# Baysian Analysis of Hospital Website and Length of Stay

## Introduction

This report presents our findings for the website efficacy analysis and factors influencing length of hospital stay. In Section 1, we explored the data collected from the hospital’s website - the new and old version - and evaluated whether or not the new page produces a higher number of conversions. In Section 2, we explored the variables provided by the hospital’s board and conducted an analysis into variables influencing the length of stay, as well as the probability that a patient stays longer than 2 days given the specified variables. The requested variables were shared by the hospital in a CSV text file format labeled mimic3d.csv. Each section is complete with: an initial data exploration and explanation, a Bayesian statistical analysis, and an interpretation and summary of the results. Section 1 uses a Baysian A/B test and Section 2 uses a Bayesian Linear regression.

## Section 1

The first task we were assigned was to evaluate the marketing team’s new webpage that ideally increased the proportion of visitors who signed up or took some other action. Our initial step was to solidify each variable’s description through a data dictionary (see Table 1 below).

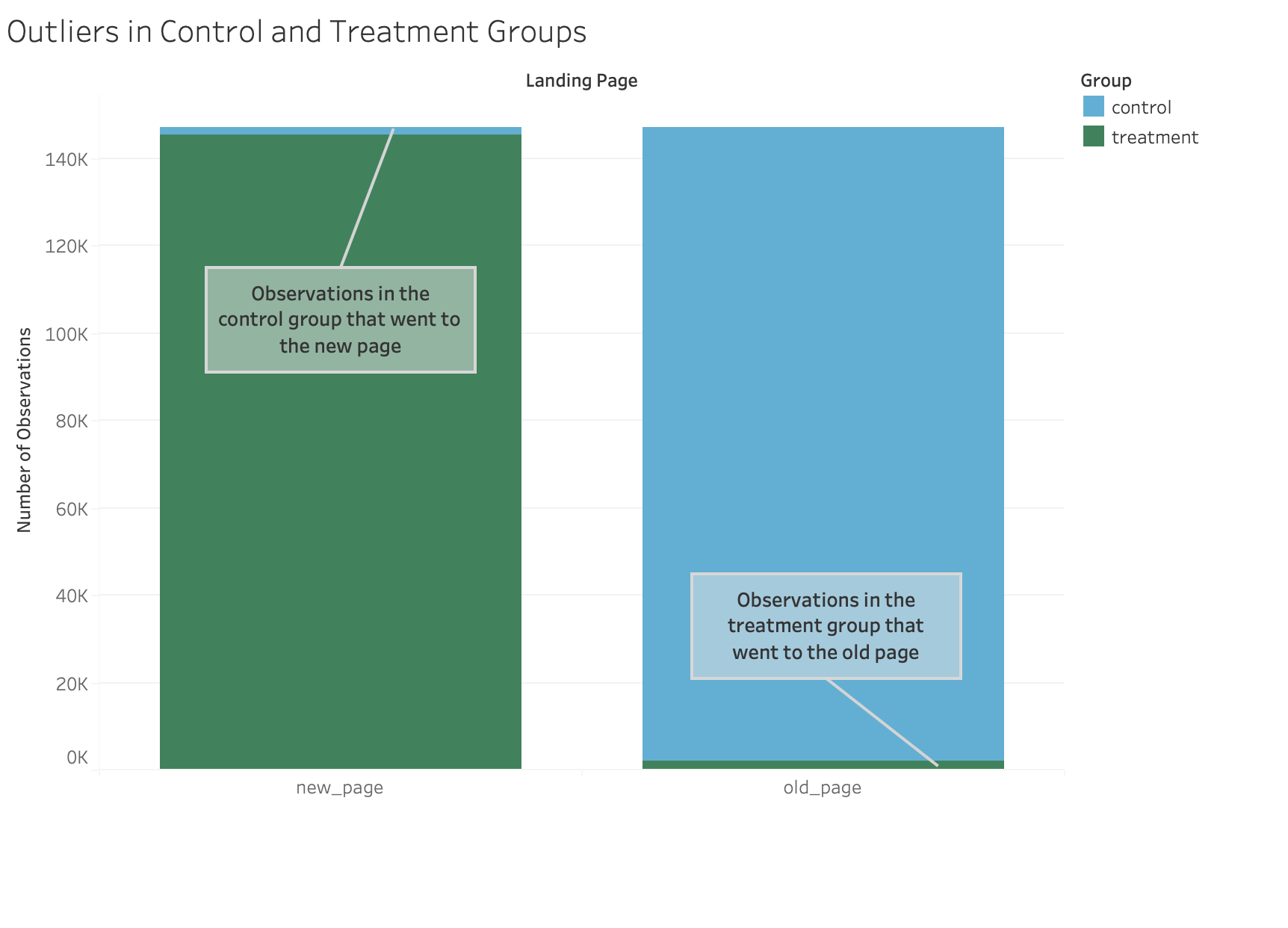
Table 1

| Column Name | Data Type | Has Nulls (Yes/No) | Description |
| --- | --- | --- | --- |
| user\_id | Integer (Numerical) | No | This represents the user’s id. |
| timestamp | DateTime | No | This represents the time and day that the user visited the web page. |
| group | Character (Categorical) | No | This defines whether the user was in the control or treatment group. |
| landing\_page | Character (Categorical) | No | This is another way of defining whether the user was in the control or treatment group, and it’s values are “old\_page” for the control group, and “new\_page” for the treatment group. |
| converted | Boolean (0/1) | No | This represents whether the user “converted” or not. 1 represents yes, and 0 represents no. |

We then began familiarizing ourselves with the data by generating summary statistics and performing exploratory data analysis in R. The dataset contained 5 features, namely, user\_id, timestamp, group, landing\_page, and converted, as outlined in Table 1. Not only did the dataset contain a large number of observations (294,478), there was also a fairly even number of observations in the control and treatment groups, which is important for the Bayesian A/B testing in the following section. From the analysis, we also found that the conversion rate for the control group was 12.04%, while the conversion rate for the treatment group was 11.89%, indicating the old page had slightly more conversions than the new page

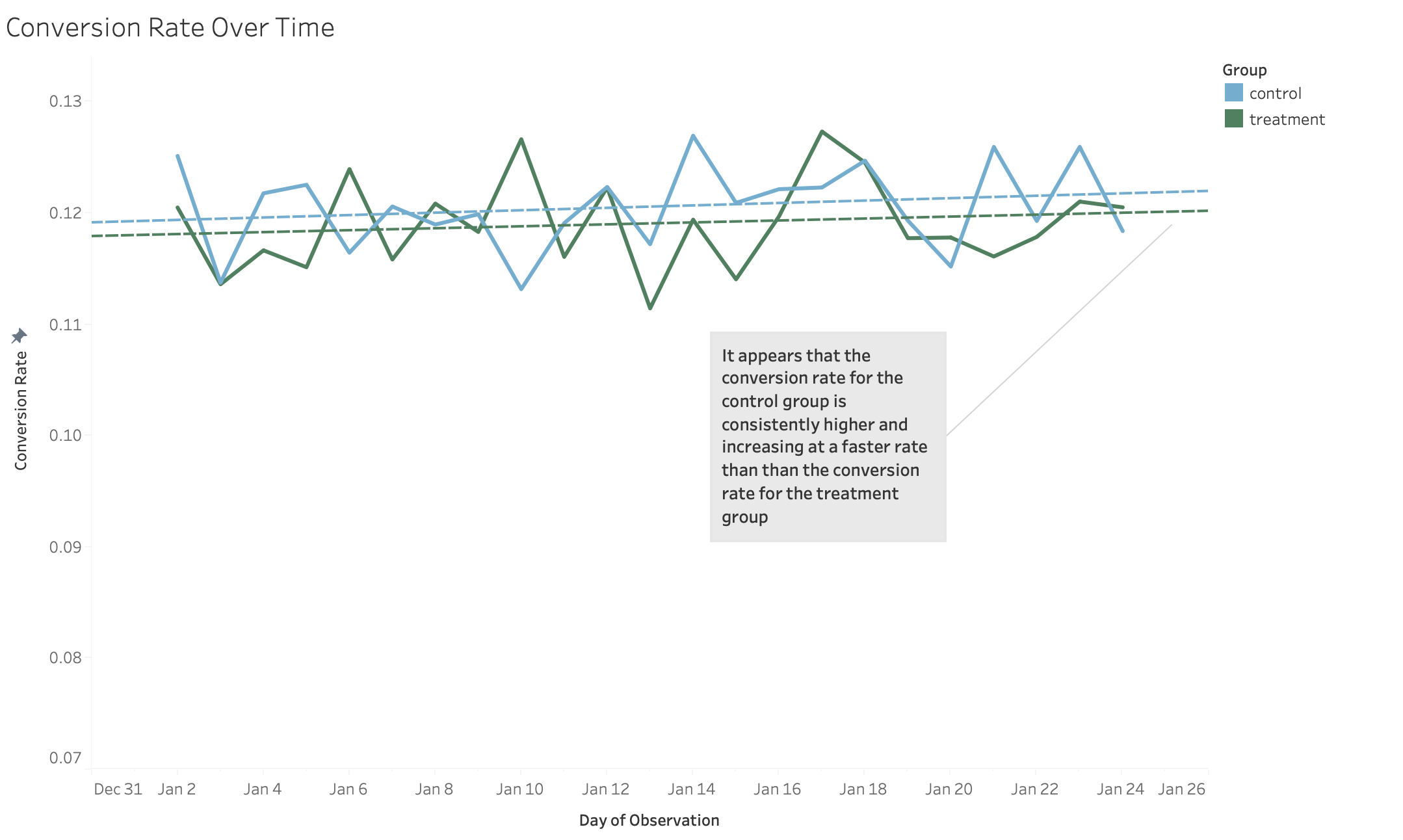
This analysis revealed that there seemed to be outliers in the treatment and control group, where 1,928 observations marked as “control” had their landing\_page variable equal to “new\_page”, and 1,965 observations labeled “treatment” had their landing\_page variable equal to “old\_page”. To clean the data, the observations in the control group that went to the new page and those in the treatment group that went to the old page were removed, leaving 145,275 and 145,311 observations in each respective group, remaining fairly evenly split. A final pass through of the data revealed one instance in which the same user\_id was assigned to two separate treatment observations. To keep the data as clean as possible, we removed both of these observations for a total of 290,584 observations - 145,275 in the control group and 145,309 in the treatment group.

