Application of machine learning to predict metastatic cancer diagnosis periods

June 2024

ABSTRACT

Utilizing a rich dataset provided by the Women in Data Science (WiDS) 2024 University Challenge, we aim to address potential issues in predicting the diagnosis period of metastatic breast cancers in the United States. Our study focuses on utilizing various machine learning models to not only predict the diagnosis period but also to explore the impact of socio-economic factors, individual patient characteristics, and geographical information in improving diagnostic accuracies. Through data analysis and model evaluation, our goal is to demonstrate how machine learning can improve predictive performance and provide valuable insights into factors influencing the length and outcome of cancer treatment. This research underscores the potential of machine learning in transforming healthcare diagnostics and offers a model for future studies aiming to optimize medical predictions and patient outcomes.

KEYWORDS: breast cancer, diagnosis period, machine learning, regression modeling, feature importance, WiDS

Author Information

George Tzimas, DePaul University, gtzimas@depaul.edu

Eric Piatek, DePaul University, epiatek@depaul.edu

Esmeralda Villela, DePaul University, evillela@depaul.edu

Sarah Hashmi, DePaul University, shashm13@depaul.edu

1 INTRODUCTION

Breast cancer remains one of the most common diseases affecting women worldwide, with early detection and treatment being crucial to improving survival rates. The ability to effectively predict the onset and progression of breast cancer can significantly influence treatment decisions and outcomes. Advances in machine learning offer promising tools for enhancing the predictive accuracy of breast cancer diagnostics. Our study, conducted as part of the Women in Data Science (WiDS) 2024 University Challenge, utilizes a comprehensive dataset provided by Health Verity, encompassing health, socio-economic, and environmental data of patients diagnosed with metastatic

triple-negative breast cancer in the United States. This project aims to predict the diagnosis period of metastatic cancer using various machine learning models, exploring the influence of different factors including demographic data, socio-economic status, and geographical variables. By integrating these diverse data streams, we seek to uncover patterns that could lead to earlier and more precise cancer diagnoses, ultimately contributing to better patient outcomes.

When it comes to the research questions, our group intends to tackle the following four questions: (1) How can we predict the response variable, estimated treatment duration, using all of the social, economic, and climate information available to us?, (2) How do socioeconomic factors and patient demographics specifically influence the length of treatment for metastatic breast cancer?, (3) How does the application of different data imputation strategies affect the predictive accuracy of treatment period models in the metastatic breast cancer dataset? More specifically, in terms of pre-processing steps and testing and training of data?, and (4) How can Bayesian modeling techniques be implemented in order to predict metastatic diagnosis period in patients with triple negative breast cancers?

2 LITERATURE REVIEW

Breast cancer (BC) is one of the most prevalent cancer types worldwide, especially in females. Being able to diagnose it during the early stages of development is crucial for successfully treating it and improving the survival rates. Early detection increases the chances of effective intervention and also significantly reduces the mortality associated with the disease.

Various studies have focused on identifying what factors contribute to diagnosis delays with the aim of improving early detection and treatment outcomes, especially in women. In recent decades, machine learning (ML) methods have revolutionized the ability to model complex relationships between various factors that may influence diagnosis delays in BC. For instance, [Dehdar et al., 2023] studied the application of various ML techniques to predict delays in BC diagnosis among 630 Iranian women. They utilized four machine learning models: Extreme Gradient Boosting (XGBoost), Random Forest (RF), Neural Networks (NN) and Logistic Regression (LR) to find which factors were significant in delaying BC diagnosis. The results showed that women with an urban residency who got married or had their first child at an age older than 30 and those without children were at a higher risk of experiencing a diagnosis delay [Dehdar et al., 2023].

Similarly, [Kourou et al., 2015] also offers an overview of various machine learning models that have been implemented for cancer survival predictions. They highlight the importance of having large and diverse datasets in order to build robust ML models that are able to effectively predict cancer susceptibility, recurrence and survival. For instance, [Lou et al., 2020] explored ML models to predict recurrence within 10 years after BC surgery. They were able to utilize a large pool of data from medical centers in Taiwan to test their model. The availability of such quality data as well as the use of effective ML techniques could effectively integrate these models into clinical practice.

A binary breast cancer classifier algorithm based on the blood test and anthropometric data was built by [Binsaif, 2022]. Six machine learning models were compared for the best performance. The models that achieved the highed AUROC in this article where the random forest model, followed by the logistic regression and SVM.

ML models have also been used in other research areas to predict health outcomes. In the study conducted by [Tate et al., 2020], the authors examined five techniques to predict mental health outcomes in adolescence. Similar to [Dehdar et al., 2023], they also explored XGBoost, RF, NN, and LG. Additionally, they looked at Support Vector Machines (SVM). Although ML has demonstrated great benefits in the area of cancer research, [Li et al., 2021] noted in their review

that it's application remains controversial. They included thirty-one studies in their review, which were comprised of 19 decision trees (DT), 18 artificial neural networks (ANN), 16 support vector machines (SVM) and 10 ensemble learning (EL) models. They found that the performance of ML models did not necessarily show any improvement compared with traditional statistical methods. This does not imply that ML models are no beneficial, but rather that additional research in this area is needed.

The benefits of ML models as a potential to predict BC based on hidden features in data was explored by [Rabiei et al., 2022]. They explored different ML approaches by applying demographic, laboratory, and mammographic data. They found that RF resulted in higher performance compared to other techniques. ML techniques could provide a model for early detection of the disease, this could help doctors flag patients at risk, based on data already available.

Given that machine learning-based predictive models promise earlier detection techniques for breast cancer diagnosis, [Rasool et al., 2022] proposed data exploratory techniques (DET) and developed four different predictive models to improve breast cancer diagnostic accuracy. They presented an SVM, LR, KNN, and ensemble classifier. The mining techniques implemented allowed them detect higher accuracy within their predictive models. They noted 99.3 % accuracy with the SVM polynomial kernel and 98.06% with recursive features elimination.

[Yang et al., 2020] explored the possibility of using ML techniques to predict the response of neoadjuvant chemotherapy (NAC) in BC treatment. Specifically, they explored a 17-gene Naive Bayes (NB) prediction model for their research.They found that the NB had a relatively high predictive value compared to SVM and KNN algorithms. However, this model was based on the qPCR gene and future models should explore a wider set of sample to further examine it's clinical application.

[Dubey et al., 2023] aimed to apply various ML methods for planning patient treatments (specifically amongst women with breast cancer), where logistic regression, random forest and KNN were utilized. Using logistic regression and RF to determine patient status (alive/dead) after 5 years based on treatment and KNN to examine alternate treatment plans, the study was able to effectively model survival probabilities to support treatment plan modeling for individuals.

Another such article by [Lee, 2023] titled "Deep Learning Techniques with Genomic Data in Cancer Prognosis" offers an insightful exploration of how deep learning technologies are transforming the field of cancer prognosis through genomic data analysis. The study identifies complex patterns within extensive, high-dimensional genomic datasets, which is pivotal for advancing our understanding and predictions regarding cancer survival. The review thoroughly evaluates various deep learning approaches and their applications in interpreting genomic information, highlighting their potential to enhance diagnostic accuracy and treatment effectiveness. This work is essential for researchers focusing on the integration of bioinformatics and machine learning to improve outcomes in breast cancer and other malignancies.

One other particular article that sticks out is the article by [Liu and Kurc, 2022]. In this article the authors examined the integration of exploratory data analysis with deep learning to enhance the processing of breast cancer datasets, specifically targeting survival analysis. The research highlights the critical roles of thorough data preprocessing and strategic visualization in boosting model efficacy and deepening insights into the survival rates among breast cancer patients.

[Choi et al., 2009] proposed the use of a hybrid Bayesian network model for the prediction of breast cancer prognosis. Of the three models tested in the study, the hybrid model achieved the highest AUC with 0.935, showing the ability for Bayesian modeling to be used in the framework of the question asked in our research. Similarly, [Kharya and Soni, 2016] utilizes a naive Bayes classifier, due to suitability when dimensionality of the data is high. In the study from [Kaur et al., 2022], a Bayesian hyperparameter optimizer BSense was developed and proposed, integrating various

deep learning models as well as the use of Bayesian Gaussian (or stochastic) processes in order to find optimal hyperparameters with less computational power needs.

When discussing Bayesian Models in the context of breast cancer datasets, we can also consider the article by [Su et al., 2023]. In this article, the researchers edeveloped a prognostic model for breast cancer using logistic regression and a Hybrid Bayesian Network. The model combined clinical data and machine learning algorithms to enhance the accuracy of predicting breast cancer outcomes. They focus on applying Bayesian network to integrate diverse data sources and improving the predictive performance of traditional logistic models.

3 DATA

3.1 Overview

The dataset consists of two files, **train.csv** and **test.csv**, with 13,173 and 5,646 records respectively. There are a total of 152 features in the training set and 151 in the testing set (the testing set does not contain the target variable, as with most Kaggle competitions). 141 of our features are numeric and 11 are categorical. Further details as to what each feature represents are available in Table 18. Our target feature is **metastatic_diagnosis_period**, which is the period (in days) in which metastatic cancer was diagnosed. A snapshot of the dataset is available in Table 3. Statistical summaries for some of our numeric and categorical features are available in Tables 4 and 5, respectively.

Table 1: Dimensions of training and testing set.

	Train	Test
# of rows	13,173	5,646
# of cols	152	151

Table 2: Count of features by data type.

	Count
Numeric	141
Categorical	11

Table 3: Random sample of 6 records for the first 10 features of the dataset.

patient_id	patient_race	payer_type	patient_state	patient_zip3	Region	Division	patient_age	patient_gender	bmi
671092	White	COMMERCIAL	TX	773	South	West South Central	56	F	
958657	Black	MEDICAID	GA	303	South	South Atlantic	83	F	
566534		MEDICAID	NY	121	Northeast	Middle Atlantic	50	F	23.13
471521	Hispanic	MEDICAID	CA	922	West	Pacific	91	F	
162323	White	MEDICARE ADVANTAGE	SC	296	South	South Atlantic	83	F	
514868	Black	MEDICAID	GA	314	South	South Atlantic	58	F	20.8

3.2 Imputing missing values

Many of our feature contain missing entries for some of our features at varying proportions. Features **metastatic_first_novel_treatment_type** and **metastatic_first_novel_treatment** are all null

Table 4: Statistical summary of the first 10 numeric features in the dataset.

