patients tested positive for C. diff., but using the entirety of the ICD9 code dataset may provide valuable information regarding which conditions can leave patients susceptible to CDI. Ultimately, our model using a Bayesian approach to logistic regression with a regularized horseshoe prior was able to make predictions about C. diff. infection, but there are numerous approaches that can produce improved predictive capability which will hopefully lead to a model robust enough to be implemented in a healthcare setting.

9. References

Blackwell, T. (2015, February 23). Toronto's largest hospital shuts down part of cardiovascular unit after C. difficile outbreak. *National Post.* https://nationalpost.com/health/torontos-largest-hospital-shuts-down-part-of-cardiovascular-unit-after-c-difficile-outbreak

Bois, Justin. (2019). Half-Cauchy distribution—Probability Distribution Explorer documentation. *GitHub.* https://distribution-explorer.github.io/continuous/halfcauchy.html

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling Sparsity via the Horseshoe. *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics in Proceedings of Machine Learning Research*, *5*, 73–80.
https://proceedings.mlr.press/v5/carvalho09a.html

C. difficile infection—Symptoms and causes. (n.d.). Mayo Clinic. Retrieved March 29, 2023, from https://www.mayoclinic.org/diseases-conditions/c-difficile/symptoms-causes/syc-20351691

Cefepime (Injection Route) Side Effects—Mayo Clinic. (n.d.). Retrieved March 29, 2023, from https://www.mayoclinic.org/drugs-supplements/cefepime-injection-route/side-effects/drg-20073408?p=1

Cowan, A. N., & Kutty, G. (2018). Clostridium difficile Colitis in a Patient With Abdominal Distention, Pain, and Severe Constipation. *Federal practitioner : for the health care professionals of the VA, DoD, and PHS, 35*(6), 44–46.

Despite Progress, Ongoing Efforts Needed to Combat Infections Impacting Hospital Patients. (2014, March 26). CDC Newsroom. https://www.cdc.gov/media/releases/2014/p0326-hospital-patients.html

Dubberke, E. R., Reske, K. A., McDonald, L. C., & Fraser, V. J. (2006). ICD-9 codes and surveillance for Clostridium difficile-associated disease. *Emerging infectious diseases*, *12*(10), 1576–1579. https://doi.org/10.3201/eid1210.060016

Dungan, W., Young, G., Collins, B., Romano, J., Honko, N., Rockey, D. (2022). Clostridioides difficile Induced Rhabdomyolysis Associated With Decompensated Cirrhosis. *Journal of investigative medicine high impact case reports, 10*, 23247096221132249.
https://doi.org/10.1177/23247096221132249

Fava, C., Cattazzo, F., Hu, Z. D., Lippi, G., & Montagnana, M. (2019). The role of red blood cell distribution width (RDW) in cardiovascular risk assessment: useful or hype?. *Annals of translational medicine, 7*(20), 581. https://doi.org/10.21037/atm.2019.09.58

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., Modrák, M. (2020). Bayesian Workflow. *arXiv.*
https://doi.org/10.48550/arXiv.2011.01808

ICD - ICD-9-CM - International Classification of Diseases, Ninth Revision, Clinical Modification. (2021, November 3). CDC. https://www.cdc.gov/nchs/icd/icd9cm.htm

Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). *PhysioNet.* https://doi.org/10.13026/C2XW26.

Kochan, T. J., Somers, M. J., Kaiser, A. M., Shoshiev, M. S., Hagan, A. K., Hastie, J. L., Giordano, N. P., Smith, A. D., Schubert, A. M., Carlson, P. E., Jr, & Hanna, P. C. (2017). Intestinal calcium and bile salts facilitate germination of Clostridium difficile spores. *PLoS pathogens, 13*(7), e1006443. https://doi.org/10.1371/journal.ppat.1006443

Larionov, M. (2019, May 16). Horseshoe priors. *Towards Data Science.*
https://towardsdatascience.com/horseshoe-priors-f97672b4f7cb

Li, B. Y., Oh, J., Young, V. B., Rao, K., & Wiens, J. (2019). Using Machine Learning and the Electronic Health Record to Predict Complicated Clostridium difficile Infection. *Open Forum Infectious Diseases*, *6*(5), ofz186. https://doi.org/10.1093/ofid/ofz186

Mayo Clinic Staff. (n.d.). *Low blood cell counts: Side effects of cancer treatment.* Mayo Clinic. Retrieved March 28, 2023, from
https://www.mayoclinic.org/diseases-conditions/cancer/in-depth/cancer-treatment/art-20046192

McElreath R. (2016). *Statistical rethinking : a bayesian course with examples in r and stan.* CRC Press/Taylor & Francis Group.