We then explored the dataset visually in Tableau. Figure 1 showcases the small number of observations whose landing page is not matching the group it should be in. Figure 2 displays the conversion rate by group by day. As we can see, the grouped conversion rates are fairly consistent over time, and the moving average dotted line indicates both conversion rates growing as the experiment goes on, though the conversion rate for the control group appeared to grow faster.



Outliers in Control and Treatment Groups

Figure 1

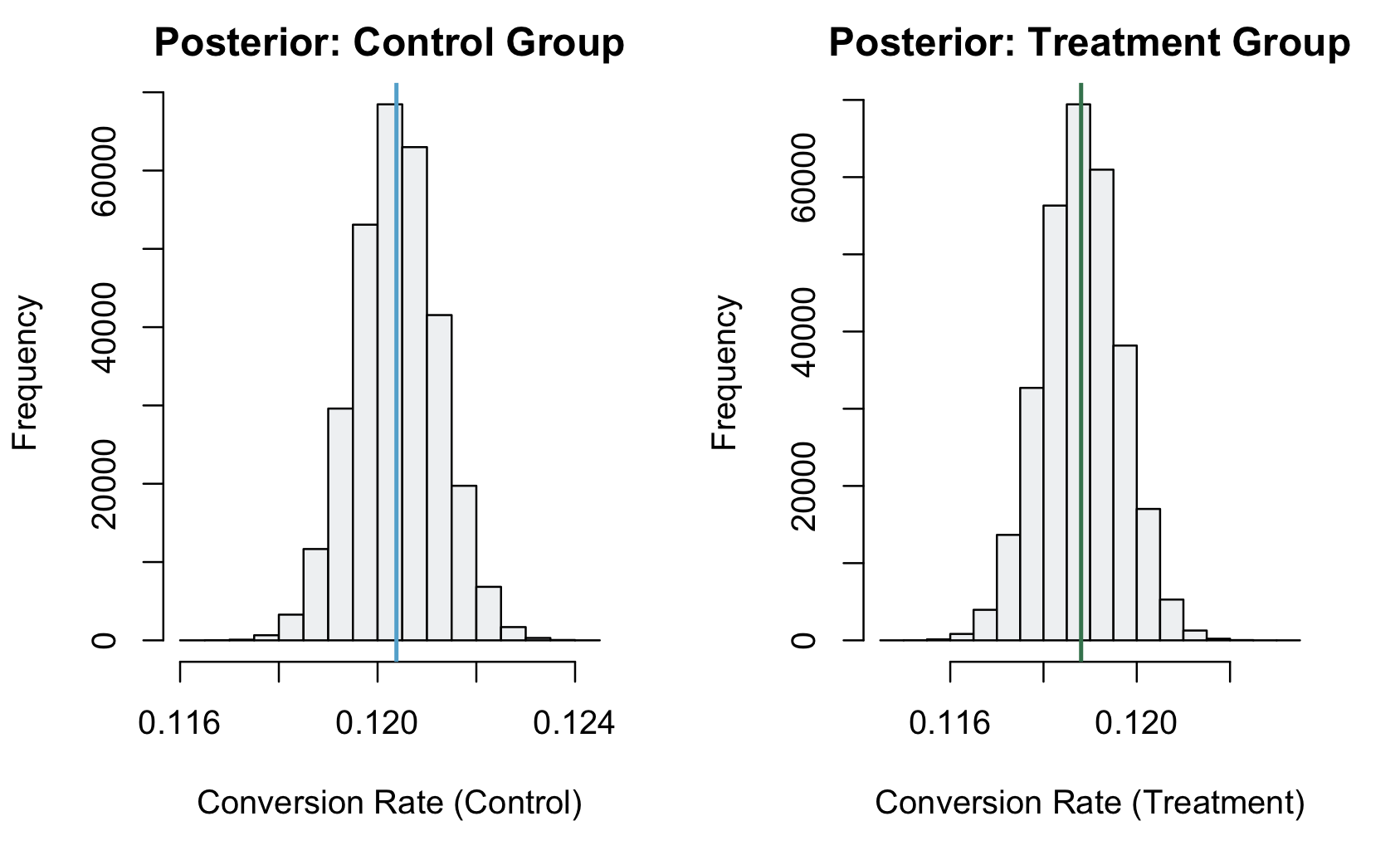


Conversion Rate Over Time

Figure 2

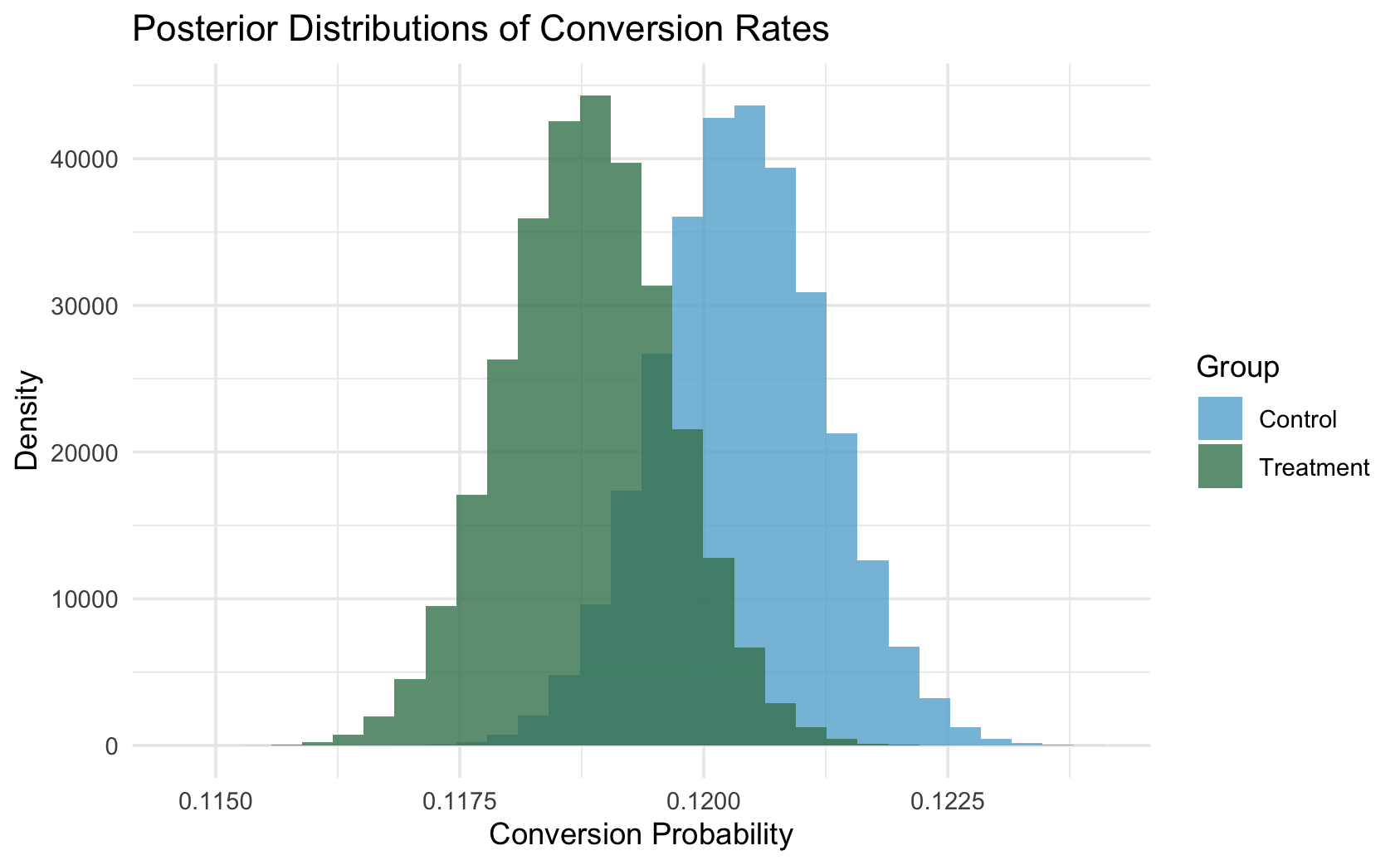
### Bayesian A/B Test

We decided to perform a Bayesian A/B test to determine how likely it is that one website page had a better conversion rate over the other. The first step was to clean the data, which we completed in our exploration. We then implemented the test with a prior of Beta(=2, =20) using code in R. We used an estimated likelihood approach for the posterior, as opposed to an exact match, which filters data with strict parameters. By utilizing a likelihood instead of exact match, we achieved more reliable results by accounting for real-world variability and error. The final results of performing the Bayesian A/B test with a prior of Beta(=2, =20) did not show that there was a large difference in conversion rate between the old page and new page. The mean conversion rate for the control group and the treatment group were 12.04%, and 11.88% respectively - a difference of only 0.16%. The credible interval for the control group was [0.1187112, 0.1220609], and the credible interval for the treatment group was [0.1171455, 0.1204646]. The produced visuals of this test are displayed in Figures 3, 4 and 5. Figure 3 displays the posterior distributions of both the control and treatment groups, and we noted the mean of the posterior for the control group is slightly higher than that of its treatment group counterpart. Figure 4 displays both posterior distributions overlapping, indicating there may be no statistically significant difference between the new web page and old web page as the conversion rates are less than 5%, the selected tolerance threshold. Figure 5 displays the posterior of control minus the posterior of treatment, and we see that it appears normally distributed around 0.0175, thus there is no major indication that one website performs better than the other.



Posterior Graphs

Figure 3



Posterior Graph

Figure 4



Posterior Graphs

Figure 5

We also explored the use of a prior other than Beta(α=2, β=20) as it assumes a low probability of success. With the hopes of comparing results, we chose an uninformative prior Beta(α=1, β=1), which minimally influences the inference. The results yielded by the uninformative prior and conjugate prior were very similar, suggesting the choice of prior is not heavily affecting the posterior conclusions. Using the uninformative prior, the credible interval for the