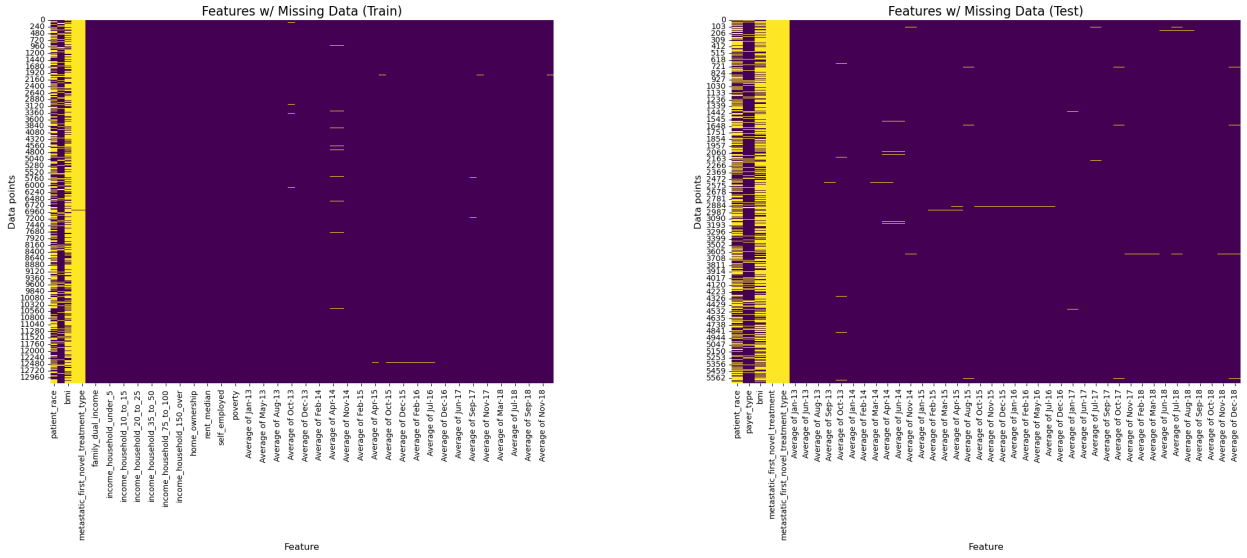
	patient_id	patient_zip3	patient_age	bmi	population	density
count	13173	13173	13173	4102	13173	13173
mean	555441.78	568.53	59.27	29.17	20651.37	1776.87
std	259476.50	275.76	13.22	5.75	13840.38	3876.06
min	100043	100	18	15	635.55	0.92
25%	335100	330	50	24.825	9160.34	163.15
50%	555769	557	59	28.58	18952.78	700.34
75%	780967	832	67	33	30021.28	1666.52
max	999982	995	91	97	71374.13	29851.69

Table 5: Statistical summary of the first 10 categorical features in the dataset.

	patient_race	payer_type	patient_state	Region	Division	patient_gender
count	6516	11408	13173	13173	13173	13173
unique	5	3	44	4	8	1
top	White	COMMERCIAL	CA	South	East North Central	F
freq	3565	6297	2377	3960	3010	13173
25%	335100	330	50	24.825	9160.34	163.15
50%	555769	557	59	28.58	18952.78	700.34
75%	780967	832	67	33	30021.28	1666.52
max	999982	995	91	97	71374.13	29851.69

entries for both sets of data, so they will be dropped. The next features with the highest percentage of missing values for both datasets are **bmi** and **patient_race**, with approximately 50% and 37% of the entries missing for both datasets. One option would be to remove these features, since they contain such a large proportion of missing values, but these features may prove to be very influential when it comes to predicting the metastatic diagnosis rate. Our next best option would be to impute the missing values.

First, we need to check whether the data for these features is missing completely at random or if there exists any pattern to the missingness, such as correlations with the absense of data in other features or the values of the other features themselves. Looking at tables 7 and 8, we can see that missing values for the **bmi** feature is not strongly correlated with missing entries in other features, but the correlation is slightly stronger when it comes to actual values in other features. For **patient_race**, we can see from tables 9 and 10 that a patient's race is also not significantly correlated with missing entries in other features, but the correlation values in regards to actual values in other features is a bit stronger. Given this lack of strong correlations for the features with the highest amount of missing entries, using an iterative imputing algorithm to fill in the missing values would be a suitable approach. This algorithm leverages the inherent relationships between features to impute the missing values using a regression model based on the non-missing values of all other features. This method may be superior to simply filling in missing values for numerical columns with the mean or median and for categorical columns with the mode, since it takes into account the inherent relationship with other features.



((a)) Training ((b)) Testing

Figure 1: Visualization of missing values from both datasets.

Table 6: Top 5 features based on percentage of missing values.

Feature	Train	Test
bmi	50.89%	51.28%
patient_race	37.35%	36.24%
payer_type	9.90%	10.21%
Average of Apr-14	1.01%	1.24%
Average of Jun-14	0.85%	1.03%

Table 7: Top 10 absolute correlation values based on presence of null values across all columns null values present on bmi.

Feature	Correlation
bmi	1.000000
payer_type	0.032058
Average of Jan-15	0.031744
Average of Feb-15	0.023167
Average of Mar-15	0.023167
Average of Jun-14	0.015860
Average of Dec-16	0.015412
Average of Apr-15	0.015238
Average of Oct-13	0.013189
Average of Aug-13	0.012358

Table 8: Top 10 absolute correlation values based on actual values across all columns and null values present on bmi.

Feature	Correlation
bmi	1.000000
Average of May-14	0.066742
Average of May-13	0.066233
Average of Oct-17	0.064767
Average of Apr-14	0.064440
Average of Apr-13	0.063058
Average of Sep-14	0.062245
Average of Oct-15	0.061854
Average of Nov-14	0.061701
Average of Oct-16	0.061678

Table 9: Top 10 absolute correlation values based on presence of null values across all columns null values present on patient race.

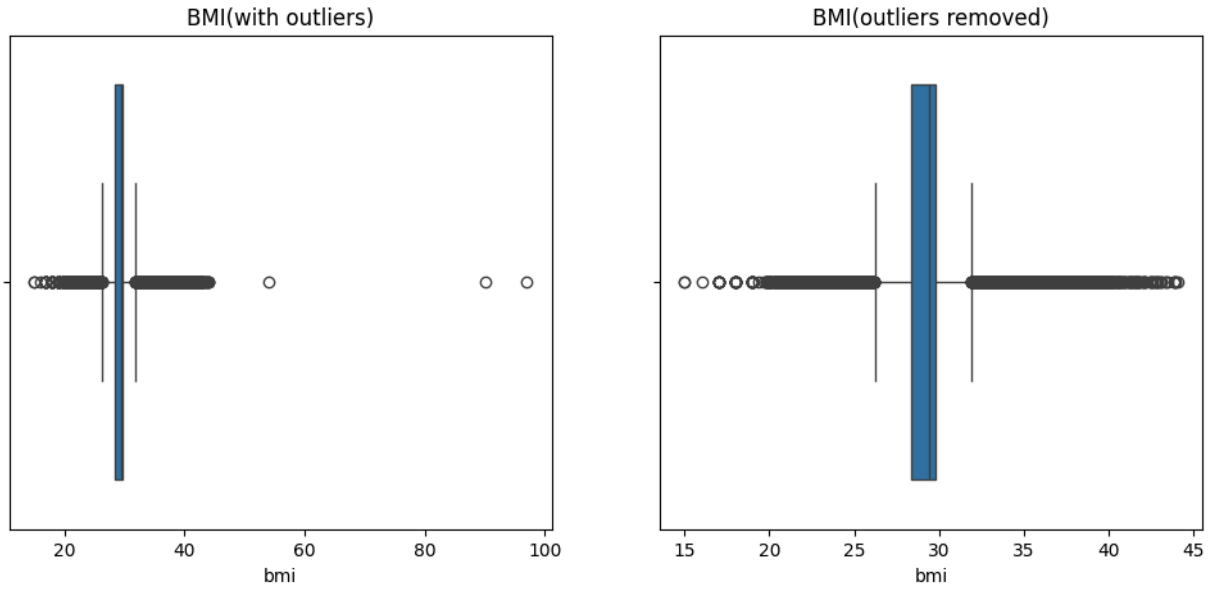
Feature	Correlation
patient_race	1.000000
payer_type	0.097592
Average of Sep-13	0.041480
Average of Mar-14	0.039992
Average of Jan-17	0.026429
Average of Dec-18	0.023317
Average of Nov-18	0.020455
income_household_under_5	0.019279
home_ownership	0.019279
family_dual_income	0.019279

Table 10: Top 10 absolute correlation values based on actual values across all columns and null values present on patient race.

Feature	Correlation
patient_race	1.000000
Average of Dec-13	0.139835
Average of Jan-18	0.139678
Average of Nov-14	0.137634
Average of Dec-17	0.136735
Average of Jan-14	0.135886
Average of Nov-18	0.133140
Average of Nov-13	0.132264
Average of Jan-15	0.131687
rent_burden	0.131678

3.3 Filtering Outliers

Some of the numeric features in the dataset contain outliers. The average **bmi** value in the dataset is 29.2, with three values falling above a 50 BMI threshold. BMI Categories according to the CDC are: Underweight = $<$ 18.5; Normal weight = 18.5–24.9; Overweight = 25–29.9; Obesity = BMI $>$ 30. Given that these values are extreme outliers, they will be removed from the dataset. See Figure 2 for the distribution of **bmi** values before and after outlier removal.



3.4 Dimensionality reduction

Given the large number of features in our dataset and the fact that most features do not pertain directly to the patients themselves but to the broader areas they live in (such as household income percentages, resident age demographics and location racial demographics), we decided to use

Principal Component Analysis (PCA). This will help us to drastically reduce the dimensionality of our dataset while also preserving as much information as possible from those combines features.

We fit the PCA model with a total of 136 features (we excluded features directly related to individual patient characteristics and categorical data). The results (Figure 2) showed that with just one principal component, we are able to captuca 99% of the variance in these features. Given this significant percentage, we can simplify our dataset by dropping all of the features that were used during PCA transformation and only retain that first principal component. This approach simplifies our data without losing any essential information, making subsequent modeling and analysis more efficient and manageable.

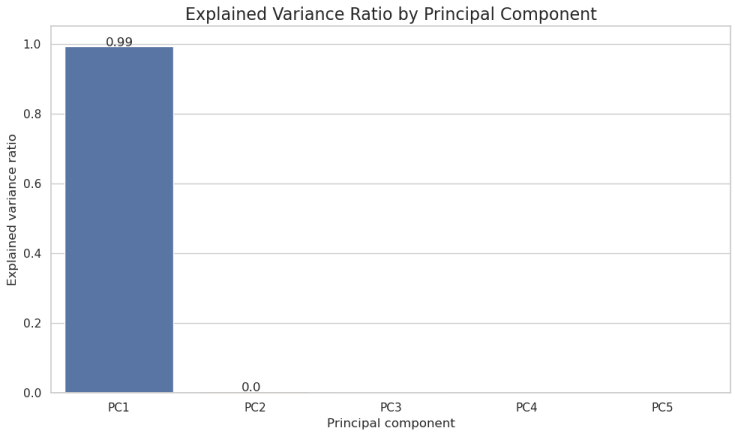


Figure 2: Explained variance ratio for the first 5 principal components.

After applying PCA, our final dataframe contains 12 features (including the target variable), which is a 92% reduction in terms of the initial number of features (152). The final features are detailed in Table 11.

Table 11: Final features after dimensionality reduction.

Feature	**Description**	**Value**
patient_zip3	Patient Zip3 (e.g. 190) on the metastatic date	Categorical
patient_age	Derived from Patient Year of Birth	Numeric
bmi	Earliest BMI recording post metastatic date	Numeric
metastatic_diagnosis_period	Period (in days) in which metastatic cancer was diagnosed	Numeric
patient_race	Patient race	Categorical
payer_type	Payer type on the metastatic date	Categorical
patient_state	Patient State on the metastatic date	Categorical
Region	Region of patient location	Categorical
Division	Division of patient location	Categorical
breast_cancer_diagnosis_code	ICD10 or ICD9 diagnoses code	Categorical
metastatic_cancer_diagnosis_code	CD10 diagnoses code	Categorical
PC1	First principal component	Numeric

3.5 Dummy encoding and numerical feature standardization

The final stage of our feature transformation involves dummy-encoding the categorical features and normalizing the numerical features. To achieve this efficiently, we created a custom transformation pipeline designed to perform each operation based on the specific type of column. Dummy-encoding was applied to all of the categorical features, converting them into binary columns. The numeric features were normalized using standard scaling, which adjusts the data to have a mean of zero and standard deviation of one. The fitting process was carried out separately on the training set to learn and adjust the necessary parameters, and the transformation was performed on the testing set.