Muldoon, E. G., Epstein, L., Logvinenko, T., Murray, S., Doron, S. I., & Snydman, D. R. (2013). The impact of cefepime as first line therapy for neutropenic fever on Clostridium difficile rates among hematology and oncology patients. *Anaerobe, 24*, 79–81. https://doi.org/10.1016/j.anaerobe.2013.10.001

Ng, Y.-L., Lo, M. C. K., Lee, K.-H., Xie, X., Kwong, T. N. Y., Ip, M., Zhang, L., Yu, J., Sung, J. J. Y., Wu, W. K. K., Wong, S. H., & Kwok, K.-W. (2021). Development of an Open-Access and Explainable Machine Learning Prediction System to Assess the Mortality and Recurrence Risk Factors of Clostridioides Difficile Infection Patients. *Advanced Intelligent Systems*, *3*(1), 2000188. https://doi.org/10.1002/aisy.202000188

Normal Calcium Levels. (n.d.). UCLA Health. Retrieved March 29, 2023, from https://www.uclahealth.org/medical-services/surgery/endocrine-surgery/patient-resources/patient-education/normal-calcium-levels

Othman, F., Crooks, C. J., & Card, T. R. (2017). The risk of Clostridium difficile infection in patients with pernicious anaemia: a retrospective cohort study using primary care database. *United European gastroenterology journal, 5*(7), 959–966. https://doi.org/10.1177/2050640617695697

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051. https://doi.org/10.1214/17-EJS1337SI

Pykes, K. (2022, July 22). *Fighting Overfitting With L1 or L2 Regularization: Which One Is Better?* Neptune.Ai. https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization

Rathod, S., McManus, D., Rivera-Vinas, J., Topal, J. E., & Martinello, R. A. (2019). 2385. Evaluating the Antibiotic Risk for Clostridioides difficile Infection (CDI): Comparing Piperacillin/-Tazobactam to Cefepime and Ceftazidime. *Open Forum Infectious Diseases, 6*(Suppl 2), S823–S824. https://doi.org/10.1093/ofid/ofz360.2063

Shen, E. P., & Surawicz, C. M. (2008). Current Treatment Options for Severe Clostridium difficile-associated Disease. *Gastroenterology hepatology, 4*(2), 134–139.

Shiftan, T. A., & Mendelsohn, J. (1978). The circulating "atypical" lymphocyte. *Human pathology, 9*(1), 51–61. https://doi.org/10.1016/s0046-8177(78)80007-0

Stan Development Team. (n.d.). *14.1 Hamiltonian Monte Carlo | Stan Reference Manual.* Retrieved March 31, 2023, from https://mc-stan.org/docs/$2_19/reference-manual/hamiltonian-monte-carlo.html$

Stool C difficile toxin: MedlinePlus Medical Encyclopedia. (n.d.). Retrieved March 26, 2023, from https://medlineplus.gov/ency/article/003590.htm

Stool Tests. (n.d.). Children's Hospital of Orange County. Retrieved March 26, 2023, from https://www.choc.org/programs-services/gastroenterology/digestive-disorder-diagnostics/stool-tests/

Stool tests. (2023, February 1). [Text/html]. Healthdirect; Healthdirect Australia. https://www.healthdirect.gov.au/stool-tests

Tang, S., Davarmanesh, P., Song, Y., Koutra, D., Sjoding, M. W., & Wiens, J. (2020). Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, *27*(12), 1921–1934. https://doi.org/10.1093/jamia/ocaa139

Tran-The, T. D. (2021, November 29). Precision-Recall Curve is more informative than ROC in imbalanced data. *Medium.* https://towardsdatascience.com/precision-recall-curve-is-more-informative-than-roc-in-imbalanced-data-4c95250242f6

WBC Count. (n.d.). Ucsfhealth.Org. Retrieved March 25, 2023, from https://www.ucsfhealth.org/Medical Tests/003643

Wiens, J., Campbell, W. N., Franklin, E. S., Guttag, J. V., & Horvitz, E. (2014). Learning

Data-Driven Patient Risk Stratification Models for Clostridium difficile. *Open Forum Infectious Diseases*, *1*(2), ofu045. https://doi.org/10.1093/ofid/ofu045

Zhang, S., Palazuelos-Munoz, S., Balsells, E. M., Nair, H., Chit, A., & Kyaw, M. H. (2016). Cost of hospital management of Clostridium difficile infection in United States-a meta-analysis and modelling study. *BMC infectious diseases, 16*(1), 447. https://doi.org/10.1186/s12879-016-1786-6

Zhang Z. (2014). Too much covariates in a multivariable model may cause the problem of overfitting. *Journal of thoracic disease, 6*(9), E196–E197. https://doi.org/10.3978/j.issn.2072-1439.2014.08.33