control group is [0.1187207, 0.1220708] and the credible interval for the treatment group is [0.1171548, 0.1204743]. The mean conversion rate for the control group and the treatment group are 12.04% and 11.88%, respectively. Clearly, the uninformative prior compared to the chosen prior of Beta(α=2, β=20), does not yield significantly different results, supporting our inclination that the prior ultimately had little to no effect on the posterior probability that one website has more conversions than the other. Lastly, we wanted to further explore the choice in priors and demonstrate why they appeared to have no effect on our results. Our goal was to determine how many observations it took for the posterior results to be unaffected by the priors. We began by taking 50 (as an estimated maximum) random samples of size 100 from our data frame. For each sample, we stored the conversion numbers for the treatment and control group in a separate data frame. To investigate this idea, we generated a loop code for a Bayesian A/B test that would take one sample of 100 and, first using the prior Beta(α=2, β=20), perform the same test as before to generate the posterior results. The loop would then update the priors with the addition of 100 more observations from the next random sample, and recalculate the posteriors. This would continue for a maximum of 25 iterations, a number we determine by trial and error, deeming that we reach the threshold when the prior becomes irrelevant far earlier than 2,500 observations.

Figure 6 visualizes the posterior distribution of the control group as we add more data to find the posteriors. The title of each subplot is the number of observations currently being used to update the posteriors. This plot is used to visualize the prior’s effect on the posterior, and to gain an understanding of at what point the prior becomes irrelevant. At first glance at Figure 6, we estimate the prior becomes irrelevant at around 1800 observations as both the distribution and mean remain little changed as observations continue to increase. To confirm this, after each iteration we calculated the posterior probability that the new web page has a higher conversion rate than the old page, using a random sample of 100 from the posterior. This posterior probability begins high at around 0.76, and remains both volatile and high, until around 1,500 observations, when it begins to decrease, and eventually ending at 0.07 at 2,500 observations, indicating that the priors become irrelevant at around 1,500 observations.

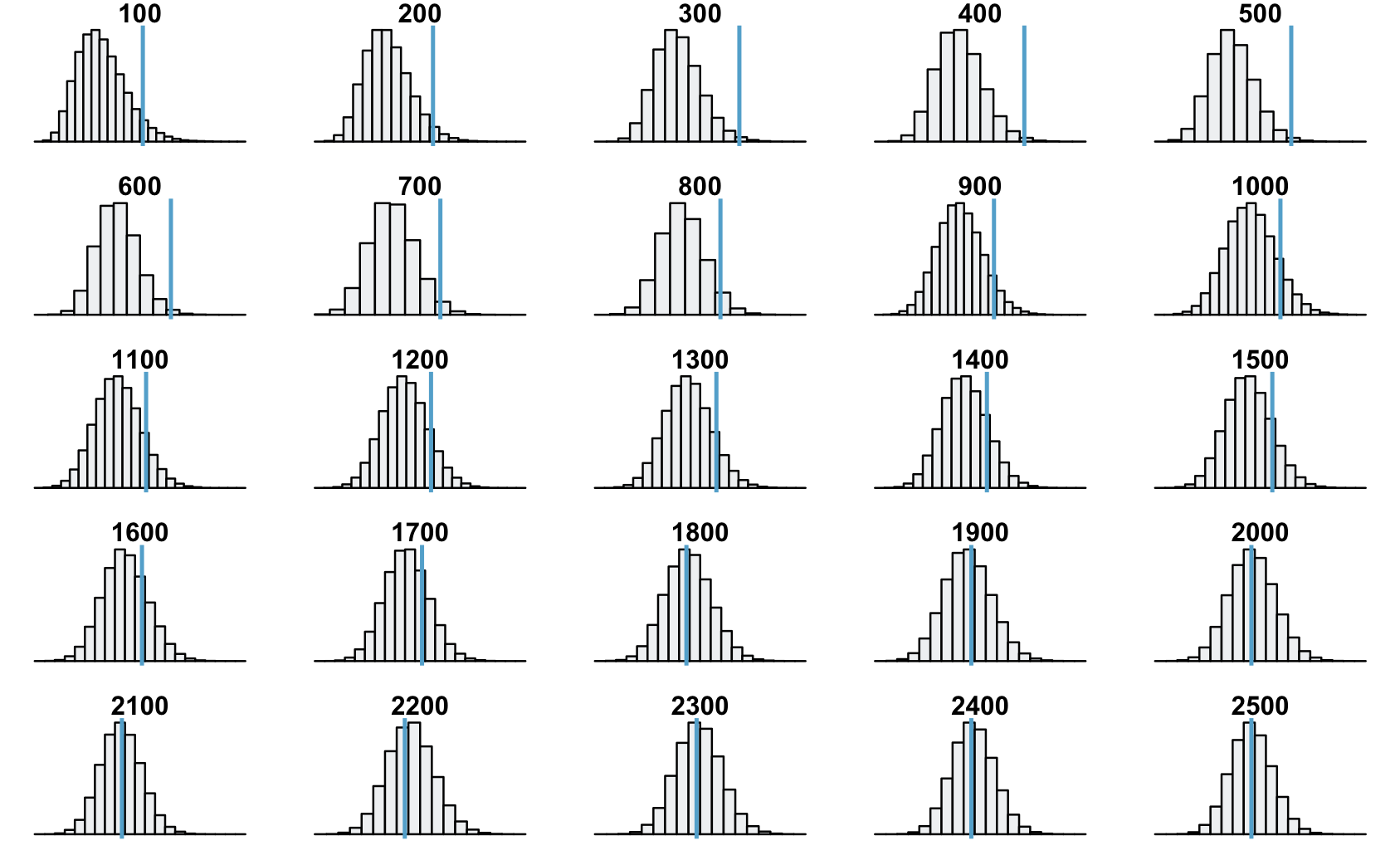


Figure 6

Figure 6

The results produced through the Bayesian A/B test substantiated our claim that while not a significant difference, the old website page is more likely to have a conversion than the new website page. We conclude that more observations may be required before finalizing the decision of a web page, though if a decision is necessary, either page is likely to produce similar conversion rates.

