Bilge K., David L., Iryna T., Liam D.

## Context

The following is split into two parts which should be treated as two distinct deliverables. We have left them together for the convenience of the TA(s) and Professor Boily. Part I is a response to a request made by the office of the Chief Data and Analytics Officer (CDAO) to analyze data provided by the marketing department. The CDAO will in turn make a recommendation to marketing department. Part II is a response to a request made by an insurance underwriting division to evaluate hospital length of stay using patient data, in order to support data-driven adjustments to insurance rates and reimbursement strategies.

## Part I – ab_data1.csv

### Data Exploration

The following is an exploration of the dataset provided by the marketing team to determine whether or not a new web page leads to an increased conversion rate (i.e. proportion of site visitors that take further action on the site). The dataset captures information about user site visits, such as assigning a unique identification number per site visitor, the time the site was visited, whether the site visitor is in the control or treatment group or landed on the new or old page, and whether a conversion occurred. Table 1 outlines the variables, their type, range, description, and examples. No datapoints were removed for the analysis.

| Variable | Type | Range | Description | Num NAs | Examples |
|---|---|---|---|---|---|
| user_id | Integer | [630000: 945999] | identification per unique site visitor (are duplicates in the dataset) | 0 | 630000, etc. |
| timestamp | Date Time | 2017-01-02 13:42:05.38 – 2017-01-24 13:41:54.46 | the time at which the site was visited | 0 | 2017-01-02 13:42:05.38, etc. |
| group | Binary Categorical | NA | whether the site visitor is in the control group or the treatment group | 0 | control, treatment |
| landing_page | Binary Categorical | NA | whether the site visitor is on the new or old page | 0 | old_page, new_page |
| converted | Integer | [0 ; 1] | whether site visitor took additional action on the site | 0 | 0, 1 |

**Table 1.** Data dictionary for ab_data1.csv

The data collection began part-way through January 2nd 2017, and finished part-way through January 24th 2017. Overall, as displayed in Figure 1, the number of site visits per day was consistent throughout the data collection period aside from the first and final days, for both new and old page site visits. The conversion proportion was also consistent throughout the data collection period, for both new and old page site visits. The dataset did not appear to have any glaring anomalies that would require considering when doing the analysis.


**Figure 1.** Number of conversions per landing page by date

**Bayesian AB Test**

The dataset consists of counts of successes in a fixed number of independent, identical, converted/not-converted trials, so the likelihood has a binomial distribution. The point of interest is the probability of success (i.e. a conversion) over some number of trials with some number of observed successes. For binomially distributed likelihood, a Beta distribution is appropriate for the prior. The old landing page was assigned a prior of Beta (α = 2, β = 20), representing what has been observed in the past. The new landing page was assigned a prior of Beta (α = 2, β = 20), the same prior. The sequence of Figures 2.1 to 2.3 shows the posterior distributions of the landing page conversion rates with a binomial likelihood and Beta priors, over a varying number of trials. As the number of trials increases, the posterior becomes more concentrated, so the shape of the posterior becomes narrower and more peaked. The probability also converges on a value, and the confidence in this value grows as the number of trials increases.