4 METHODOLOGY

4.1 Exploratory Data Analysis

In this section, we will visually explore all the features related to individual patients in the dataset with the goal of gaining insights into the relationships and patterns of these features. This process will guide us towards.

Figure 3 shows patient counts by race. The largest racial category in the dataset is White with 3,565 patients, followed by Black and Hispanic. Asian is the smallest category, with 373 patients. This disproportionate distribution could potentially be due to the particular demographics of the locations that the data was acquired from.

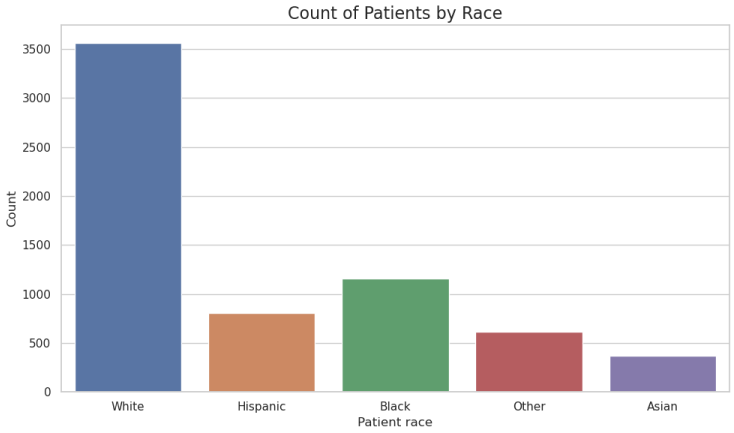


Figure 3: Count of patients by race.

When looking at the count of patients by health insurance type (Figure 4), we can see that most patients use commercial insurance, with Medicaid and Medicare Advantage having almost equal proportions. Commercial insurance is typically provided by private companies and is often provided by employers. It is generally more expensive in terms of premiums and out-of-pocket costs compared to other plans. Medicaid is a state and federally funded program that is typically used to cover individuals and families in lower-income households, with costs being generally very low or free. Medicare Advantage is typically provided by private companies approved by Medicare, with the additional benefits like dental, vision, and hearing.

We can further explore how these insurance categories are distributed based on racial category (Figure 5) in order to see if any patterns emerge. We can see that the largest demographic for all

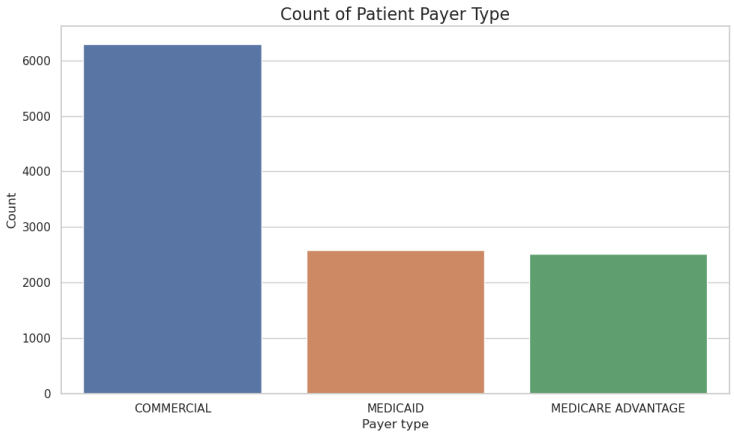


Figure 4: Count of patients by payer type.

three insurance types is White. For Blacks, Hispanics and Asians, the most common insurance types based on proportion are Medicaid. Those that are categorized as ”Other” in contrast tend to favor Commercial insurance above all else.

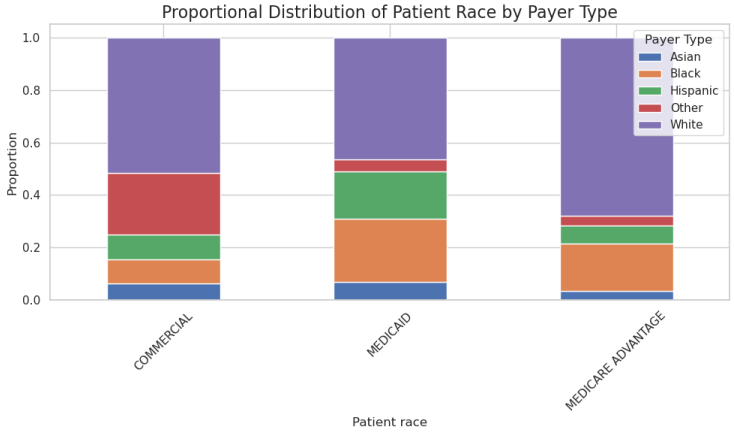


Figure 5: Proportional stacked bar chart of race by health insurance type.

When it comes to the location of the patients, we can see that the majority are located in California, followed by New York, Texas, and Illinois (Figure 6). Patients from other States are not as heavily represented compared to the aforementioned States.

The correlation matrices seen in the Figure 6 and 7 indicate very weak linear relationships between the variables BMI, metastatic diagnosis period, patient age, and median household income. Notably, there is a slightly negative correlation between BMI and median household income (correlation: -0.19), suggesting that higher BMI may be associated with lower income, though the relationship is weak. Other correlations, such as between BMI and age (correlation: -0.02), metastatic diagnosis period and age (correlation: -0.06), and between metastatic diagnosis period and household income (correlation: -0.02), also show very weak negative relationships. These weak correlations imply that no strong linear dependencies exist between these variables within the dataset. The findings suggest that additional variables or more complex statistical methods may be necessary to uncover more significant patterns or influences affecting these health-related

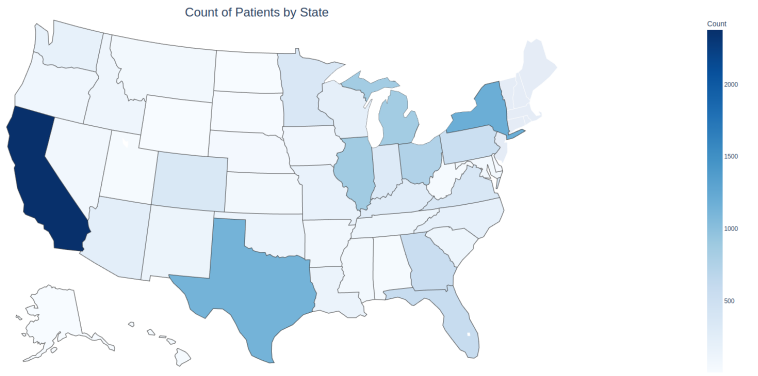


Figure 6: Choropleth visualization of the count of patients by State.

factors.

Figure 6 from the dataset highlights significant socio-economic and demographic relationships, particularly involving race and income. Strong positive correlations exist between being classified as white or Asian with higher household median incomes (0.65 and 0.70 respectively), while a moderate positive correlation is noted for being identified as black (0.34). Inter-racial relations reveal a strong negative correlation between white and Asian categories (-0.61), suggesting these groups are rarely represented together, contrasted by a strong positive overlap between the Asian and other categories (0.82). Regarding health metrics, there's a moderate negative correlation between BMI and being white (-0.42), indicating lower BMI in whites, and a moderate positive correlation with being black (0.33), suggesting higher BMI prevalence among blacks. These insights emphasize the intersection of race, economic status, and health, indicating potential areas for further socio-economic disparity and public health studies.

Next we will explore some of the numeric features that pertain to the individual patients. First we will explore our target variable, **metastatic_diagnosis_period** (MDP), which is the time (in days) until the metastatic cancer was diagnosed (Figure 9). Interestlingly, the distribution of MDP is highly right-skewed, with 25% of the patients having an MDP at or below 11 days and 50% of the patients at or below 44 days. This shows that most patients receive a diagnosis relatively early. However, there are cases where the diagnosis takes much longer. When comparing the distribution of MDP by race (Figure 10), we see that Hispanic patients tend to have the lowest median MDP at 32 days, while the highest one for "Other" patients is at 49.5 days.

Age on the other hand appears relatively bell-shaped, with a skew of 0.16, suggesting this feature is approximately normally distributed. The average age for a patient is 59 years old with a standard deviation of 13 years. This spread of age is to be expected, since the incidence of cancer is typically higher in older populations as opposed to younger individuals. When comparing across race (Figure 12), Hispanic patients have the lowest median age at 56 years, while White patients have the highest at 60.

Likewise, the distribution of patient BMI appears to also follow the normal distribution, with the exception of some outlier values which will be removed later (Figure 13). The average BMI for a patient is 29, which is close to the overweight threshold according to WHO standards. When we compare across the different racial groups (Figure 14), the differences in median BMI values are not significantly different from each other, with Black patients having the highest median BMI of 30.1 and Asian patients the lowest with a median BMI of 25.

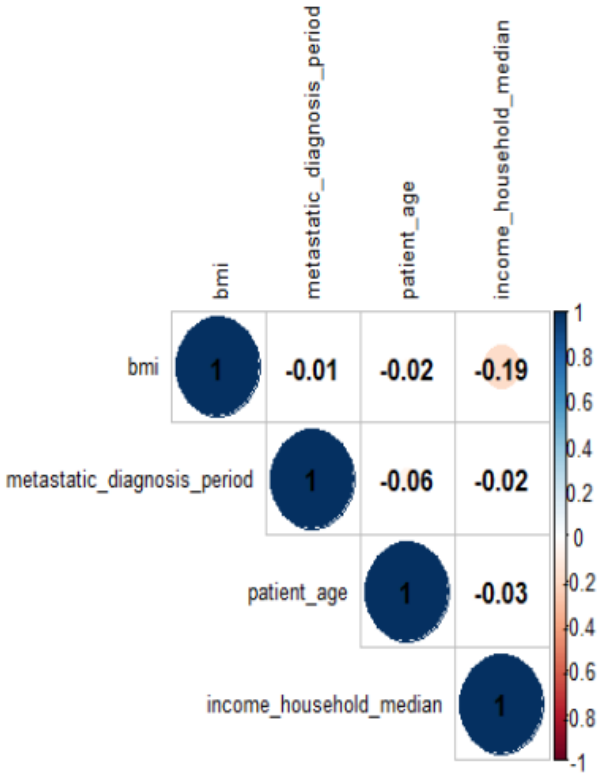


Figure 7: Correlation Matrix

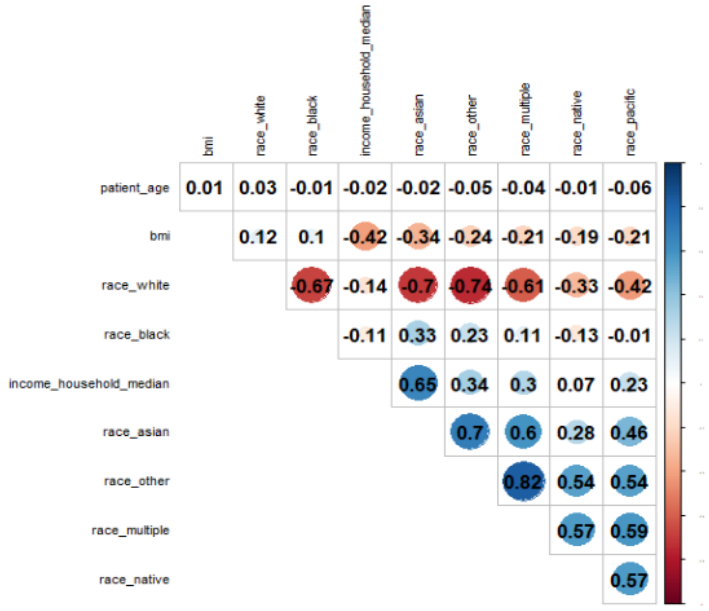


Figure 8: Correlation Matrix

4.2 Bayesian model development

In order to build an effective Bayesian model for predicting metastatic diagnosis period in patients with triple negative breast cancers, preliminary analysis of the data to determine key factors for the