A. STAN: Normal Prior

```
data {
  int<lower=0> n_train;
  int<lower=0> n_test; // number of units
  int<lower=0> d; // number of covariates
  int<lower=0,upper=1> y_train[n_train]; // binary responses
  matrix[n_train, d] x_train; // covariates for each entry, including the
      ↪ intercept covariate
  matrix[n_test, d] x_test; // covariates for each entry, including the intercept
      ↪  covariate
  int do_prior_predictive;

}
parameters {
  vector[d] beta; // the coefficients
  real beta0;
}
transformed parameters {
  vector[n_train] eta_train; // linear predictors
  eta_train = beta0 + x_train * beta;
  vector[n_test] eta_test; // linear predictors
  eta_test = beta0 + x_test * beta;
}
model {
  beta ~ normal(0, 2);
  beta0 ~ normal(0, 5);
  if (do_prior_predictive != 1) {
    y_train ~ bernoulli_logit(eta_train);
  }
}
generated quantities {
  vector[n_test] p_test;
  p_test = inv_logit(eta_test);
}
```

B. STAN: Double Exponential Prior

```
data {
  int<lower=0> n_train;
  int<lower=0> n_test; // number of units
  int<lower=0> d; // number of covariates
  int<lower=0,upper=1> y_train[n_train]; // binary responses
  matrix[n_train, d] x_train; // covariates for each entry, including the
      ↪ intercept covariate
  matrix[n_test, d] x_test; // covariates for each entry, including the intercept
      ↪  covariate
  int do_prior_predictive;

}
parameters {
  vector[d] beta; // the coefficients
  real beta0;

}
transformed parameters {
  vector[n_train] eta_train; // linear predictors
  eta_train = beta0 + x_train * beta;
  vector[n_test] eta_test; // linear predictors
  eta_test = beta0 + x_test * beta;
}
model {

    beta ~ double_exponential(0,8);
    beta0 ~ normal(0, 5);
  if (do_prior_predictive != 1) {
    y_train ~ bernoulli_logit(eta_train);
  }
}
generated quantities {
  vector[n_test] p_test;
  p_test = inv_logit(eta_test);
}
```

C. STAN: Horseshoe Prior

```
data {
   int < lower =0 > n_train ; // number of observations
   int < lower =0 > n_test ; // number of observations
   int < lower =0 > d ; // number of predictors
   int <lower = 0, upper = 1> y_train[n_train]; // outputs
   matrix [n_train ,d] x_train; // inputs
   matrix [n_test ,d] x_test; // inputs
   real < lower =0 > scale_global ; // scale for the half -t prior for tau
   real < lower =1 > nu_global ; // degrees of freedom for the half -t prior
                                 // for tau
   real < lower =1 > nu_local ; // degrees of freedom for the half - t priors
                                 // for lambdas
   real < lower =0 > slab_scale ; // slab scale for the regularized horseshoe
   real < lower =0 > slab_df ; // slab degrees of freedom for the regularized
                                 // horseshoe
   int do_prior_predictive;
}

parameters {
   vector [ d] z;
   real beta0;
   vector < lower =0 >[ d] lambda ; // local shrinkage parameter
   real < lower =0 > caux ;
   real <lower = 0> tau;

}

transformed parameters {
   vector < lower =0 >[ d] lambda_tilde ; // 'truncated' local shrinkage
       ↪ parameter
   real < lower =0 > c; // slab scale
   real <lower = 0> tau0;
   vector [ d] beta ; // regression coefficients
   vector [n_train] f_train; // latent function values
   vector [n_test] f_test; // latent function values
   c = slab_scale * sqrt ( caux );
   lambda_tilde = sqrt ( square(c) * square ( lambda ) ./ (square(c) + square(
       ↪ tau)* square ( lambda )) );
   tau0 = scale_global ;
   beta = z .* lambda_tilde * tau0 ;
   f_train = beta0 + x_train * beta ;
   f_test = beta0 + x_test * beta ;
}

model {
   // half -t priors for lambdas and tau , and inverse - gamma for c ^2
   z ~ normal (0 , 1);
   lambda ~ student_t ( nu_local , 0, 1);
   tau ~ student_t (nu_global , 0, scale_global);
   caux ~ inv_gamma (0.5* slab_df , 0.5* slab_df );
   beta0 ~ normal(0, 5);
   if (do_prior_predictive != 1){
       y_train ~ bernoulli_logit(f_train);
       }
}
```

```

generated quantities {
  vector[n_test] p_test;
  p_test = inv_logit(f_test);
}
(Piironen \& Vehtari, 2017)
```