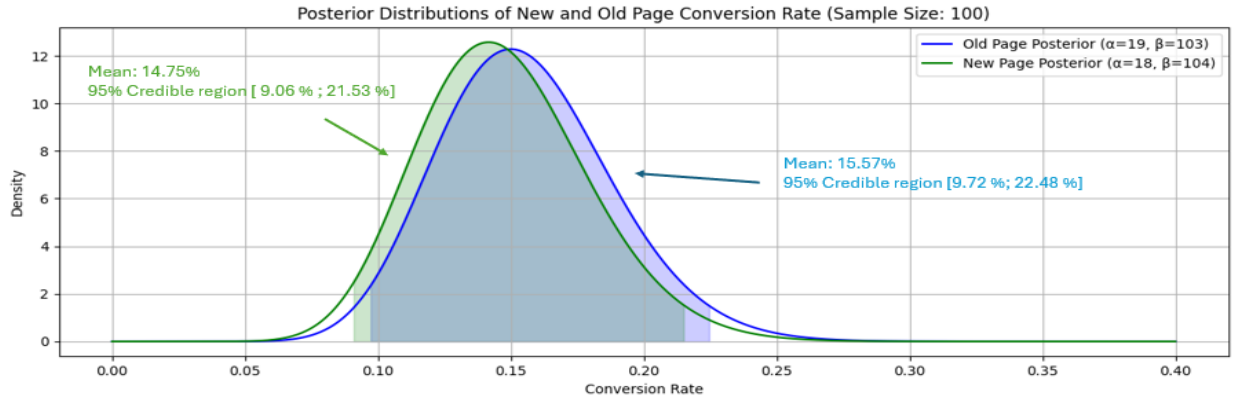


**Figure 2.1** Posterior distributions of landing pages conversion rates after n=100 trials

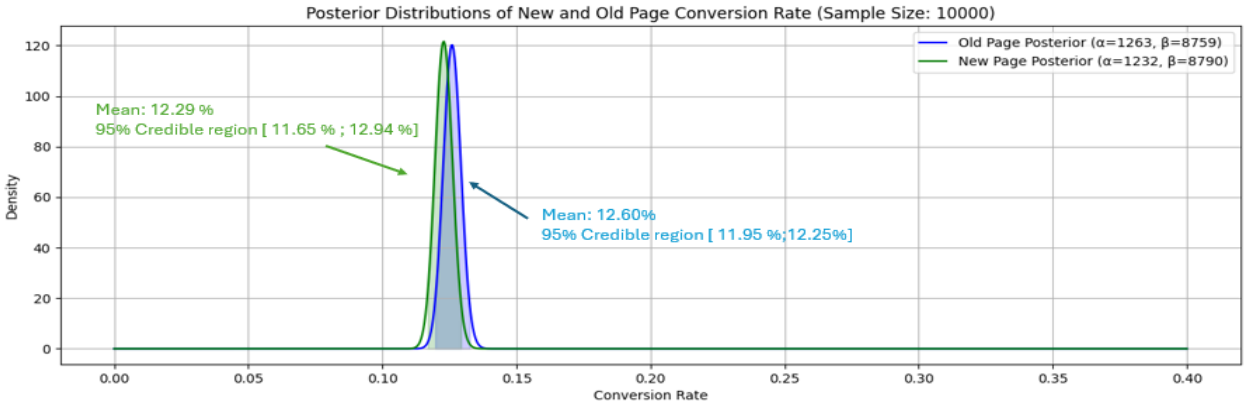


**Figure 2.2** Posterior distributions of landing pages conversion rates after n=10,000 trials

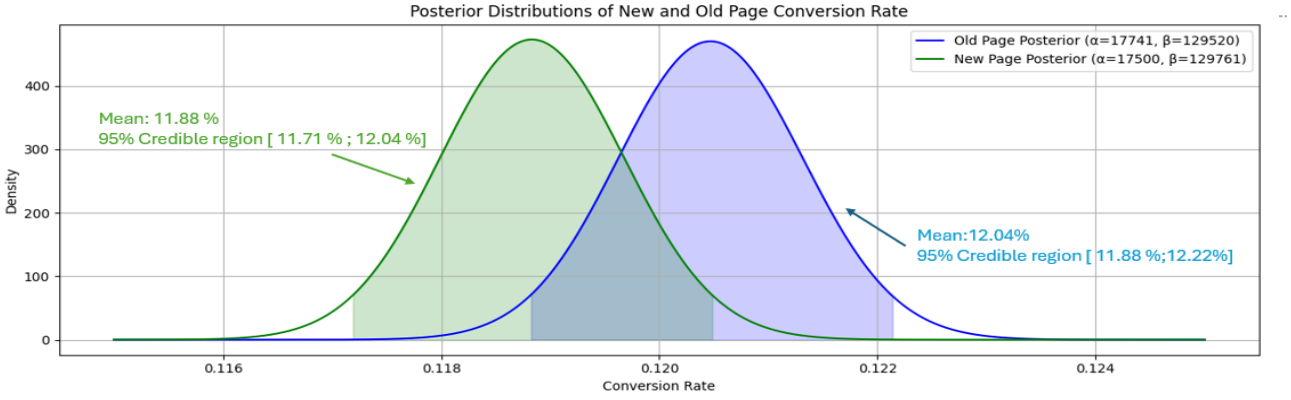


**Figure 2.3** Posterior distributions of landing page conversion rates using entire dataset (note the x-axis has changed)

Notice the intersection in credible regions in Figures 2.3, this suggests possibility that conversion rates are equal. To confirm the old landing page does in fact have a greater conversion rate than the new landing page, we will conduct a simulation based on these two posteriors. We seek to find the posterior distribution for P(new > old).

Given that the marketing department has invested time and additional resources into creating a new landing page, it follows that it's believed the conversion rate for the new lading page is greater than the conversion rate of the old landing page. Thus, we assume the probability that the new conversion rate is greater than the old conversion rate, following a Beta(12,4) distribution (the prior distribution). To build the posterior probability P(new > old), the following algorithm is run:

1. sample n=100 observation from P(new) and P(old).
2. Compare new and old rates to build a binary variable called "success"
3. Update the posterior probability using successes
4. Calculate Kullback–Leibler divergence between current posterior and previous posterior
5. Check the criterion KL-d < 0.01
6. If step 5 is false, iterate again. If step 5 is true stop – this is the posterior probability.

Using the informed prior, it's determined that the informed posterior probability has an expected value of P(new > old | informed prior) = 12.02% with a 95% Credible Interval of [9.35%, 14.95%] - visualized in Figure 3.1. Immediately after updating the posterior only once with n=100, it is evident the prior is not in the right ballpark. Given the belief is not met, it needs to be determined to what extent the prior is influencing the posterior. To evaluate this, the posterior is reconstructed using an uninformed prior, Beta(1,1). The algorithm returns an uninformed posterior with an expected value of P(new > old | uninformed prior) = 9.12 % with 95% Credible Interval of [7.10%, 11.35%] - visualized in Figure 3.2. Looking at the informed posterior, it is determined that the informed posterior is certainly biased to the prior belief as the informed prior is shifted to the right. Both posteriors can be compared visually in Figure 3.3.