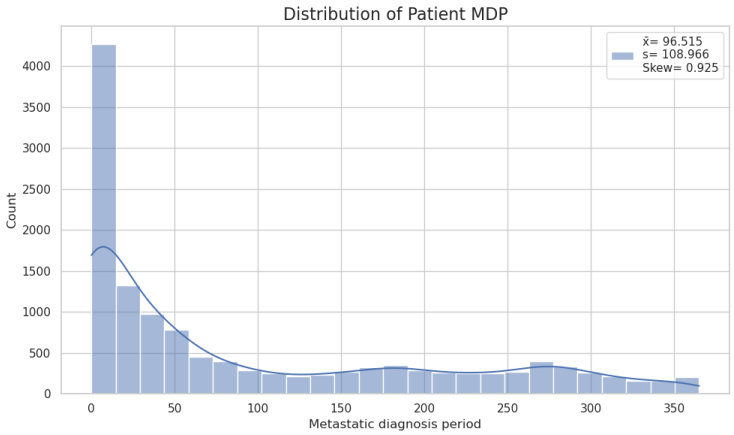


Figure 9: Distribution of patient metastatic diagnosis period (in days).

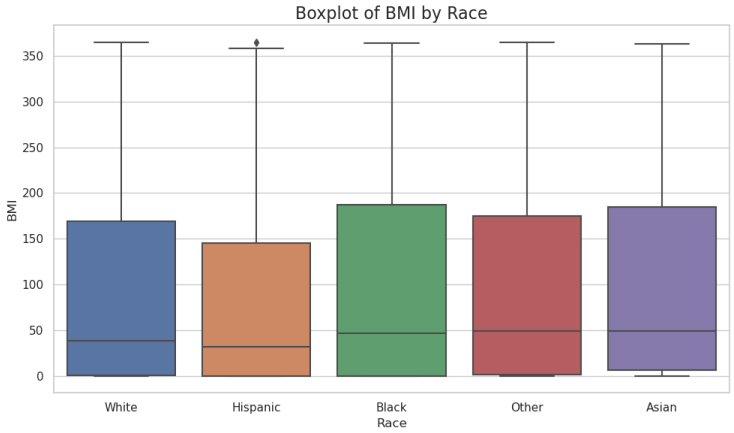


Figure 10: Distribution of patient metastatic diagnosis period by race.

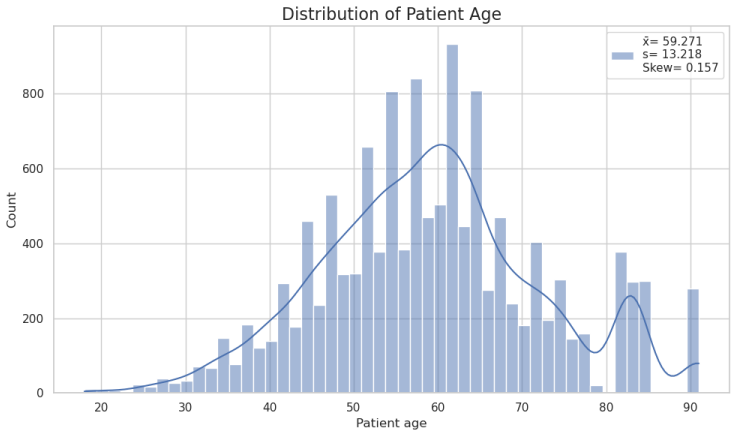


Figure 11: Distribution of patient age.

model was done; this, post exploratory analysis and data cleaning.

First, a correlation matrix between numerical variables was performed using the `cor()` func-

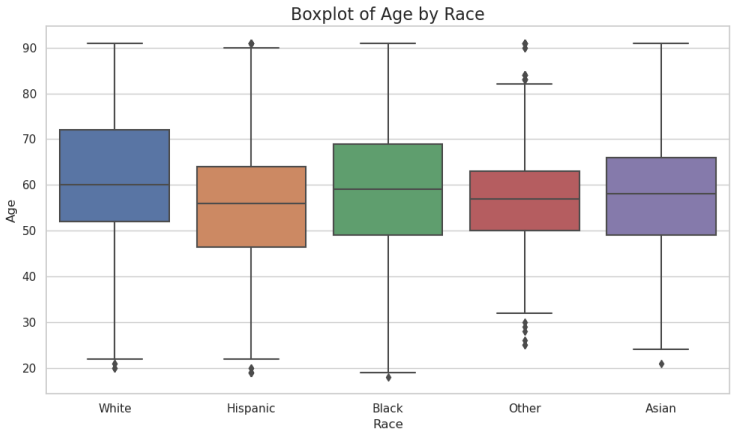


Figure 12: Distribution of patient age by race.

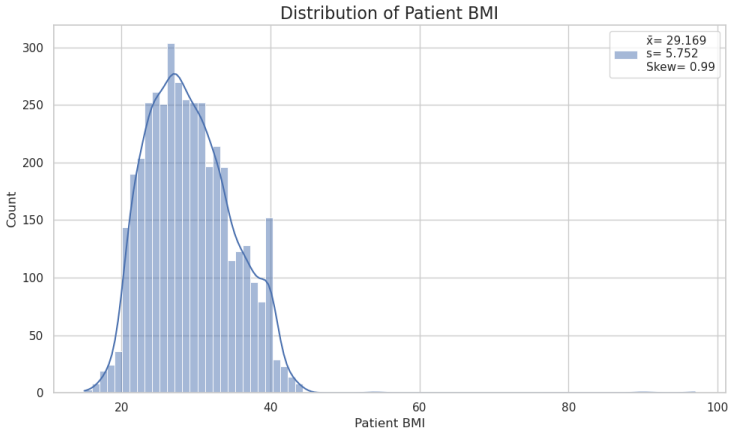


Figure 13: Distribution of patient BMI.

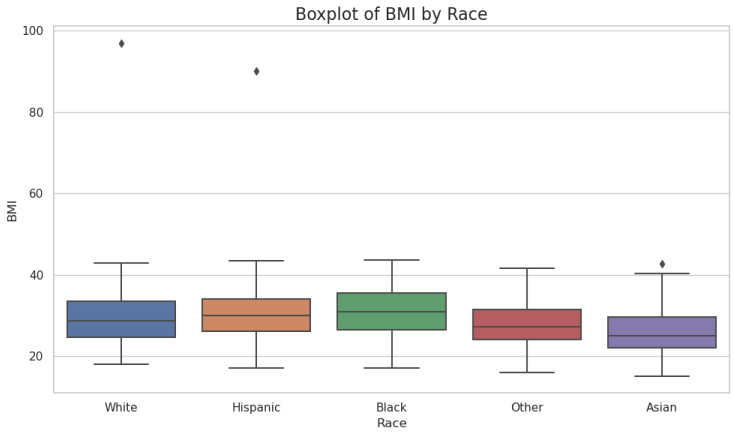


Figure 14: Distribution of patient BMI by race.

tion in RStudio to determine which of the explanatory variables were most highly correlated with the response. These tests determined that, of all numerical variables, **patient_age**, **bmi_null**, **la-**

bor_force_participation, **race_Asian**, **family_dual_income**, **education_bachelors**, **race_Hispanic**, and **income_household_10_to_15** had the highest correlation with the response **metastatic_diagnosis_period**. As such, these features were shortlisted for model development.

ANOVA (Analysis of Variance) testing was also necessary to test for statistical significance within categorical groups. This was accomplished using `aov()` in RStudio. The features **patient_state**, **breast_cancer_diagnosis_code**, and **metastatic_cancer_diagnosis_code** were tested; all showed significant differences between groupings. As such, Tukey HSD tests were implemented on the three ANOVA tests with `TukeyHSD()` in RStudio, to test differences in means between each group within their respective categorical variable. Due to output size from the large number of categories per variable, the mean of each Tukey test was calculated, whereby we can check for outliers within the Tukey tests using z-scores. For our purposes, a standard z-score of 2 was used to signify outliers from the mean. For these tests, we limited output to categoricals with over 100 observations due to the fact that categoricals with a low number of observations could show significance due to skewed values. For example, the **metastatic_cancer_diagnosis_code** feature showed significantly lower means for the values **C7880** and **C7900**; upon further inspection, it was found that both values appeared only once respectively in the data. As such, the distinction was made.

The results of the comparison of the filtered Tukey means returned no significance between categories between the categorical variables that had more than 100 observations. As such, none of these variables, **patient_state**, **breast_cancer_diagnosis_code**, and **metastatic_cancer_diagnosis_code** were considered for the model.

Finally, stepwise regression was performed on all numerical explanatory variables against our response, **metastatic_diagnosis_period** to determine significance. This was done using `step(lm())` in RStudio, with forward direction chosen. Of all variables, those that showed enough significance (p-score below .01) to be considered were: **patient_age**, **income_household_100_to_150**, **income_household_150_over**, **income_household_six_figure**, **education_less_highschool**, **education_highschool**, **education_some_college**, **Average.Of.Jul18**, **bmi_null**, **race_Asian**, **race_Hispanic**, **payer_COMMERCIAL**, and **payer_MEDICAID**, some seen above in the correlation matrix, corroborating previous findings. These were all added (if not previously present) into the shortlist of variables for model development.

Post-testing, the full list of variables chosen for the model is as follows: **patient_age**, **family_dual_income**, **income_household_10_to_15**, **income_household_100_to_150**, **income_household_150_over**, **income_household_six_figure**, **education_less_highschool**, **education_highschool**, **education_some_college**, **education_bachelors**, **labor_force_participation**, **Average.Of.Jul18**, **bmi_null**, **race_Asian**, **race_Hispanic**, **payer_COMMERCIAL**, and **payer_MEDICAID**.

In order to better develop a model with reduced noise, the dataframe post-PCA was chosen for the final model. In order to develop the model effectively, mean encoding was utilized for the categorical variables in the dataframe. This is a strategy in which categorical variables are replaced with the mean of the specific categorical for a feature. After encoding, a final stepwise regression was conducted to measure significance. Of the 12 features, which can be seen in Table 11, the final list of variables chosen was **PC1**, **patient_age**, **patient_state_mean**, **breast_cancer_diagnosis_code_mean**, and **metastatic_cancer_diagnosis_code_mean**.

In order to properly develop priors for the Bayesian model, the chosen variables were further analyzed. Summary statistics, box plots, and histograms were analyzed individually.