## Section 2

The next area of analysis is in regards to data collected from the hospital. First we will be defining the data, and explaining what some of the values mean. In the second portion we wil be analyzing. We have divided up the dataset into three distinct sections; Precursor Data, Geographical Data, Variable Data. These are divisions made by the type of the data provided (Category vs Numeric) as well as the type of actionable markers we can recommend.

### Data Dictionary

#### Precursor Data

Each patients arrives to the hospital in a given state. Each patient is assigned a gender, age, marital status, etc. These are data points which cannot be controlled by the hospital. Below is the data dictionary for this section.

| Name | Variable Type | Description |
| --- | --- | --- |
| gender | char | Gender expressed as M or F |
| age | integer | Age of patient expressed at integer |
| insurance | string | Self paid, Private, Medicaid or Medicare |
| religion | string | Self identified religion of patient |
| marital\_status | string | Self identified stats (Divorced, Single, Married, etc) |
| ethnicity | string | Reorganized to Categories White, Black, Asian, Hispanic, Other and Unavailable |

Note: Originally ethnicity was expressed as subcategories ( eg. ASIAN - Chinese ). Little insight could be drawn from such a small subdivision so it was decided to categorize them into broader categories. Unavailable has remained the same.

#### Positional Data

Each hospital is located in different parts of the United States, and will be pushed by different factors due to that location. Further more each patient will be inducted into different parts of the hospital. A patient entering emerge will have a different experience to those entering newborn.

| Name | Variable Type | Description |
| --- | --- | --- |
| hadm\_id | integer | Identifier for particular visit to hospital |
| admit\_type | string | Categorized admission as - Emergency, Elective, Newborn Urgent |
| admit\_location | string | Physical location the patient was admitted to (Ex. Transfer from HOSP/Extram) |

With this section of data we will look into patients records across time, and throughout different parts of the hospital.

#### Variable Data

This is data that in some way could change via workflow changes within the hospital, and as mistakes are made by staff. The number of labs for instance. It is possible that every hospital currently operates with the utmost efficiency with no wasted lab results. This is incredibly unlikely however, especially when excessive lab assessment’s lengthen the stay of the patient. What will be shown later in this report is a statistical approach to which variables may be of higher importance to the length of each patients stay.

| Name | Variable Type | Description |
| --- | --- | --- |
| AdmitDiagnosis | string | Diagnosis upon admission |
| NumCallouts | integer | Number of clinicians *called out* of their working hours |
| NumDiagnosis | float | Aggregate Number of Procedures |
| NumProcs | float | Aggregate Number of Procedures |
| AdmitProcedure | string | Procedure performed upon admission |
| NumCPTevents | float | CPT Code labeled events |
| NumInput | float | Aggregate Number of Inputs ( eg Medication Administered ) |
| NumLabs | float | Aggregate Number of Labs |
| NumMicroLabs | float | Aggregate Number of Procedures Micro Labs |
| NumNotes | float | Aggregate Number of Clinical Notes |
| NumOutput | float | Aggregate Number of Procedures Outputs |
| NumRx | float | Number of prescribed medications |
| NumProcEvents | float | Aggregate Number of Procedures |
| NumTransfers | float | Aggregate Number of Transfers |
| NumChartEvents | float | Aggregate Number of Procedures |
| ExpiredHospital | boolean | Whether the patient died in hospital |
| TotalNumInteract | float | Aggregate Number of Procedures |

### Analytics

#### Precursor Data

Due to the fact that this data is prior to the patients arrival at the hospital, there is little that can be said for actionable results. Instead we will seek to understand more about the patient base.

**Gender**

The key not is the greater proportion of men making up 55% of the patient base.

| Men | Women |
| --- | --- |
| 32950 | 26026 |

This proportion is interesting as it is 5% offset from the national average.

**Age**

For ages strictly greater than 0 we have a skewed distribution shown in figure ## with mean equal to 61 and standard deviation equal to 16; comparing this to the average age in the United States being 38. This is not unexpected as people of more age are more likely to posses ailments requiring hospital visitation.

Once we incorporate the data for ages less than 1 ( eg newborn / infant ), we can see age no longer follows a normal distribution. Over 8000 newborns or infants were admitted though the duration of the study.

**Insurance**

The most common form of insurance at 48% is Medicare, followed closely by Private insurance at 38%. Below is a pie-chart breaking down the insurance categories.