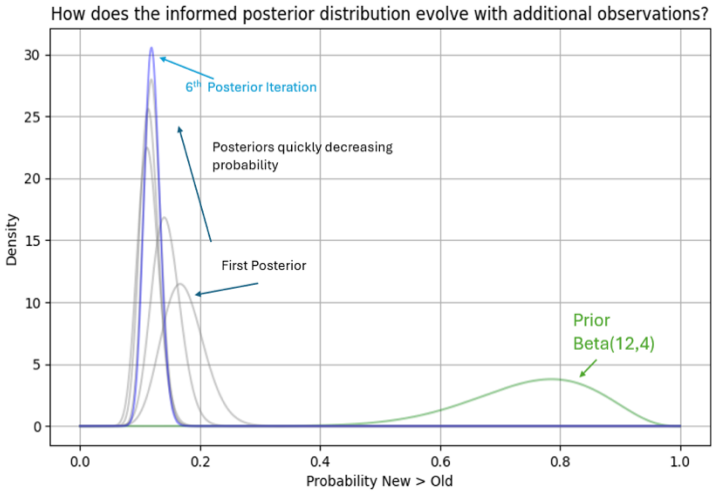


**Figure 3.1** Simulation of posterior distribution P(new > old) using informed prior, Beta(12,4)

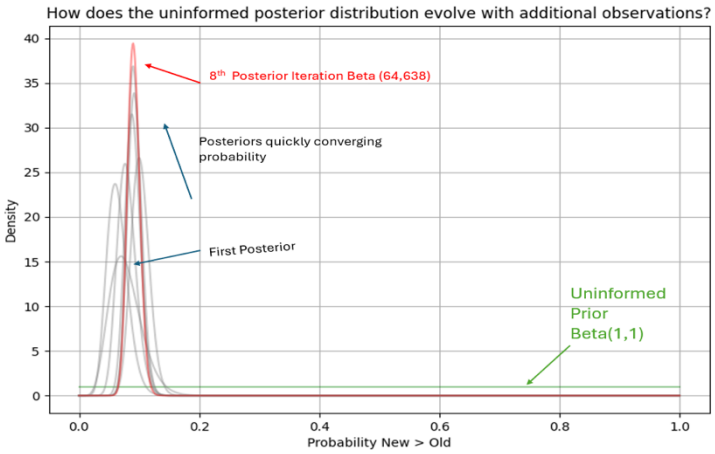


**Figure 3.2** Simulation of posterior distribution P(new > old) using uninformed prior, Beta(1,1)

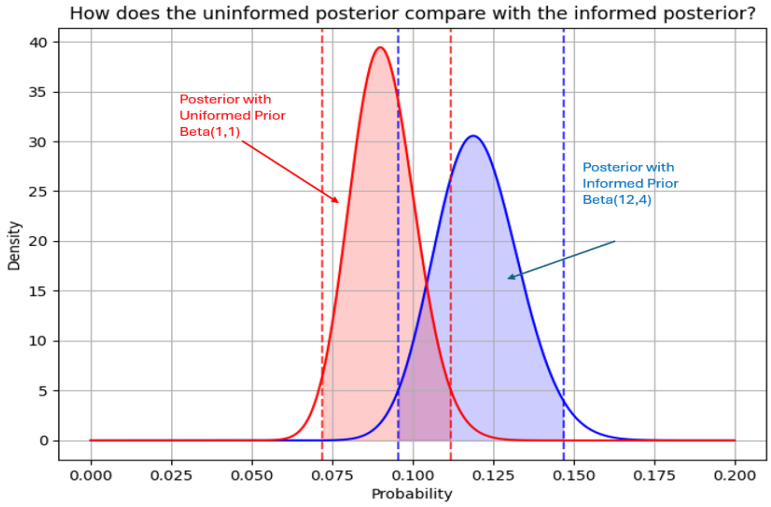


**Figure 3.3** Comparing Simulated Posteriors using different priors

How many iterations would it take for the posterior to "forget" the prior? This question can be reframed as "when will the informed prior be similar to the posterior with an uninformed prior?" To answer this, the informed posterior will be rebuilt with the same algorithm, except instead of comparing to the previous iteration in the KL-d criterion, it will be compared to the uninformed posterior. Once the criterion is met, both posteriors are considered similar enough, implying the informed posterior has "forgotten" the prior. This is visualised in Figure 4.2.