For **PC1**, the histogram showed a left-skew, although attempted transformations such as logarithmic, square root, and Box-Cox were unsuccessful. Due to the nature of PC1 containing the principal component which holds 99% of the variance from the original dataset, we chose to keep

the base values for the model.

patient_age showed a normal distribution, with **patient_state_mean** being skewed slightly to the left. Transformations, similarly to PC1, were unsuccessful in normalizing further but due to the variable significance it was decided to keep the slightly skewed values for the model.

The variables **breast_cancer_diagnosis_code_mean** and **metastatic_cancer_diagnosis_-code_mean** both showed non-normal distributions, which was expected due to the nature of the features (various different codes for different types of cancer). To normalize this, the diagnoses codes were categorized into low, medium, or high (1, 2, 3) based on the quartiles for the respective variables. Low was assigned to values between the minimum value and 1st quartile, medium for values between the 1st - 3rd quartile, and high for values between the 3rd quartile and maximum value.

For **PC1**, we used a Cauchy prior due to the wide range of values and large variance present in the principal component. The Cauchy prior allows us to express the belief that the parameter estimate is likely to be near the median, but with significant uncertainty, which we believe fits the characteristics of the principal component. An uninformative Cauchy was used to model uncertainty.

For **patient_age** and **patient_state_mean**, the mean and standard deviation were calculated due to the normal or near-normal distributions exhibited. For **patient_age**, the mean was 59.27 with a standard deviation of 13.21888. For **patient_state_mean**, the mean was 96.53 with a standard deviation of 10.60562. These were the parameters chosen to inform the priors given their distributions, allowing us to implement our beliefs on the distributions. Specifically choosing the mean and standard deviation of a variable allows us to inform the model with our prior belief about the distribution of the data, which is why these statistics were chosen for their respective variables.

For the remaining variables, **breast_cancer_diagnosis_code_mean** and **metastatic_cancer-_diagnosis_code_mean**, ordinal logistic regression was chosen to be implemented directly in the Bayesian model due to the categorical implementation discussed previously. For each category (low, medium, high) for the variables, the mean and standard deviation were calculated and implemented into the model to allow us to inform the prior, similarly to above.

The final model was developed with the specifications above, using the JAGS package in RStudio along with the CODA package in order to sample for convergence post-model updating. For this model, we decided upon 4 MCMC chains, with burn-in of 10,000 and iterations of 20,000 to allow for effective sampling and convergence towards our posterior distributions. Note that prior and posterior distributions were developed with the syntax ”beta_variable”; for example, moving forward with the model, **PC1** will appear as **beta_PC1**. Note that for the categorical variables e.g. **metastatic_cancer_diagnosis_code_mean**, there were three separate priors developed based on the means and standard deviations of each category. For example, the low category for the variable above is encoded as **beta_mcd[1]**, etc. For testing purposes, there was an alternative model developed in order to see if accuracy could be improved; the second model contained all of the variables as the first model, although used all uninformative priors as opposed to those determined by means. This way, the model would allow the data to have an uninterrupted effect on the posterior means and thus the predictions.

To assess convergence, trace plots, Geweke diagnostics, and effective sample sizes were checked. Trace plots indicate the mean for each MCMC chain, where plots that appear closer together indicate better convergence towards some posterior. For an example, please see **beta_PC1**'s trace plot as compared to **beta0**'s trace plot. Of note is that in Bayesian models, beta0 indicates the base estimate when all other parameters are zero, similar to linear regression models. The trace plots for all parameters can be found at... where convergence looks likely for all parameters,

although the plots for **beta0**, **beta_age** and **beta_state** indicate lower confidence in convergence. Next, Geweke diagnostics were assessed for each chain, for each variable. Geweke diagnostics diagnose convergence by comparing the distribution means between the first 10% and the final 50% of an MCMC chain as per standard procedure. These percentages may vary but for this model there was no reason to deviate. Z-scores are calculated for each variable, where a z-score close to 0 indicates effective convergence and a z-score greater than 2 indicates a lack of convergence. The full diagnostics for each chain can be found at... where we can see some sort of convergence for all parameters in all chains, except for **beta_mcd[3]** in chain number 4. Lastly, effective sample sizes were checked, where a higher ESS indicates more effective sampling and likelihood for convergence by comparing the number of independent samples from the posterior distribution needed to equate to the correlated samples obtained from the MCMC. Higher sample sizes are desirable due to improved precision (i.e. posterior measures of mean, variance, etc.), reduced autocorrelation, and better assessment of convergence. For this model, **beta0**, **beta_age** and **beta_state** all exhibited lower sample sizes than the remaining variables indicating lower effective sample sizes. However, when taken into context with the acceptable trace plots and Geweke diagnostics, we decided these variables did not need to be removed from the model. Convergence for our alternative model showed similar results, with acceptable and comparable plots and statistics.

In order to model the predicted values using JAGS, we sampled for **mu[i]** where mu is the linear predictor based on the priors and i corresponds to each individual observation from our data, **train_pca**. Note that due to needing comparison against observed values for calculation of accuracy statistics such as RMSE, MAE, and MSE, **test_pca** was not utilized. As mentioned previously, the model used 4 MCMC chains at 20,000 iterations, leaving us with 80,000 total posterior samples to draw from for predictions (Where each sample has 13,173 predictions for each observation, for a total of 1,053,840,000 predicted values. Due to limitations in computational power, using all of these values was not possible. Of tests with 1,000, 5,000, and 20,000 samples, 1,000 and 5,000 sampled correctly but RStudio encountered fatal errors at 20,000; this could be remediated with larger, more powerful devices in future studies. In order to calculate each **mu[i]**, each prediction for the respective observation was placed into a data frame, where each row held the predictions of one of the MCMC chains. So, row [1,1] corresponds to the first MCMC chain's prediction for observation 1, [1,2] corresponds to the first MCMC chain's prediction for observation 2, etc. Then, column means were taken to determine the final predicted value of all of the sampled chains. The first testing runs were implemented with 1,000 posterior samples and the respective RMSE, MAE, and MSE of each model can be seen below:

Sample Size	Model	RMSE	MAE	MSE
1000	Model 1	411.5105	312.7993	169340.9
1000	Model 2	86.7966	66.90932	7533.649

Table 12: Comparison of Bayesian Accuracy Statistics for Models with 1000 Samples

As can be seen, the uninformative model which allows the data to more naturally determine the convergence of the MCMC chain vastly outperforms the original model, with informed priors. For this reason, we tested 5,000 samples only with this model, in order to see if the statistics were in-line with expectations and could be improved. The following table shows the results:

The results remained similar to those with 1,000 posterior samples, which indicates that the MCMC chains were effectively sampling from one another along the 20,000 iterations. A large variance in the statistics would have been an indicator of ineffective sampling across chains. A final test was conducted with model 2 at 5,000 samples with similar results, confirming model

Sample Size	Model	RMSE	MAE	MSE
5000	Model 2	86.79611	66.91311	7533.566

Table 13: Bayesian Accuracy Statistics for Model 2 with 5000 Samples

compilation and sampling consistency.

4.3 Model fitting

After our dataset was cleaned and transformed, it was divided into a training and testing set, with 20% of the data allocated for testing. Additionally, an extra validation split was created comprising of 25% of the training data to be used at the final stage of model evaluation.

We fitted and compared eight different regression models: Linear Regression (LR), Support Vector Regressor (SVR), Random Forest Regressor (RFR), Gradient Boosting Regressor, AdaBoost Regressor (ABR), MLP Regressor (MLPR), XGBoost Regressor (XGBR), and CatBoost Regressor (CBR). Each model was evaluated using their default hyperparameters. To assess the performance of these models, we used four different metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared (R^2). Results of this initial model evaluation stage are listed in Table 14.

Out of all models tested, Gradient Boosting Regressor had the best predictive performance. Therefore, this model was selected as the final model for further development.

4.4 Hyperparameter tuning

The next step involves further tuning the GBR model using a grid search method to identify the optimal hyperparameters. Table 15 describes the purpose of each of the tuned hyperparameters. The grid search model used threefold cross-validation, resulting in a total of 3,888 fits to evaluate each hyperparameter. The final tuned GBR model was then used to make the final predictions.

Table 14: Evaluation metrics for different regression models.

Model	MSE	RMSE	MAE	R-Squared
Gradient Boosting Regressor	6853.742	82.787	64.399	0.418
MLP Regressor	6866.456	82.864	62.735	0.417
CatBoost Regressor	6873.768	82.908	63.200	0.416
Random Forest Regressor	7187.995	84.782	64.256	0.390
XGB Regressor	7278.157	85.312	64.782	0.382
Bayesian Regression	7533.566	86.796	66.913	NA
AdaBoost Regressor	8000.410	89.445	76.525	0.321
Support Vector Regressor	12047.027	109.759	77.223	-0.023
Linear Regression	1.49E+21	3.86E+10	1.06E+09	-1.26E+17

Table 15: Hyperparameters and values used for grid search.

Hyperparameter	Tested Values	Description
learning_rate	0.01, 0.05, 0.1	Scales the contribution of each tree in the ensemble
max_depth	3, 4, 5, 6	Limits the maximum depth of each tree
max_features	'sqrt', 'log2', None	Number of features to consider when looking for the best split
min_samples_leaf	1, 2, 4	Minimum number of samples required to be at a leaf node
min_samples_split	2, 5, 10	Minimum number of samples required to split an internal node
n_estimators	100, 200, 300, 500	Number of sequential trees to be modeled
subsample	0.6, 0.8, 1.0	The fraction of samples to be used for fitting each individual base learner

5 RESULTS

5.1 Final model performance

After completing the hyperparameter tuning, the final model was trained on the training set and evaluated on the validation and test sets. The performance metrics are summarized in Table 16.

Table 16: Final model metrics for Gradient Boosting Regressor model.

Set	MSE	RMSE	MAE	R-Squared
Train	5816.047	76.263	58.809	0.516
Validation	6403.681	80.023	61.598	0.452
Test	6737.072	82.080	62.850	0.419

5.2 Feature importance

To understand what features affect the model's decisions the most, we analyzed feature importance. Since GBR is an ensemble model consisting of multiple sequential decision trees, the importance of a feature is determined based on the effectiveness of each feature in splitting the dataset during the tree-building process.

The most significant features in our model are breast cancer diagnosis codes. This result indicates that specific diagnoses or conditions related to breast cancer play a crucial role in influencing the time it takes to reach a diagnosis. The characteristics of the breast cancer itself could potentially influence the diagnostic process and introducing time delays. A patient's age is another significant factor that influences the duration of the diagnosis period. Older or younger ages may be associated with different speeds in diagnosis. BMI also plays a crucial role due to the health risks associated with being overweight, which may complicate the detection and diagnosis of breast cancer. The geographical location of a patient, as indicated by the ZIP code, is another important factor. This suggests that the accessibility and quality of healthcare services might affect the diagnosis process.