**Religion**

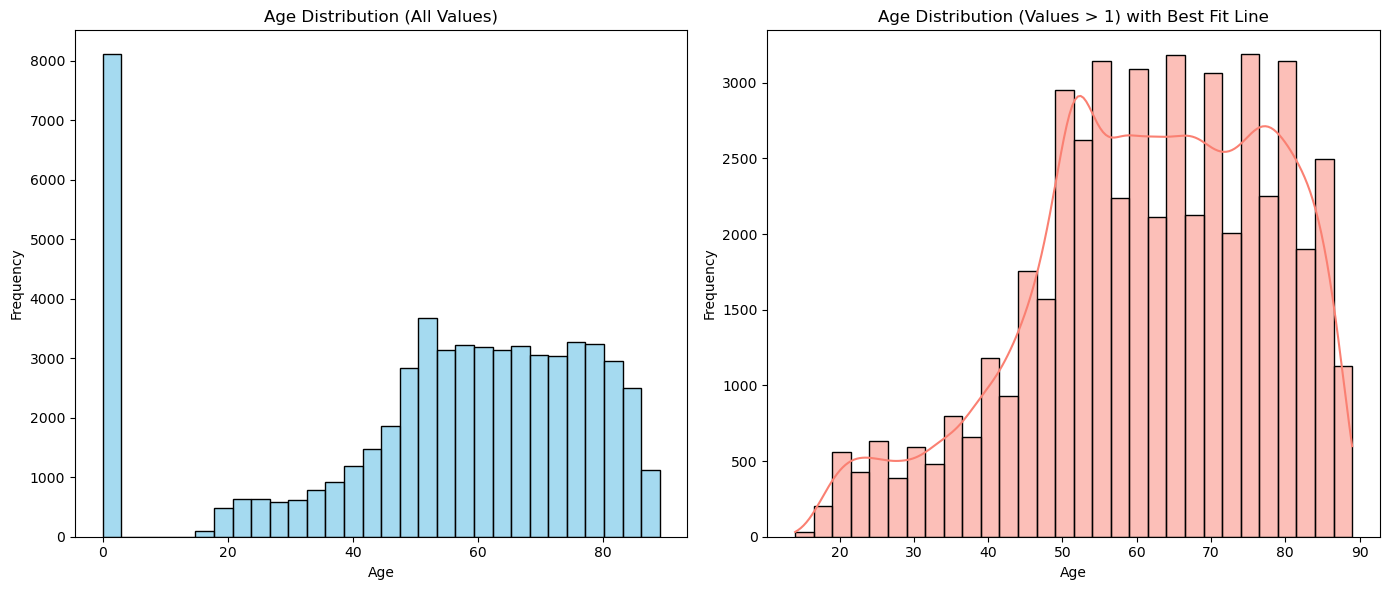
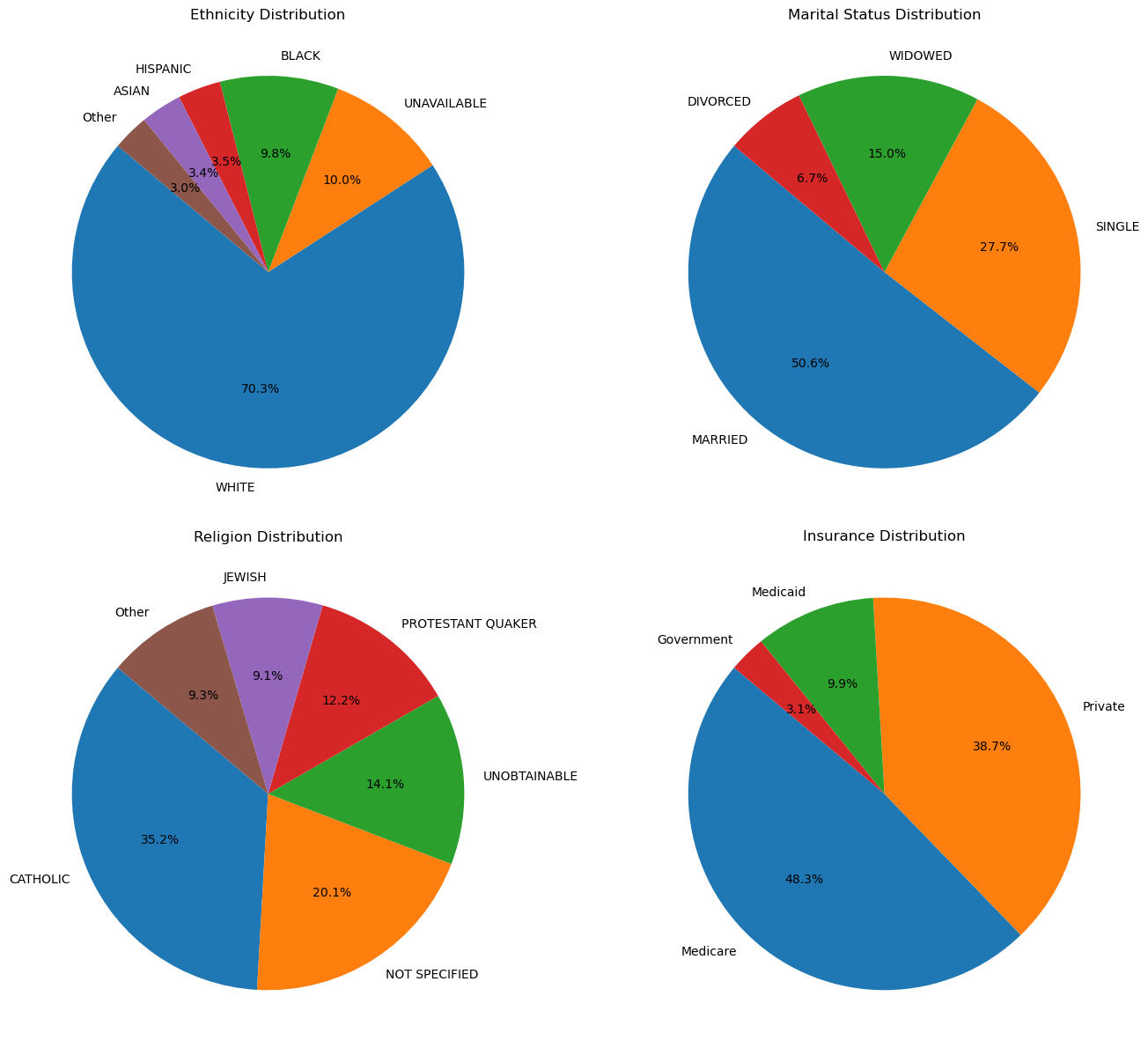
Along the trend of trying to determine patient data, we compare the distribution of religious secs compared to the national average. We see that compared to the national average, 3% less patients are Catholic. This falls below statistical significant range. We can say then that the religious distribution of patient data follows that of national averages.

**Marital Status**

As shown in the pie-charts below, 50% of patients are married. This is similar to 53% national average of married individuals.

**Ethnicity**

As mentioned above, the ethnicity was simplified to 6 categories. According to the pie-chart below, 70% of patients identify as white. This is 10% higher than the national average. This idea is supported by the difference between patient ethnicity percentage and the national average.

#### Positional Data

**Admit Type**

Admission types will follow one of four choices, Emergency, Newborn, Elective, and Urgent. The pie-chart can be found below.

The vast majority of patients enter as emergency (71%).

**Admit Location**

Within the Hospital there are 8 known locations a patient can be admitted to. It follow from the previous section that the majority of patients will be entering via the Emergency Room, this corroborates that theory, in total 38% of patients.

#### Variable Data

It is important to note that any correlational data will be done in the next section as a precursor to the Bayesian Linear Regression Analysis.

Until then we will look into the frequencies of some key data points withing the Variable Data.

**AdmitDiagnosis**

Though the duration of the data over 15 thousand unique diagnosis’s were found, nearly half of which were Newborn. Due to the nuances many diagnosis’s entail, we will analysis the diagnosis further by filtering by keywords. If a keyword is detected, the entry will be flagged and categorized accordingly.

In total this system flagged:

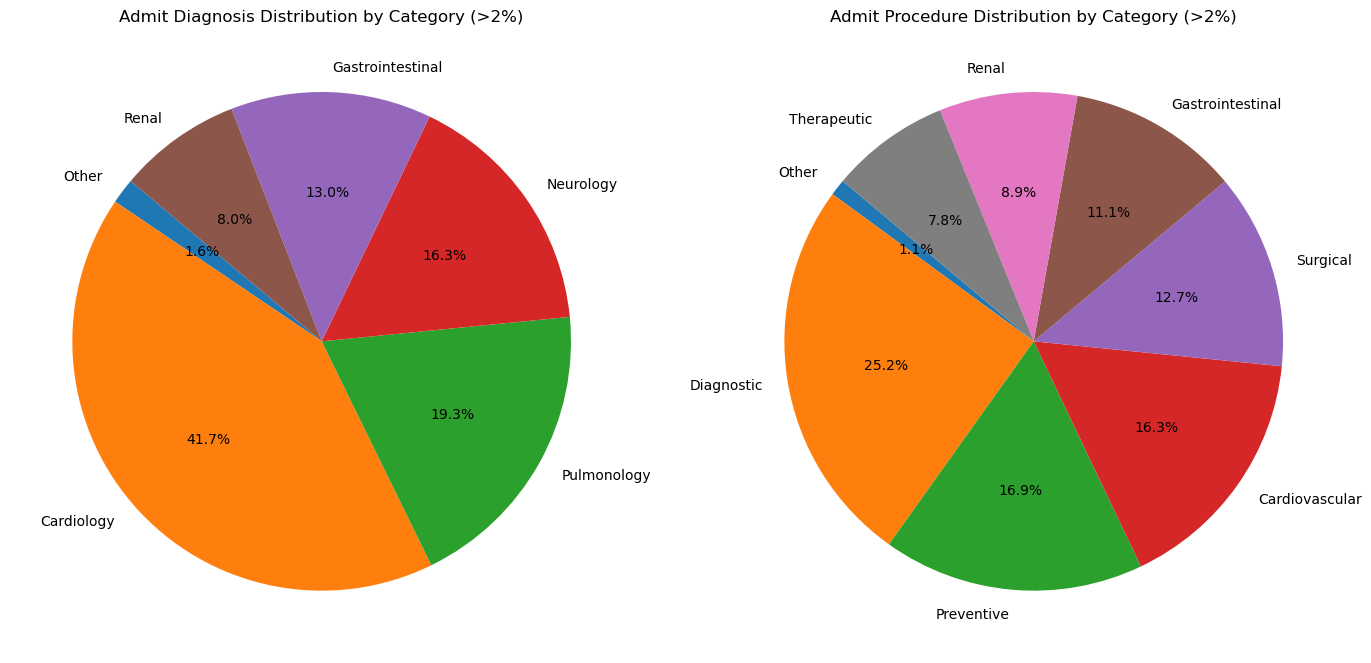
| Category | NumFlagged\_Items | Keywords |
| --- | --- | --- |
| Other | 43133 | N/A |
| Diagnostic | 3991 | scan, x-ray, ultrasound, mri, ct, imaging, biopsy, diagnostic |
| Preventive | 2683 | vaccination, screening, check-up, preventive, immunization |
| Cardiovascular | 2587 | angioplasty, stent, bypass, pacemaker, cardiac, heart |
| Surgical | 2008 | surgery, operation, resection, excision, laparotomy, amputation, biopsy |
| Gastrointestinal | 1760 | endoscopy, colonoscopy, gastrectomy, biopsy, bowel, liver |
| Renal | 1417 | dialysis, nephrectomy, catheter, kidney, bladder, renal |
| Therapeutic | 1229 | therapy, treatment, chemotherapy, radiation, rehabilitation, transfusion |
| Respiratory | 154 | ventilation, tracheotomy, bronchoscopy, thoracotomy, oxygen, pulmonary |
| Musculoskeletal | 14 | arthroscopy, joint, spine, orthopedic, tendon, ligament |