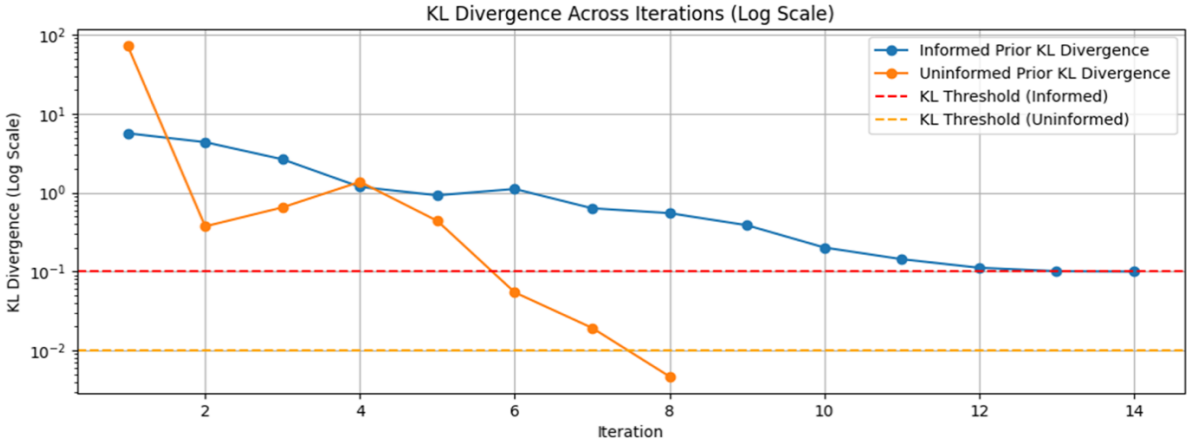


**Figure 4.1** Number of iterations for the posterior to "forget the prior

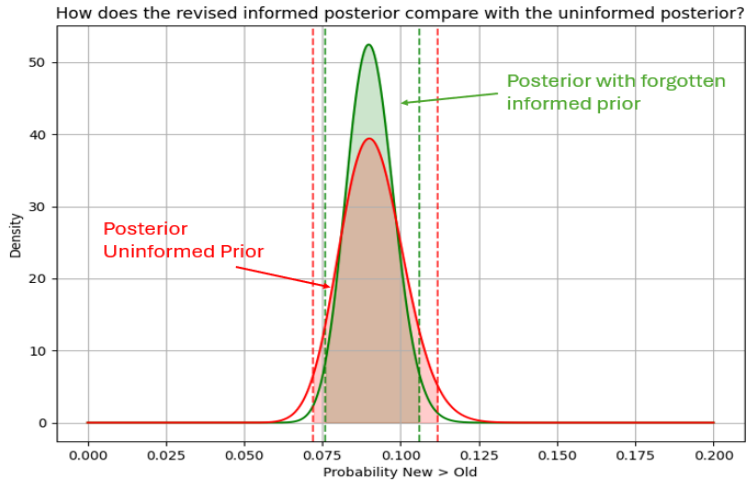


**Figure 4.2** Uninformed Posterior vs Informed Posterior with forgotten prior

| Posterior | Prior Alpha | Prior Beta | Posterior Alpha | Posterior Beta | Mean | Credible Region | Sample size n=100 |
|---|---|---|---|---|---|---|---|
| Informed Prior | 12 | 4 | 74 | 542 | 12.02% | [9.56%; 14.69%] | 6 |
| Uninformed Prior | 1 | 1 | 73 | 729 | 9.10% | [7.21%; 11.18%] | 8 |
| Revised Informed Prior | 12 | 4 | 128 | 1288 | 9.03% | [7.60%; 10.5%] | 14 |

**Table 2.** Comparison of the three posterior distributions

After 14 iterations (1400 samples), the prior Beta(12,4) has been forgotten (Figure 4.1). Note a flaw with the algorithm is the KL-d criterion is not monotonically decreasing, thus the threshold is adjusted appropriately. Table 2 compares all three posterior distributions. Note how similar the mean and credible region are.

## What is the Conclusion?

In conclusion, it is not recommended that the marketing department replace the old landing page with the new landing page. This change is unlikely to result in an increased conversion rate. However, not all customers need to see the same landing page. It is conceivable that the new landing page could have a higher conversion rate among particular subsets of consumers. Further customer segmentation analysis would be required, but it is worth investigating, especially given investments already made into developing the new page.

## Part II – Mimic3d1 & Patients.csv

### Data Exploration

Insurance policymaking is dominated with actuarial and data science, with companies examining risk models to manage client risks and reimbursements. Using a patient dataset and applying Bayesian reasoning, allows the exploration of prior assumptions and how they influence length of stay (LOSdays).