To further understand the individual impact of each feature on the predictions made by the GBR model, we can utilize partial dependence plots (PDP) to illustrate that effect. From Figure 16, we can observe that diagnosis codes have an increasing trend, indicating that as the value of these codes increases, so does the predicted time to diagnose the cancer. For patient age, there are relatively stable predictions for a range of ages, but with sharp increases or peaks at specific ages. This could imply that specific age ranges may effect the diagnosis speed of breast cancer. The plot for BMI shows very little change in the model's prediction across different BMI values,

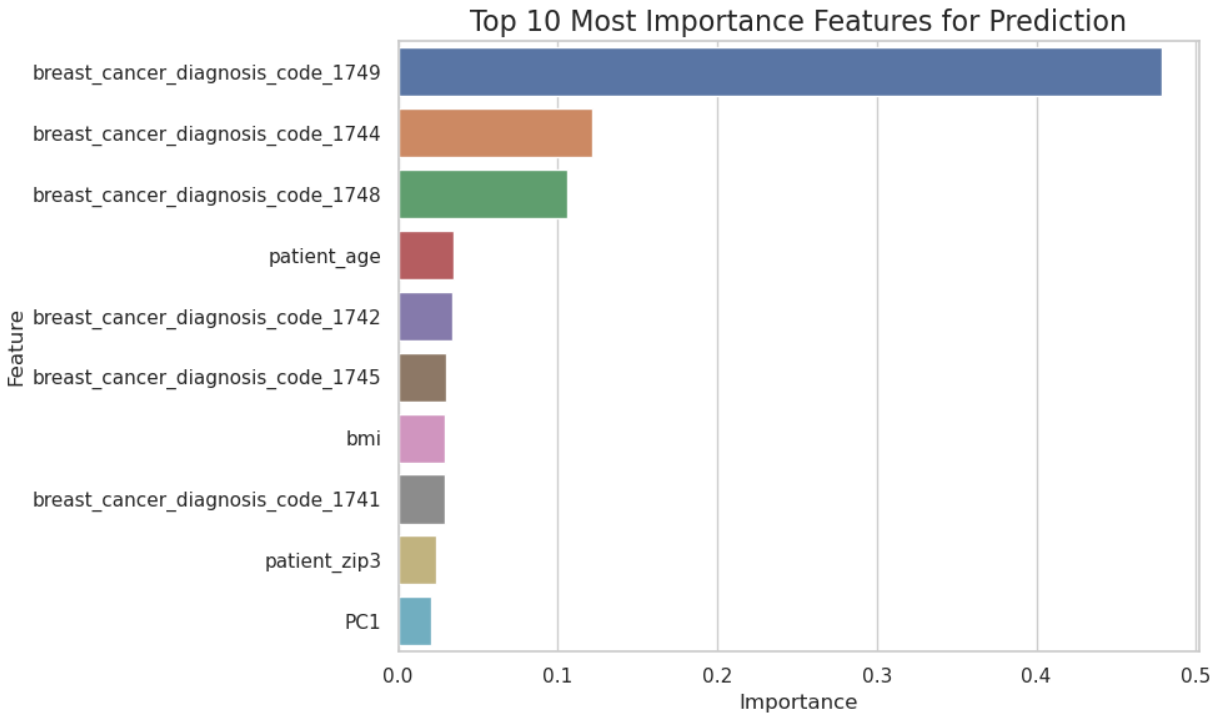


Figure 15: Top 10 most important feature of the final GBR model.

Table 17: Descriptions of the diagnosis code features.

Code	Description
1749	Malignant neoplasm of breast (female), unspecified
1748	Malignant neoplasm of other specified sites of female breast
1744	Malignant neoplasm of upper-outer quadrant of female breast
1741	Malignant neoplasm of central portion of female breast
1745	Malignant neoplasm of lower-outer quadrant of female breast
1742	Malignant neoplasm of upper-inner quadrant of female breast

meaning it does not have a strong independent effect on the model's predictions. However, this doesn't mean BMI is not important, but its effect could be non-linear or dependent on interactions with other features. The plot for zip3 is also mostly flat, meaning geographic location has little to no direct influence on the predictions, assuming other factors are constant.

After examining the individual impacts of each feature, we then utilized SHAP plots to explore the features with interactions taken into account. This will enable us to visualize how these interactions between different features might influence the outcome (Figure 17). Similar to the partial dependence analysis, breast cancer diagnosis codes have the highest positive impact on the model output, further confirming that they are highly predictive of the time until diagnosis. In contrast, patient age has a wide distribution of effects across the model output. Large ages tend to decrease the diagnosis period and in some cases increase it (this would explain that upward trend on the partial dependence plot).

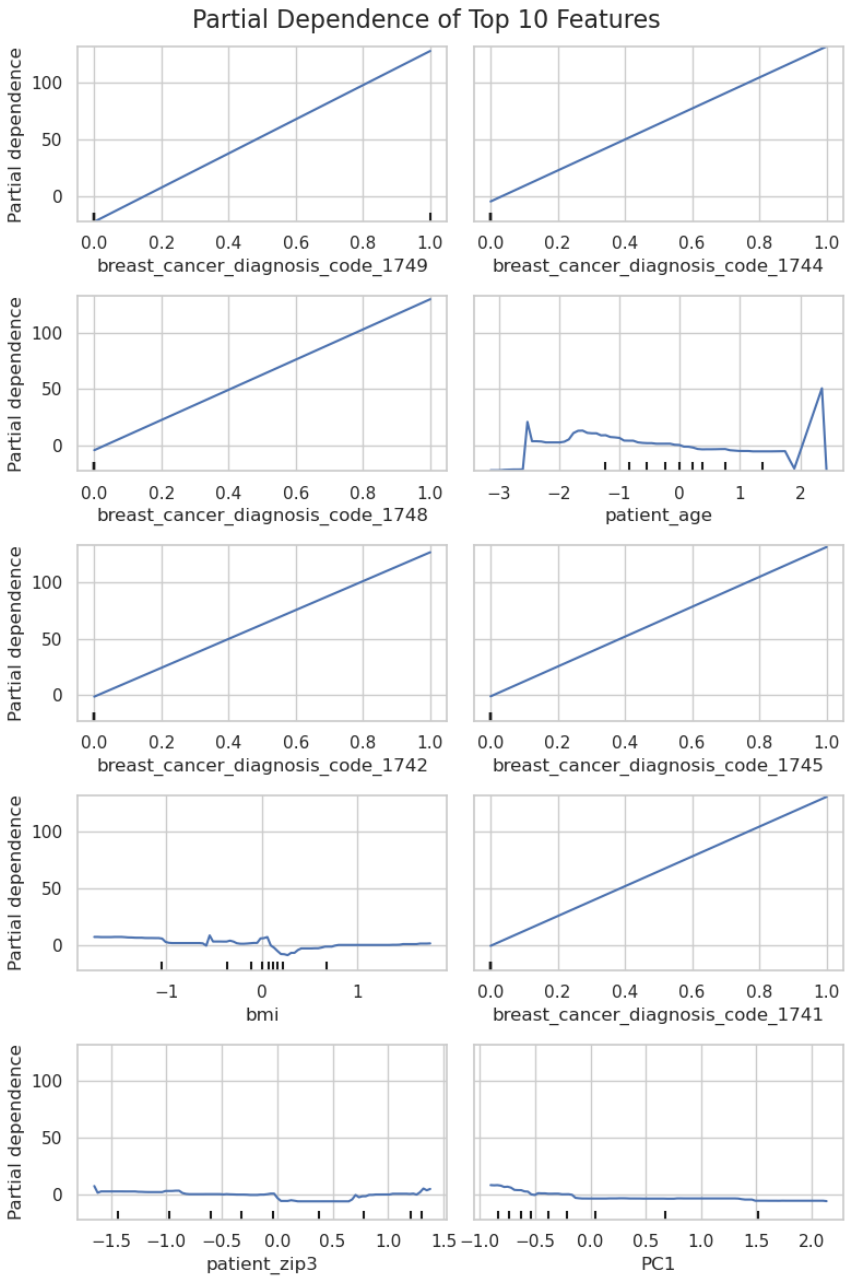


Figure 16: Partial dependence of different features on the model outcome.

6 DISCUSSION

In this paper, we conducted an extensive study that aimed to improve delays in breast cancer diagnosis time utilizing various machine learning methods. Using a rich dataset from the Women in Data Science (WiDS) datathon, our primary objective was to investigate the impact of socioeconomic, demographic, and clinicopathologic factors on the time to diagnosis of metastatic triple-negative breast cancer. After preprocessing of the data, feature engineering, and selection of the model parameters, our results illustrate the complexity of the problems related to cancer diagnosis and

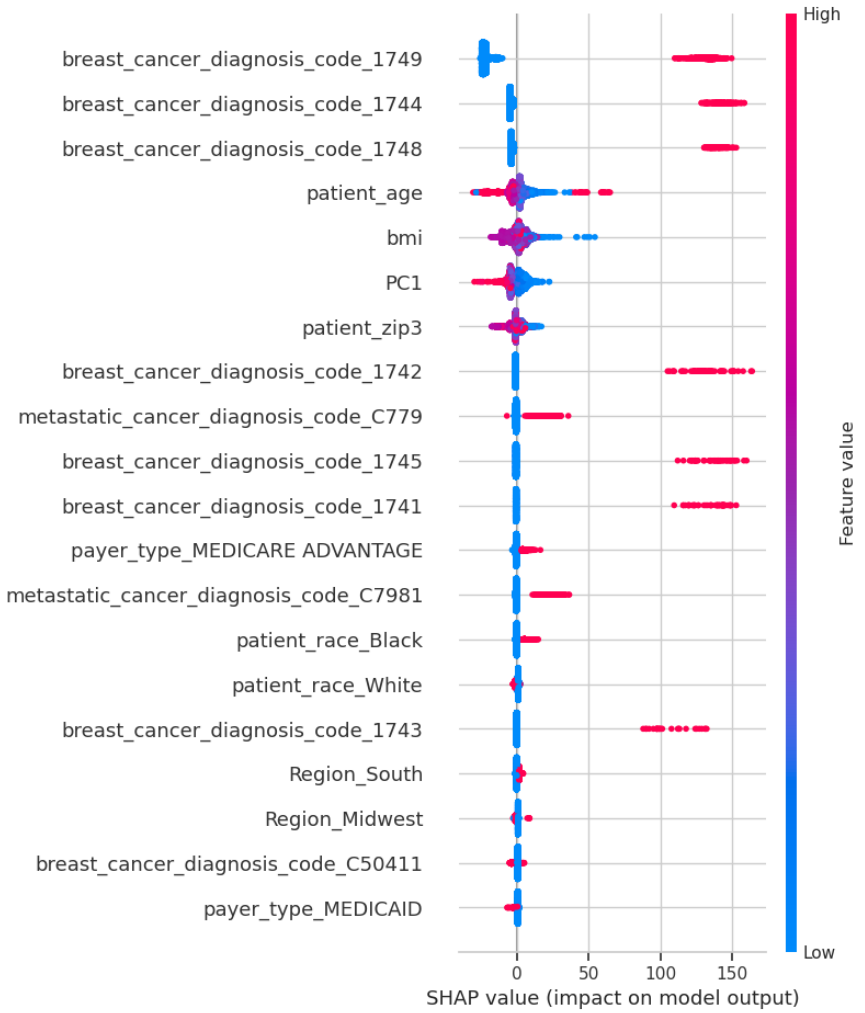


Figure 17: Shapley values of different features on the model outcome.

the potential of machine learning in healthcare.

After testing various regression models, Gradient Boosting Regressor proved to be a powerful predictor, outperforming the other models reviewed. In addition, partial dependence and SHAP analyses were used to explore, in greater detail, factors that were more influential in the prediction of diagnosis periods (eg, age, geography, specific diagnostic codes). The results highlight the potential that machine learning offers in the realm of healthcare decision making. These advanced analytic methods if used in healthcare can help providers in increasing accurate and timely diagnosis and can result in enhanced patient care leading to better outcomes.