**AdmitProcedure**

We use a similar filter as above to analyze the procedures recorded.

| Category | NumFlagged\_Items | Keywords |
| --- | --- | --- |
| Other | 43133 | N/A |
| Diagnostic | 3991 | scan, x-ray, ultrasound, mri, ct, imaging, biopsy, diagnostic |
| Preventive | 2683 | vaccination, screening, check-up, preventive, immunization |
| Cardiovascular | 2587 | angioplasty, stent, bypass, pacemaker, cardiac, heart |
| Surgical | 2008 | surgery, operation, resection, excision, laparotomy, amputation, biopsy |
| Gastrointestinal | 1760 | endoscopy, colonoscopy, gastrectomy, biopsy, bowel, liver |
| Renal | 1417 | dialysis, nephrectomy, catheter, kidney, bladder, renal |
| Therapeutic | 1229 | therapy, treatment, chemotherapy, radiation, rehabilitation, transfusion |
| Respiratory | 154 | ventilation, tracheotomy, bronchoscopy, thoracotomy, oxygen, pulmonary |
| Musculoskeletal | 14 | arthroscopy, joint, spine, orthopedic, tendon, ligament |

Although the majority of procedures could not be flagged, this tables still gives us an understanding how many of each type of procedure are completed and should be used to find proportions of categories, as opposed to the raw size.

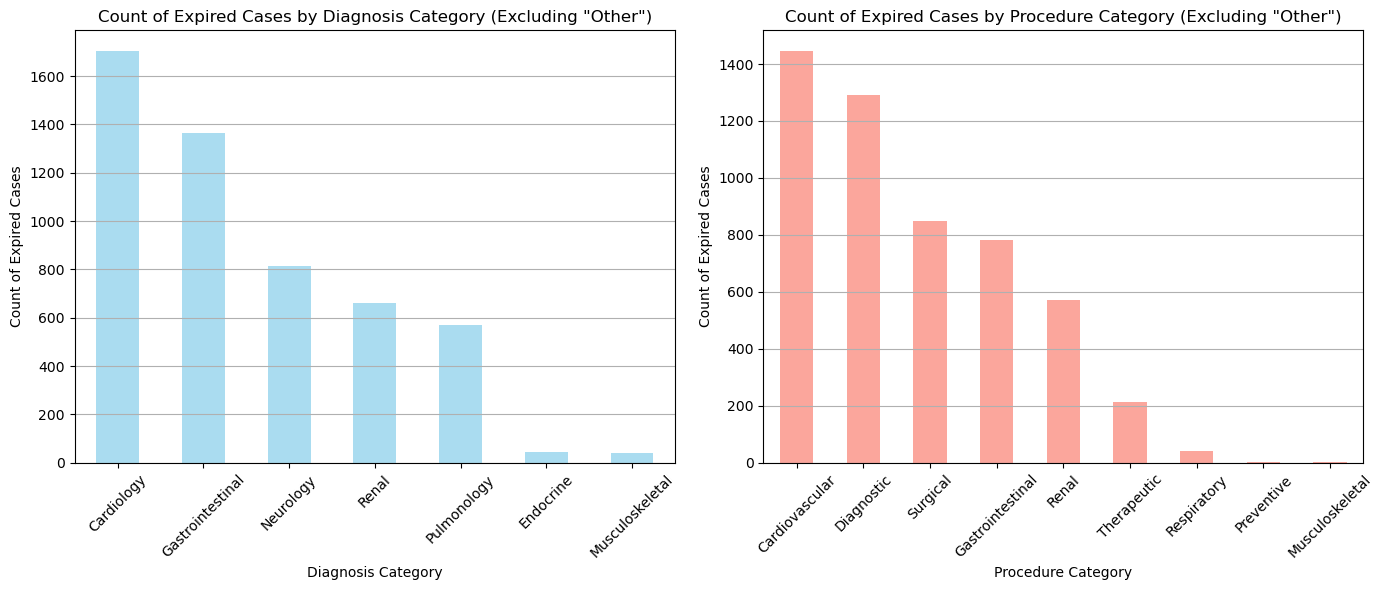
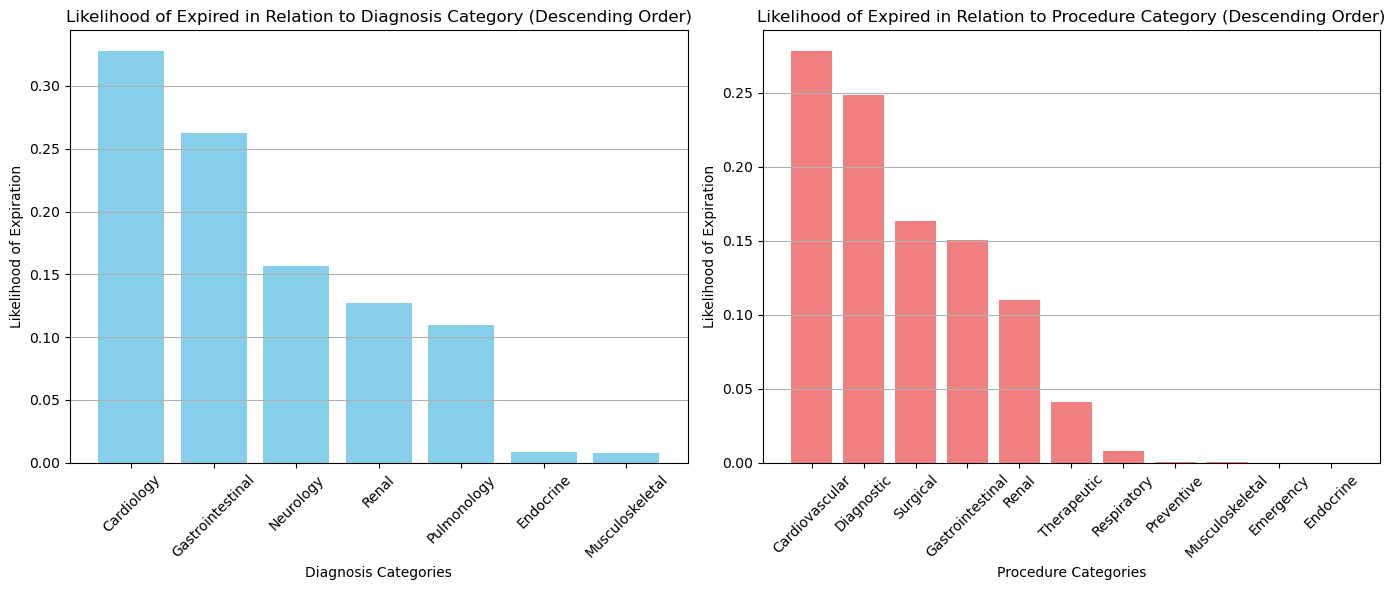


Admit & Procedure Piechart

**Expired in Hospital**

Thanks in part to the categories we defined above, we can go through and analyses which groups of procedures and diagnoses have the highest mortality rate associated with them. It is important to note that this does not imply correlation. A particular ailment may have a high lethality despite the procedures complete not because of them. Below are two bar plots showing the number of occurrences than a particular diagnosis or procedure lead to expiry. However more importantly, as we establish above, the categories are designed to be a guideline for probabilities. As such the second bar graph shows the likelihood that given a particular diagnosis or procedure that you would expire within the hospital.

#### Figures

### Bayesian Analysis

Next, we performed Bayesian linear regression on the patient data provided. The original data had 28 variables. Note that, at this point, we were using the entire patient dataset and had not filtered for the patients in patients.csv. We took LOSdays as our target, and all other variables except hadm\_id and LOSgroupNum as predictors, but left hadm\_id in the dataframe for now to facilitate splitting the dataset later. Religion and ethnicity were dropped because we did not believe them to be significant. ExpiredHospital was also dropped because predicting how long a patient will stay at the hospital after their death seemed irrelevant. To reduce the risk of multicollinearity and high variance inflation factor when computing pairwise interactions, numeric variables were centered. High-cardinality categorical variables were encoded using binary encoding, and low-cardinality categorical variables were one-hot encoded. After encoding, we had 56 predictors.