| VARIABLE | TYPE | DESCRIPTION | RANGE | MEAN | num NAs | num DISTINCT VALUES | EXAMPLES |
|---|---|---|---|---|---|---|---|
| hadm_id | Integer | Unique hospital stay ID | 100001-199999 | 149970.81 | 0 | 58976 | 100001 |
| gender | Categorical | Patient sex. | NA | NA | 0 | 2 | F |
| age | Integer | Age at admission | 0-89 | 53.01 | 0 | 77 | 35 |
| LOSdays | Float | Hospital stay length (days) | 0-294.63 | 10.11 | 0 | 1884 | 2.33 |
| admit_type | Categorical | Admission type | NA | NA | 0 | 4 | EMERGENCY |
| admit_location | Categorical | Source of admission | NA | NA | 0 | 9 | CLINIC REFERRAL/PREMATURE |
| AdmitDiagnosis | Character | Initial diagnosis | NA | NA | 25 | 15691 | CORONARY ARTERY DISEASE |
| insurance | Categorical | Insurance type | NA | NA | 0 | 5 | Medicaid |
| religion | Categorical | Patient's self-reported religion | NA | NA | 458 | 20 | PROTESTANT QUAKER |
| marital_status | Categorical | Patient's marital status | NA | NA | 10128 | 7 | MARRIED |
| ethnicity | Character | Patient's recorded ethnicity. | NA | NA | 0 | 41 | BLACK/AFRICAN AMERICAN |
| NumCallouts | Float | ICU callouts count | 0-4.76 | 0.1 | 0 | 127 | 0.1 |
| NumDiagnosis | Float | Total diagnoses | 0-450 | 2.68 | 0 | 1259 | 1.09 |
| NumProcs | Float | Total procedures | 0-275 | 0.79 | 0 | 594 | 0.17 |
| AdmitProcedure | Character | Primary procedure | NA | NA | 0 | 1277 | Non-invasive mech vent |
| NumCPTevents | Float | CPT-coded events count | 0-225 | 1.07 | 0 | 661 | 2.43 |
| NumInput | Float | Input events count | 0-6825 | 30.38 | 0 | 11174 | 11.46 |
| NumLabs | Float | Lab tests count | 0-5175 | 46.42 | 0 | 10517 | 1.88 |
| NumMicroLabs | Float | Microbiology tests count | 0-375 | 1.22 | 0 | 1246 | 0.81 |
| NumNotes | Float | Clinical notes count | 0-7500 | 6.46 | 0 | 1183 | 1.59 |
| NumOutput | Float | Output events count | 0-375 | 7.11 | 0 | 3160 | 16.19 |
| NumRx | Float | Prescriptions issued | 0-750 | 9.59 | 0 | 3494 | 10.05 |
| NumProcEvents | Float | All procedure events count | 0-100 | 0.69 | 0 | 856 | 4.88 |
| NumTransfers | Float | Ward/service transfers count | 0-125 | 1.09 | 0 | 433 | 0.49 |
| NumChartEvents | Float | Clinical observations count | 0-49325 | 528.51 | 0 | 43628 | 1221.39 |
| ExpiredHospital | Integer | In-hospital death (1=yes, 0=no) | 0-1 | 0.1 | 0 | 2 | 0 |
| TotalNumInteract | Float | Patient-care interactions count | 0-68600 | 636.12 | 0 | 46891 | 1482.53 |
| LOSgroupNum | Integer | Stay length group (categorical) | 0-3 | 1.36 | 0 | 4 | 1 |

The dataset contains information on hospital admissions and patients' demographic characteristics with no missing values in most variables, including 58,976 unique hospital stay ID's and spanning a variety of demographics and clinical events. The patient age at admission ranges from 0 to 89 with a mean of 53. Length of hospital stay (LOSdays) ranges widely from less than a day to nearly 295 days, with a mean of 10.1 days, indicating a skew towards shorter stays. Some categorical variables include gender, insurance type and admission type. Even though most of the dataset is complete, there are some

variables such as religion and marital status which contain missing values. Significantly, clinical event counts such as NumChartEvents, show large variances well reflecting the complexity of patients' care needs. Overall the data appears to be structured well enough to model and analyze from the perspective of an insurance company.

## Basic Variable Relationships

The usage of linear regression to gain an idea about priors in Bayesian modelling can lead to overconfidence, especially when the same dataset is used to both construct the prior and compute the likelihood. However, it can still help develop a better prior by testing assumptions and suspicions.