In future work, several key areas can be explored to further enhance the predictive accuracy and utility of our findings in metastatic breast cancer diagnosis.

A Appendix A: Feature Descriptions

Table 18: Features in the dataset and their descriptions.

Feature	Description
patient_id	Unique identification number of patient
patient_race	Patient race
payer_type	Payer type on the metastatic date
patient_state	Patient State on the metastatic date
patient_zip3	Patient Zip3 (e.g. 190) on the metastatic date
patient_age	Derived from Patient Year of Birth
patient_gender	F, M on the metastatic date
bmi	Earliest BMI recording post metastatic date
breast_cancer_diagnosis_code	ICD10 or ICD9 diagnoses code
breast_cancer_diagnosis_desc	ICD10 or ICD9 code description
metastatic_cancer_diagnosis_code	ICD10 diagnoses code
metastatic_first_novel_treatment	Generic drug name of the first novel treatment after metastatic diagnosis
metastatic_first_novel_treatment_type	Description of first novel treatment after metastatic diagnosis
region	Region of patient location
division	Division of patient location
population	An estimate of the zip code's population.
density	The estimated population per square kilometer.
age_median	The median age of residents in the zip code.
male	The percentage of residents who report being male (e.g. 55.1).
female	The percentage of residents who report being female (e.g. 44.9).
married	The percentage of residents who report being married (e.g. 44.9).
family_size	The average size of resident families (e.g. 3.22).
income_household_median	Median household income in USD.
income_household_six_figure	Percentage of households that earn at least $100,000 (e.g. 25.3)
home_ownership	Percentage of households that own (rather than rent) their residence.
housing_units	The number of housing units (or households) in the zip code.
home_value	The median value of homes that are owned by residents.
rent_median	The median rent paid by renters.
education_college_or_above	% of residents with at least a 4-year degree.
labor_force_participation	% of residents 16 and older in the labor force.
unemployment_rate	% of residents unemployed.
race_white	% of residents who report their race White.
race_black	% of residents who report their race as Black or African American.
race_asian	% of residents who report their race as Asian.
race_native	% of residents who report their race as American Indian and Alaska Native.
race_pacific	% of residents who report their race as Native Hawaiian and Other Pacific Islander.
race_other	% of residents who report their race as Some other race.
race_multiple	% of residents who report their race as Two or more races.
hispanic	% of residents who report being Hispanic.
age_under_10	% of residents aged 0-9.
age_10_to_19	% of residents aged 10-19.
age_20s	% of residents aged 20-29.
age_30s	% of residents aged 30-39.
age_40s	% of residents aged 40-49.
age_50s	% of residents aged 50-59.
age_60s	% of residents aged 60-69.
age_70s	% of residents aged 70-79.
age_over_80	% of residents aged over 80.
divorced	% of residents divorced.
never_married	% of residents never married.
widowed	% of residents never widowed.
family_dual_income	% of families with dual income earners.
income_household_under_5	% of households with income under $5,000.
income_household_5_to_10	% of households with income from $5,000-$10,000.

Table 18 continued from previous page

income_household_10_to_15	% of households with income from $10,000-$15,000.
income_household_15_to_20	% of households with income from $15,000-$20,000.
income_household_20_to_25	% of households with income from $20,000-$25,000.
income_household_25_to_35	% of households with income from $25,000-$35,000.
income_household_35_to_50	% of households with income from $35,000-$50,000.
income_household_50_to_75	% of households with income from $50,000-$75,000.
income_household_75_to_100	% of households with income from $75,000-$100,000.
income_household_100_to_150	% of households with income from $100,000-$150,000.
income_household_150_over	% of households with income over $150,000.
income_individual_median	The median income of individuals in the zip code.
poverty	The median value of owner occupied homes.
rent_burden	The median rent as a % of the median renter's household income.
education_less_highschool	% of residents with less than a high school education.
education_highschool	% of residents with a high school diploma but no more.
education_some_college	% of residents with some college but no more.
education_bachelors	% of residents with a bachelor's degree (or equivalent) but no more.
education_graduate	% of residents with a graduate degree.
education_stem_degree	% of college graduates with a Bachelor's degree or higher in a STEM field.
self_employed	% of households reporting self-employment income on their 2016 IRS tax return.
farmer	% of households reporting farm income on their 2016 IRS tax return.
disabled	% of residents who report a disability.
limited_english	% of residents who only speak limited English.
commute_time	Median commute time of resident workers in minutes.
health_uninsured	% of residents who report not having health insurance.
veteran	% of residents who are veterans.
Average of Jan	72 columns showing the zip 3 Monthly Average Temperature

B Appendix B: Author Contributions

Sarah Hashmi Helped with the exploratory data analysis work by creating correlation coefficient matrixes, contributed to the making of the presentation, and literature review.

George Tzimas Missing data imputation, dimensionality reduction, detailed test of regression models, hyperparameter tuning, feature importance analysis.

Eric Piatek conducted data transformations through dummy variables and mean encoding, exploratory analysis with stepwise regression, and the full development of the Bayesian model as an alternative to traditional ML methods.

Esmeralda Villela presented the WiDS 2024 University Challenge Project to the team after the group was formed, performed data exploration, preliminary regression models (Random Forest, XGBoost, and Ridge Regression) for the project updates.