Next, we computed pairwise interactions. The addition of pairwise interactions increased the number of predictors from 56 to 1596. The next task was feature selection, for which we used mltxend’s SequentialFeatureSelector. We initially tried a number of other feature selection techniques, including assessing mutual information scores, random forest models, assessing correlation coefficients, inspecting covariance matrix visualizations, computing VIF, and stepwise selection on subsets of variables. This added an extra layer of complexity to the task by necessitating decision making and a lot of trial and error. Using a sequential feature selector allowed us to outsource some of the decision-making without compromising the quality of our feature selection. We ran deterministic linear regression to compare the quality of the features selected by different methods, and the sequential feature selector yielded the highest R2 score. It is worth noting that the R2 was still not particularly high (0.1631), but we attribute this to the constraint on the number of features and the general trend of lower R2 values in social science research compared to other domains. The five features selected were ‘NumTransfers’, ‘NumCallouts admit\_type\_EMERGENCY’, ‘NumDiagnosis NumTransfers’, ‘NumChartEvents admit\_type\_NEWBORN’, ‘admit\_type\_NEWBORN admit\_location\_CLINIC, and REFERRAL/PREMATURE’.

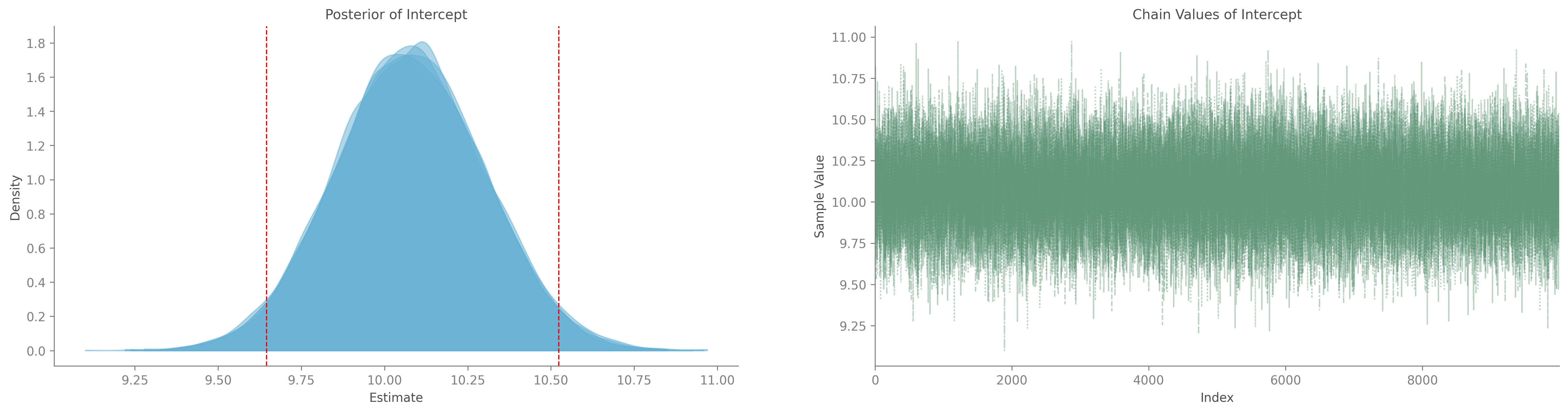
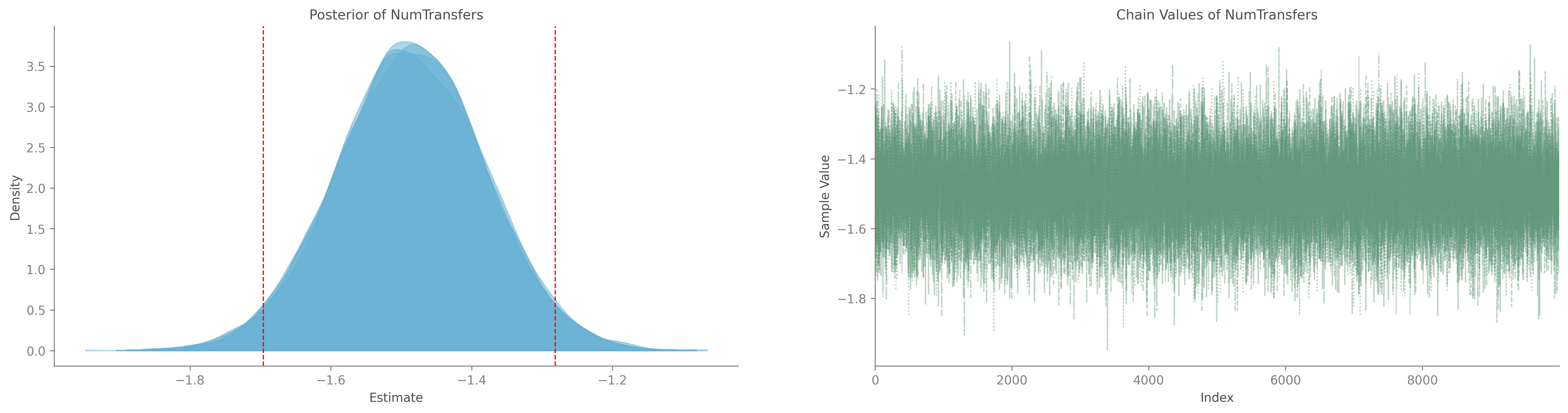
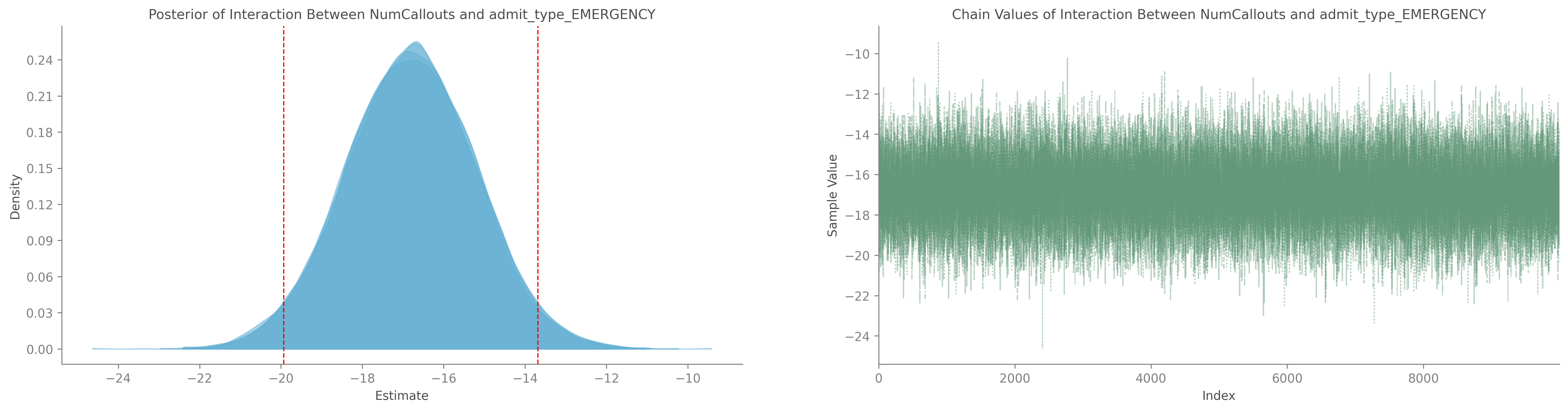
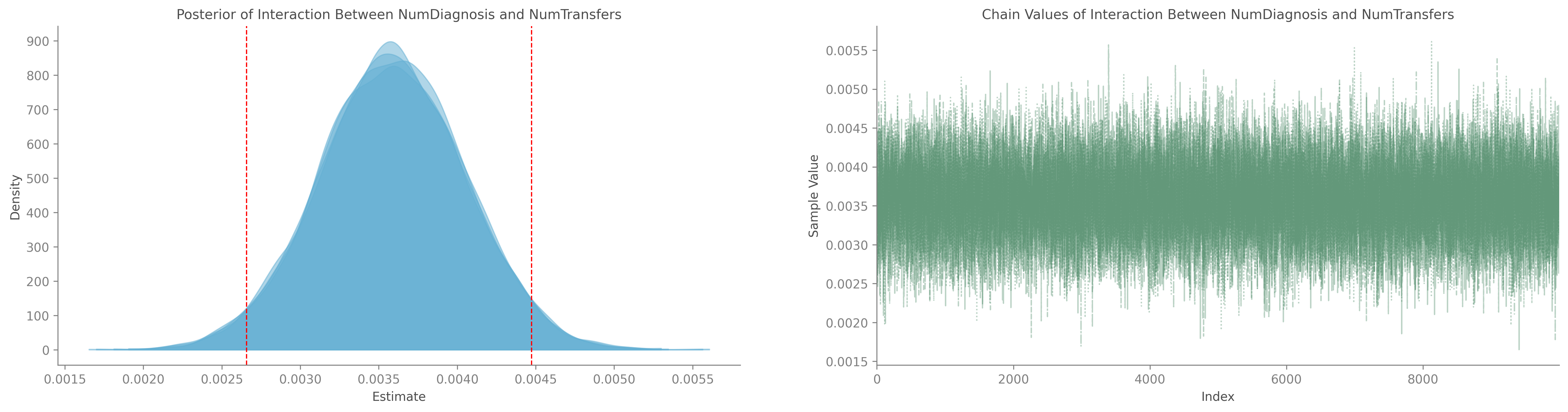
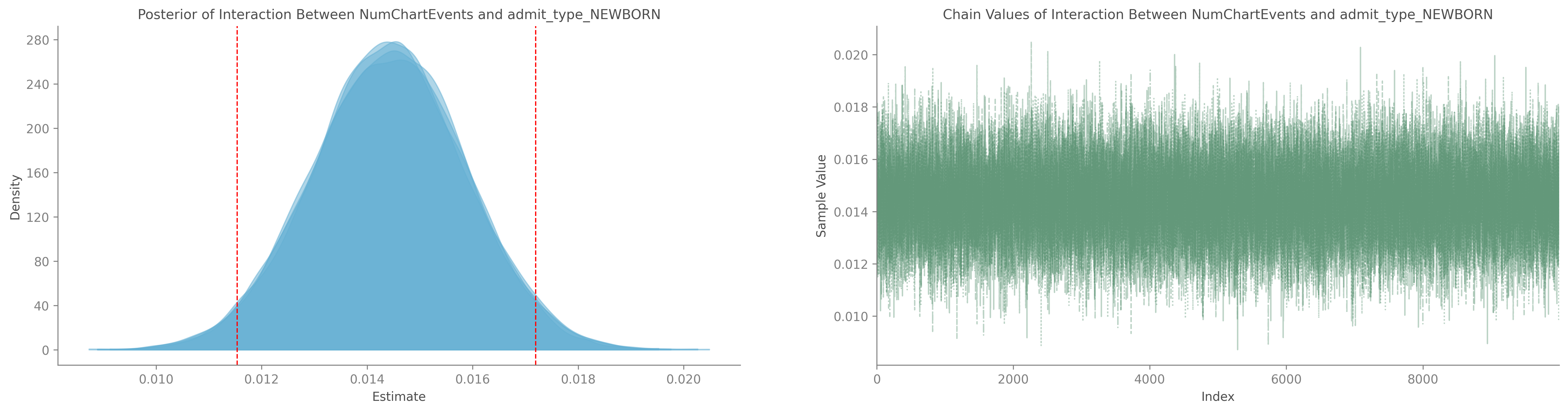
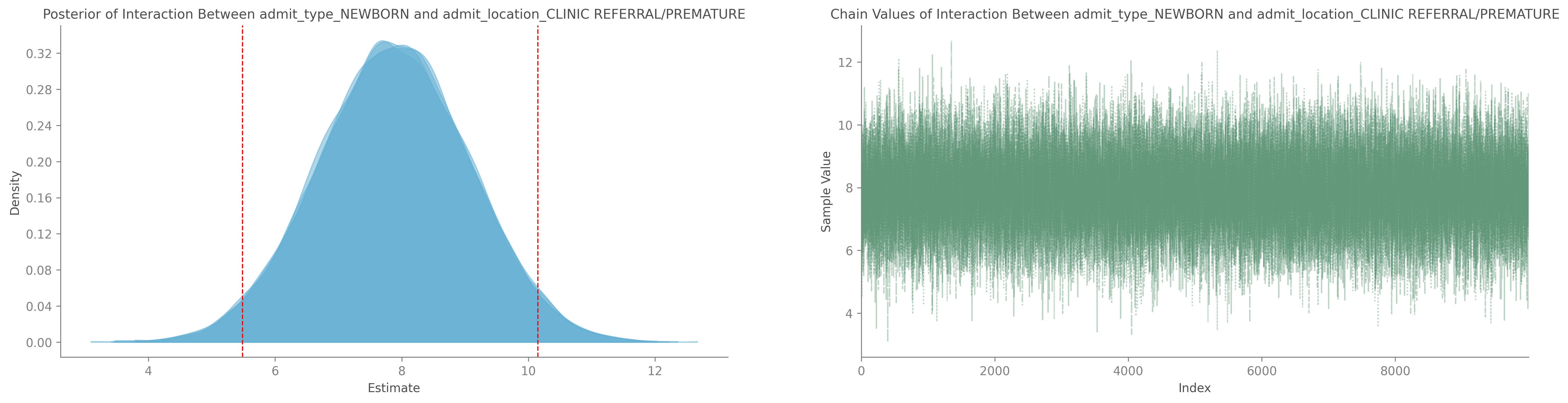
Our dataset was then split into two subsets: the data on which we performed Bayesian inference (patients in patients.csv) and the rest of the data. We performed deterministic linear regression on the latter, using our five selected features as predictors and LOSdays as the target. The purpose of this was to use the intercept and coefficients to inform the means of our normal priors when we perform Bayesian linear regression. This approach falls under the larger umbrella of empirical Bayesian methods (Casella 1992). As stated previously, it is important to note that here we did not use the data on which we then performed Bayesian inference. Below is a summary of the results.