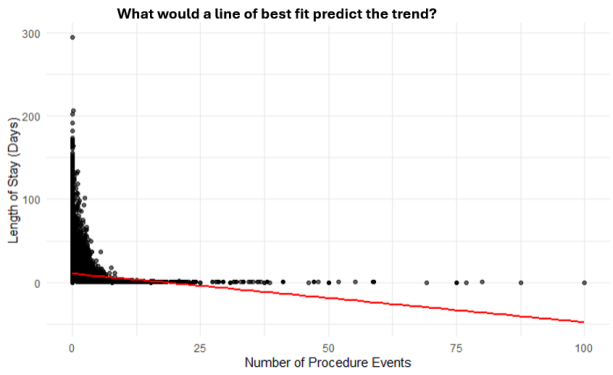


**Figure I** What would a line of best fit predict the trend?

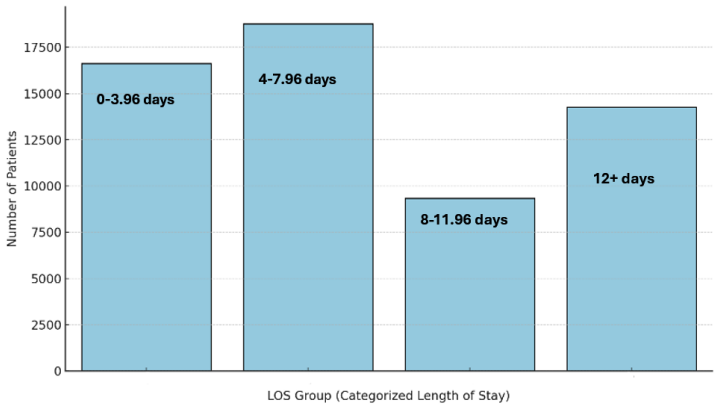


**Figure II** Grouping people by bins, 0-3 based on their length of stay

Beginning with traditional linear regression, for the relationship shown in **Figure I**, the eye test doesn't do any favours, giving a declining slope. This is the case for ALL variables when used as a predictor for LOSdays. This is because of the high volume of data points in mimic3d1, with the vast majority of data points being in the 0-14 day range for LOSgroupNum, as shown in **Figure II**. This makes lines of best fit decline regardless of numerical relationships. While 11 of the 15 numerical variables were statistically significant via p-values, only 4 of them were able to explain 1% or more of the variation in LOSdays. These were NumProcEvents (1%), NumCallouts(4.26%), NumDiagnosis (2.43%), NumTransfers (2.82%), and NumRx (2.87%). It is easy to imagine how these variables relate to the length of stay, because they all reflect a patient's engagement with the hospital, rather than prior demographic information that is unrelated to their visit to the hospital, such as ethnicity, marital status, and insurance coverage.

But things that happen during the stay are of little to no use for pre-visit actuarial data, because insurance policies are written and sold before hospital visits, which further emphasizes and justifies the use of Bayesian analysis, which allows insurance companies to incorporate prior beliefs and quantify uncertainty in predictor importance. However, hospital activity data can tell something insightful with regards to forecasting in-hospital activities costs.

## Probability Distribution

Having examined how individual variables relate to length of stay through traditional linear regression, now shifting the focus from estimating average outcomes to postulating a distribution that considers the full range of outcomes. OLS regression analysis may help in understand which variables are strongly associated with LOSdays, but it does not capture the overall shape or uncertainty of patient stay durations. The next step of data exploration is to construct an unbiased posterior distribution for LOSdays, using uninformative priors that reflect the most neutral approach possible, minimizing assumptions and ensuring that inferences are driven entirely by the observed data.

According to the Kolmogorov-Smirnov (KS) statistic, the distribution that fit best to the values of LOSdays is the log-normal distribution, which is used as the likelihood function. The log-normal distribution's good fit is demonstrated even more clearly in (**Figure III)**, because the curve reflects the key visual characteristic of the data: a steep peak

concentrated around lower LOSdays values, with a long tapering tail toward higher values. This mirrors the real distribution, where the majority of patients have short hospital stays.

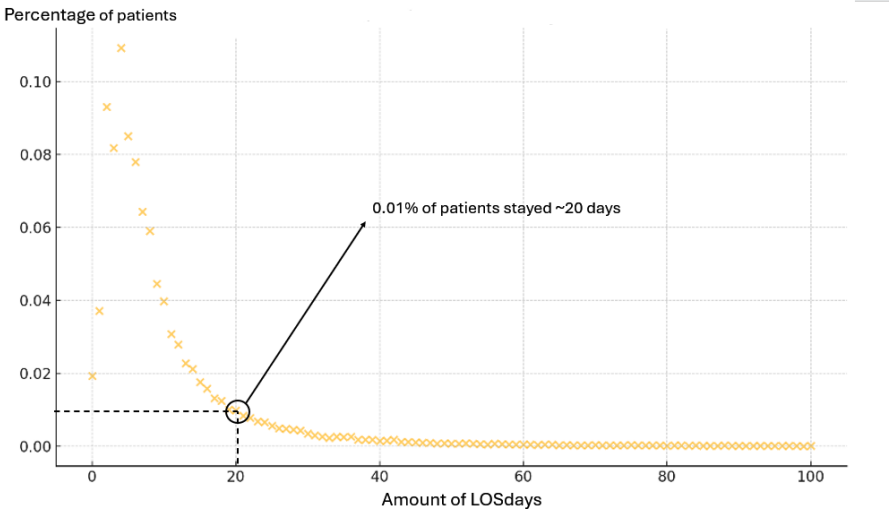


**Figure III** Proportion of Patients by Length of Stay

To avoid introducing strong assumptions, constructing an uninformative prior using the **Maximum Entropy Principle** would ensure the start with the least biased distribution. This would be the uniform distribution, where insurance companies are essentially clueless about the value of the parameters of the distribution within a certain range. Since the log-normal likelihood function depends on the parameters mean and standard deviation $(\mu, \sigma)$, insurance companies will place a uniform prior distribution $\mu \sim U(1,20)$ and $\sigma \sim U(0.1,10)$, which is a very wide range to stick to the methodology of minimal biasedness. Using MCMC sampling with uninformative priors, 12,000 were drawn for posterior samples for both the mean and standard deviation of the log-transformed LOSdays.