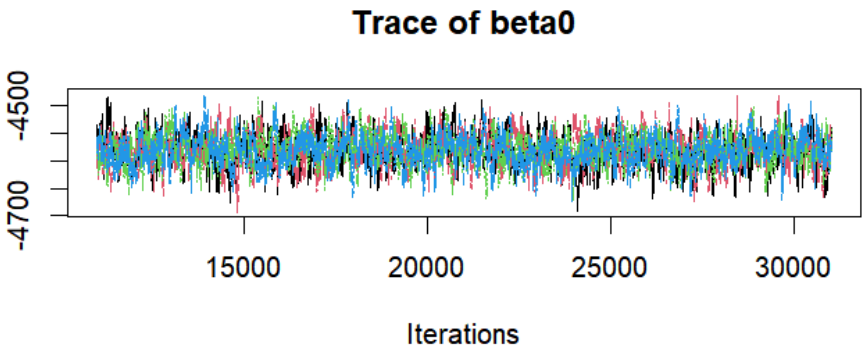


Figure 18: Trace plot for Parameter β_0

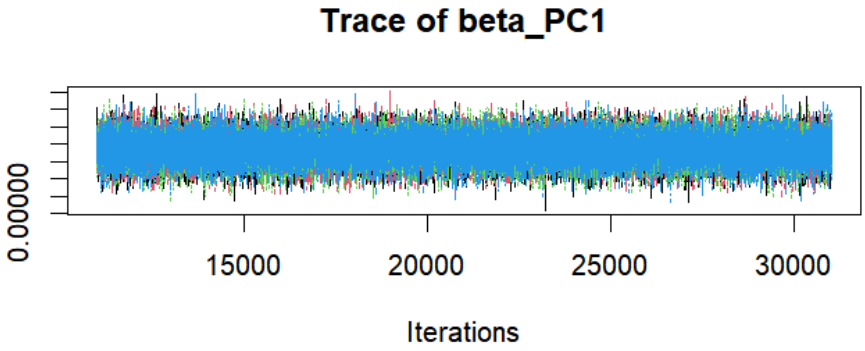


Figure 19: Trace plot for Parameter β_{PC1}

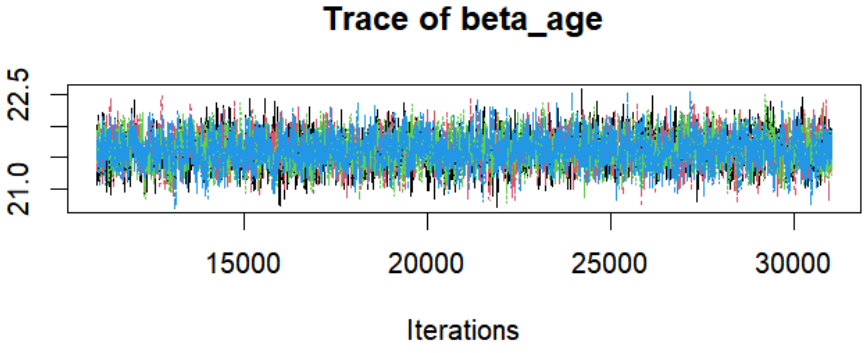


Figure 20: Trace plot for Parameter β_{age}

C Appendix C: Bayesian Model Output

C.1 Trace Plot for Parameter β_0

C.2 Trace Plot for Parameter β_{PC1}

C.3 Trace Plot for Parameter β_{age}

C.4 Trace Plot for Parameter $\beta_{\text{bcd}[1]}$

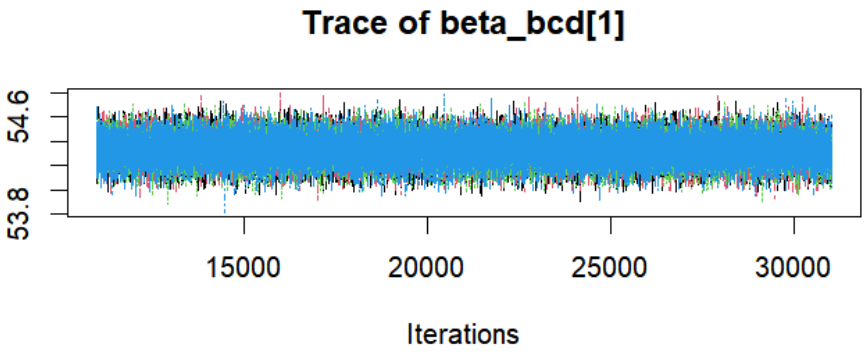


Figure 21: Trace plot for Parameter $\beta_{\text{bcd}[1]}$

C.5 Trace Plot for Parameter $\beta_{\text{bcd}[2]}$

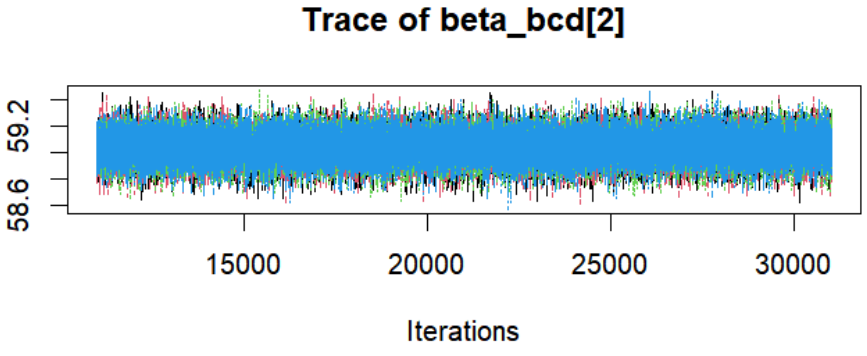


Figure 22: Trace plot for Parameter $\beta_{\text{bcd}[2]}$

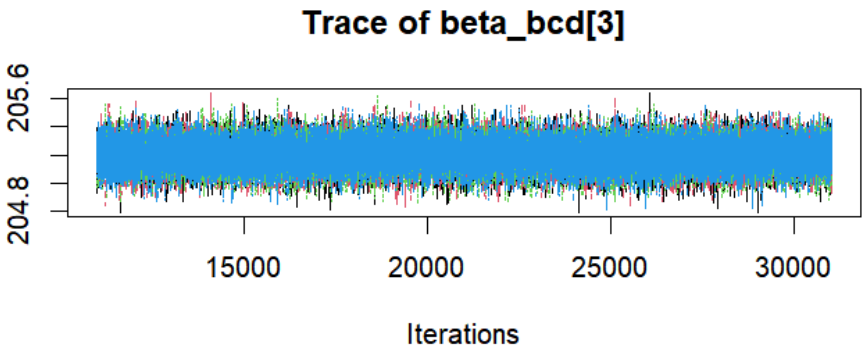


Figure 23: Trace plot for Parameter $\beta_{\text{bcd}[3]}$

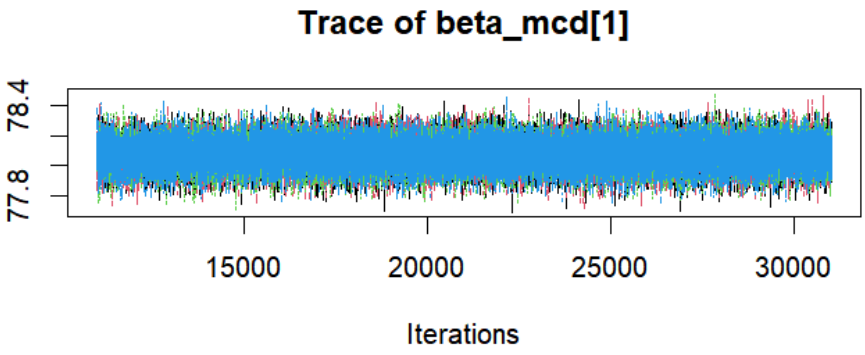


Figure 24: Trace plot for Parameter $\beta_{\text{mcd}[1]}$

C.6 Trace Plot for Parameter $\beta_{\text{bcd}[3]}$

C.7 Trace Plot for Parameter $\beta_{\text{mcd}[1]}$

C.8 Trace Plot for Parameter $\beta_{\text{mcd}[2]}$

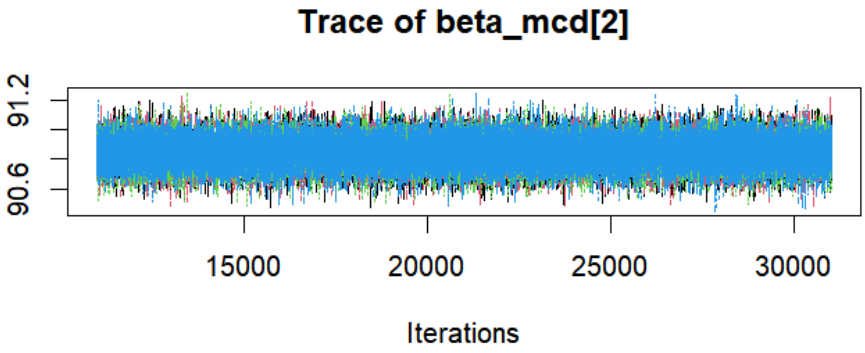


Figure 25: Trace plot for Parameter $\beta_{\text{mcd}[2]}$

C.9 Trace Plot for Parameter $\beta_{\text{mcd}[3]}$

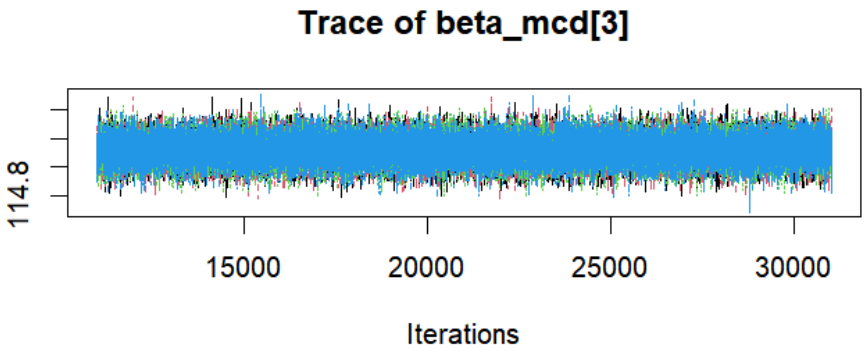


Figure 26: Trace plot for Parameter $\beta_{\text{mcd}[3]}$

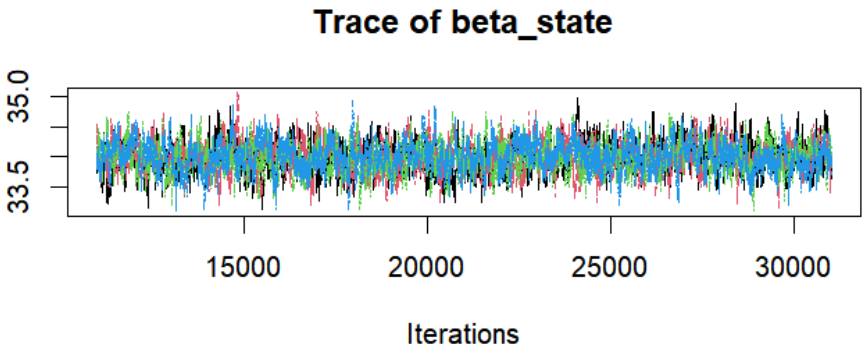


Figure 27: Trace plot for Parameter β_{state}

Table 19: Geweke output for MCMC 1

β_0	β_{PC1}	β_{age}	$\beta_{bcd[1]}$	$\beta_{bcd[2]}$	$\beta_{bcd[3]}$	$\beta_{mcd[1]}$	$\beta_{mcd[2]}$	$\beta_{mcd[3]}$	β_{state}
1.5445	0.0073	-1.7956	0.6790	1.0577	0.6250	-0.9549	1.3697	0.4983	-1.3386

Table 20: Geweke output for MCMC 2

β_0	β_{PC1}	β_{age}	$\beta_{bcd[1]}$	$\beta_{bcd[2]}$	$\beta_{bcd[3]}$	$\beta_{mcd[1]}$	$\beta_{mcd[2]}$	$\beta_{mcd[3]}$	β_{state}
-0.6756	-0.1335	0.7907	0.7439	1.1260	0.4175	-0.6133	0.4801	0.6981	0.5180

Table 21: Geweke output for MCMC 3

β_0	β_{PC1}	β_{age}	$\beta_{bcd[1]}$	$\beta_{bcd[2]}$	$\beta_{bcd[3]}$	$\beta_{mcd[1]}$	$\beta_{mcd[2]}$	$\beta_{mcd[3]}$	β_{state}
-0.1846	0.1304	0.0301	0.0771	-0.6232	-0.1403	-0.0681	0.0574	-1.1163	0.2598

C.10 Trace Plot for Parameter β_{state}

C.11 Geweke Output for MCMC 1

C.12 Geweke Output for MCMC 2

C.13 Geweke Output for MCMC 3

C.14 Geweke Output for MCMC 4

Table 22: Geweke output for MCMC 4

β_0	β_{PC1}	β_{age}	$\beta_{bcd[1]}$	$\beta_{bcd[2]}$	$\beta_{bcd[3]}$	$\beta_{mcd[1]}$	$\beta_{mcd[2]}$	$\beta_{mcd[3]}$	β_{state}
-1.2915	-1.7880	0.4436	-0.2340	0.3163	-0.0749	-1.2940	0.1266	-2.5412	1.4694

C.15 Effective Sample Sizes (ESS)

Table 23: Effective Sample Sizes (ESS)

β_0	β_{PC1}	β_{age}	$\beta_{bcd[1]}$	$\beta_{bcd[2]}$	$\beta_{bcd[3]}$	$\beta_{mcd[1]}$	$\beta_{mcd[2]}$	$\beta_{mcd[3}$
1061.472	49201.427	3295.205	80927.098	80000.000	80000.000	79326.926	79288.571	80000.0

References

[Binsaif, 2022] Binsaif, N. (2022). Application of machine learning models to the detection of breast cancer. *Mobile Information Systems*.

[Choi et al., 2009] Choi, J. P., Han, T. H., and Park, R. W. (2009). A hybrid bayesian network model for predicting breast cancer prognosis. *Journal of Korean Society of Medical Informatics*, 15(1):49–57.

[Dehdar et al., 2023] Dehdar, S., Salimifard, K., Mohammadi, R., Marzban, M., Saadatmand, S., Fararouei, M., and Dianati-Nasab, M. (2023). Applications of different machine learning approaches in prediction of breast cancer diagnosis delay. *Frontiers in Oncology*, 13:1103369.

[Dubey et al., 2023] Dubey, S., Tiwari, G., Singh, S., Goldberg, S., and Pinsky, E. (2023). Using machine learning for healthcare treatment planning. *Frontiers in Artificial Intelligence*, 6:1124182.

[Kaur et al., 2022] Kaur, P., Singh, A., and Chana, I. (2022). Bsense: A parallel bayesian hyperparameter optimized stacked ensemble model for breast cancer survival prediction. *Journal of Computational Science*, 60:101570.

[Kharya and Soni, 2016] Kharya, S. and Soni, S. (2016). Weighted naive bayes classifier: a predictive model for breast cancer detection. *International Journal of Computer Applications*, 133(9):32–37.

[Kourou et al., 2015] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17.

[Lee, 2023] Lee, M. (2023). Deep learning techniques with genomic data in cancer prognosis: A comprehensive review of the 2021–2023 literature. *Biology*, 12(7):893.

[Li et al., 2021] Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., and Peng, X. (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review. *PLOS ONE*, 16.

[Liu and Kurc, 2022] Liu, H. and Kurc, T. (2022). Deep learning for survival analysis in breast cancer with whole slide image data. *Bioinformatics*, 38(14):3629–3637.

[Lou et al., 2020] Lou, S.-J., Hou, M.-F., Chang, H.-T., Chiu, C.-C., Lee, H.-H., Yeh, S.-C. J., and Shi, H.-Y. (2020). Machine learning algorithms to predict recurrence within 10 years after breast cancer surgery: A prospective cohort study. *Cancers*, 12:3817.

[Rabiei et al., 2022] Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., and Atashi, A. (2022). Prediction of breast cancer using machine learning approaches. *Journal of Biomedical Physics and Engineering*, 12.

[Rasool et al., 2022] Rasool, A., Bunterngchit, C., Tiejian, L., Islam, M. R., Qu, Q., and Jiang, Q. (2022). Improved machine learning-based predictive models for breast cancer diagnosis. *International Journal of Environmental Research and Public Health*, 19:3211.

[Su et al., 2023] Su, F., Chao, J., Liu, P., Zhang, B., Zhang, N., Luo, Z., and Han, J. (2023). Prognostic models for breast cancer: based on logistics regression and hybrid bayesian network. *BMC Medical Informatics and Decision Making*, 23(1):120.

[Tate et al., 2020] Tate, A. E., McCabe, R. C., Larsson, H., Lundstro¨m, S., Lichtenstein, P., and Kuja-Halkola, R. (2020). Predicting mental health problems in adolescence using machine learning techniques. *PLOS ONE*, 15.

[Yang et al., 2020] Yang, L., Fu, B., Li, Y., Liu, Y., Huang, W., Feng, S., Xiao, Lin andSun, L., Deng, L., Zheng, X., Ye, F., and Bu, H. (2020). Prediction model of the response to neoadjuvant chemotherapy in breast cancers by a naive bayes algorithm. *Computer Methods and Programs in Biomedicine*, 192.