#### Table 1: Coefficients for Empirical Bayesian Analysis

| Coefficients | Estimate | Pr(>|t|) |
| --- | --- | --- |
| (Intercept) | 1.016e+01 | < 2e-16 |
| x1 | -1.469e+00 | < 2e-16 |
| x2 | -1.424e+01 | < 2e-16 |
| x3 | 6.560e-03 | < 2e-16 |
| x4 | 1.465e-02 | < 2e-16 |
| x5 | 1.421e+01 | < 2e-16 |

**Residual standard error:** 11.4 on 55994 degrees of freedom  
**Multiple R-squared:** 0.1632, **Adjusted R-squared:** 0.1631

We performed Bayesian linear regression using Python’s pymc library. For the intercept and coefficients, we used normal priors with the above coefficient estimations for the means. To reflect uncertainty about our previous computations, we set variances to 100 to make the priors less informative. An exponential prior was selected for the standard deviation. The likelihood is defined as a normal distribution with mean computed by our regression formula. Below are the plots of our posterior distributions and sampling chains. The red lines indicate the high density intervals.

For the following, we use “coefficient” to represent the overall posterior distribution. The negative coefficient for number of transfers suggests that more transfers was correlated with a shorter overall stay. This seemed surprising at first, but perhaps patients who were frequently transferred were less likely to stay in one place for a long period of time. The interaction between NumCallouts and admit\_type\_EMERGENCY had a large negative coefficient relative to the other variables. This indicates that, when a patient entered the hospital via emergency admission, more callouts was correlated with a shorter hospital stay. We hypothesize that this could be for two reasons. Firstly, that more callouts indicated that a patient received more immediate care, or that patients with more callouts were more likely to die shortly after reaching the hospital. The slightly positive coefficient for the interaction between NumDiagnosis and NumTransfers was interesting, as transfers alone had a negative coefficient. Perhaps transfers due to new diagnoses had an inverse effect compared to transfers alone. The small positive coefficient for the interaction between NumChartEvents and admit\_type\_NEWBORN suggested that, for newborn admissions, more chart events were correlated with a slightly longer stay. This makes sense, as more chart events could be indicative of premature births or other complications.. The interaction between admit\_type\_NEWBORN and admit\_location\_CLINIC REFERRAL/PREMATURE also had a higher positive coefficient, supporting our hypothesis that premature births led to longer hospital stays.

By sampling from the posterior distributions, we computed the probability of a patient staying in the hospital for more than two days to be 73.7%.

## Conclusion

The key findings of both analyses completed in this report are: neither the old website nor the new website provided better conversion results, therefore either may be used with very similar results, and, using a Bayesian linear regression model, we estimated that the probability of a patient staying in the hospital for more than 2 days, and therefore very likely missing work, is 73.7%.