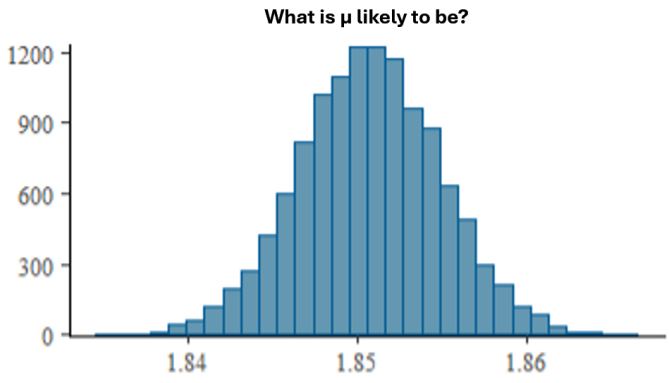


**Figure IV** Posterior Distribution of $\mu$

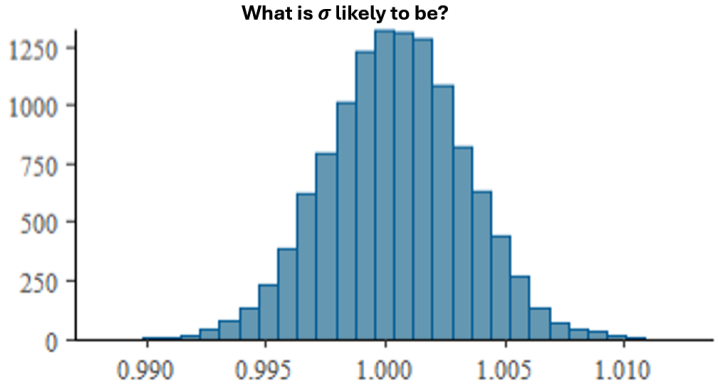


**Figure V** Posterior Distribution of $\sigma$

As shown in **Figure IV** and **V**, the values center at approximately $\mu = 1.85$ $\sigma = 1$ within a very narrow interval, ensuring confidence that the values of the parameters for the posterior distribution are $X \sim LogNormal(1.85,1)$. Thus, with minimal assumptions about the data during the exploratory process, it was possible to construct not only a general distribution but also one with a mean and variance that can be estimated with reasonable confidence. But realistically, a distribution no matter how unbiased and well-fitted, can offer more than a few promising insights about predicting patient stays and costs without complementing it with incredibly complex existing actuarial data.

However, findings about related data to LOSdays can be telling counterintuitively about stay and in-hospital activity, which is explored in the next section.

**Bayesian Regression analysis**

Bayesian regression analysis allows the exploration of the posterior distribution of predictor coefficients $\beta_i$, which in Layman's terms, leads to the estimation of how much each factor likely affects the outcome, and the certainty around those estimates.

In the dataset, all numerical variables with the exception of age, are about patient activity after admission. When it comes to the discussion about predicting LOSdays, one would intuitively assume that the higher the activity, the longer the stay, but tests state otherwise.

Maintaining a cautious approach with minimal prior assumptions, prior distributions were set for the parameters $\beta_i \sim N(0,10^2)$ and $\sigma \sim Exponential(1)$, which are considered uninformative. It is worth mentioning a log-normal model was assumed for the regression analysis here, because traditional regression alone cannot capture the heavy right-tail that is seen in LOSdays distribution. Although neutrality in the methodology has been emphasized, this assumption reflects a practical and data-driven choice, the log-normal model better handles skewed data, ensuring no predictions are negative, and it offers more realistic estimates for patients with longer stays.

For the first posterior distribution of predictor coefficients, the focus was placed on in-hospital activity variables that had low collinearity, such as NumCallouts, NumInput, NumMicroLabs, NumCPTevents, NumChartEvents, and NumTransfers. Different selections of variables for posterior distributions reflect the purpose of the variables and how they relate to each other. For instance, the usage of TotalNumInteract wouldn't be preferred in this regression test because it represents a sum of many of these variables, making it redundant. Among these variables, only NumCallouts and NumTransfers had an estimated coefficient higher than |0.01|, with NumCallouts having a -1.40 coefficient within a 95% confidence interval of $\pm$ 0.2, and NumTransfers having a -0.19 coefficient within a 95% confidence interval of $\pm$ 0.01 from **Figure VI**. This indicates that the number of callouts has a very strong negative effect, with increasing callouts giving decreasing LOSdays, and that the number of transfers has a small yet consistent effect on reducing LOSdays. Though there is little to interpret from this, the results contribute to a surprising point: all variables about hospital activity have a "reducing" effect on LOSdays.

Regression Coefficients:

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 2.19 | 0.02 | 2.14 | 2.23 | 1.00 | 10138 | 9277 |
| NumCallouts | -1.40 | 0.10 | -1.60 | -1.19 | 1.00 | 9683 | 8436 |
| NumInput | -0.00 | 0.00 | -0.00 | 0.00 | 1.00 | 11900 | 8061 |
| NumMicroLabs | -0.01 | 0.00 | -0.01 | -0.00 | 1.00 | 16570 | 9096 |
| NumCPTevents | -0.01 | 0.01 | -0.03 | 0.02 | 1.00 | 8098 | 7507 |
| NumChartEvents | -0.00 | 0.00 | -0.00 | 0.00 | 1.00 | 12279 | 9717 |
| NumTransfers | -0.19 | 0.01 | -0.20 | -0.18 | 1.00 | 10154 | 8056 |

**Figure VI** What Were the Numerical Relationships of in-Hospital Activity and LOSdays?

This gives us an important observational insight, which is that severity of a patient's condition is not necessarily reflected in the volume of in-hospital activity. In fact, patients who stay longer may do so not because of intense intervention, but due to prolonged observation, systemic delays, or non-procedural conditions. This goes against the initial assumption that high activity equates to high risk. Relying solely on activity volume to determine LOS-related insurance coverage may lead to erroneously predicting costs and risks related to patient condition severity.

For the second posterior distribution, variables that more broadly categorize in-hospital activity were used. These two were TotalNumInteract, and ExpiredHospital. Surprisingly, though in-hospital activity variables in the first posterior distribution showed them to have a moderately impactful relationship, the second regression analysis says otherwise with TotalNumInteract having an almost absent effect on LOSdays via. However, ExpiredHospital seemed to have a major impact on LOSdays, with a coefficient of -0.34 with a 95% confidence interval of $\pm$ 0.13 via **Figure VII.**

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 1.95 | 0.02 | 1.91 | 2.00 | 1.00 | 8322 | 8510 |
| TotalNumInteract | -0.00 | 0.00 | -0.00 | -0.00 | 1.00 | 10955 | 8731 |
| ExpiredHospital | -0.34 | 0.07 | -0.47 | -0.21 | 1.00 | 6928 | 5909 |

**Figure VII** What Were the Numerical Relationships Between Interactions, Deaths, and LOSdays?

This further drives home the point that a patient's condition severity is not reflected in longer stays, as this strong relationship between shorter stays and deaths suggest that shorter stays are associated with more severe admissions.

Following, an important concern that was brought by the insurance company, which was the probability that patients stay longer than two days and miss work. With only the patient admissions from patients.csv, the decision of using the regression model constructed in the first posterior distribution to find the posterior probability was made **(Figure VIII)**, and found that staying more than 2 days was a very confident prediction, and that probabilities for shorter stays were much lower.

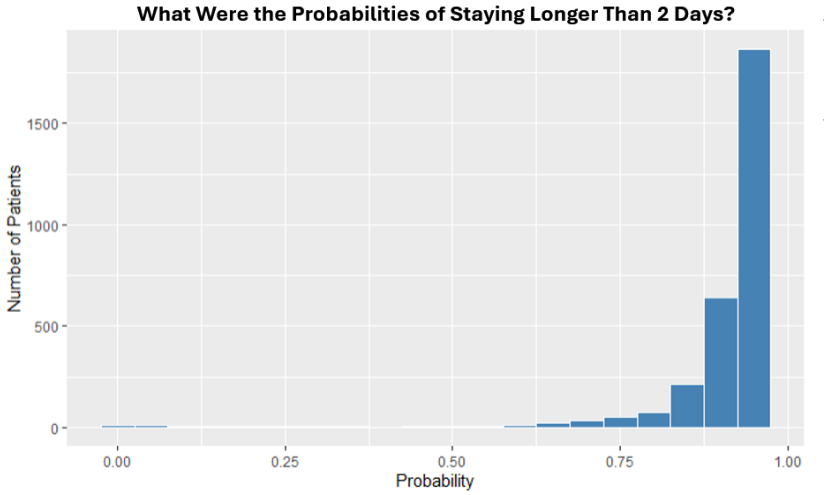


**Figure VIII** What Were the Probabilities of Staying Longer Than 2 Days?

This task was significant, as missing work is an important consideration for hospital reimbursements because of the financial consequences for patients who have longer stays in the hospital. The strong skew toward high probabilities of staying more than 2 days may reflect the fact that many hospital admissions follow structured care pathways, rather than the magnitude of patients' conditions. For example, scheduled procedures, post-surgical monitoring, and medical workups all inherently require multi-day observation. This means that longer hospital stays may not always signal higher medical severity, but rather planned and predictable protocols.

## Conclusion

In conclusion, analysis reveals that hospital length of stay (LOSdays) is shaped less by individual procedures or activity volume than expected. Contrary to assumptions that more activity signals longer stays, some of the most severe patient outcomes, including death, were associated with shorter stays and fewer recorded interactions. The probability modeling showed that most patients are expected to stay beyond two days with high certainty, suggesting that structured care pathways play a significant role in LOS outcomes. These findings highlight a critical consideration for insurance policy design: reimbursement strategies should not rely solely on surface-level indicators like procedural count or LOS duration. Instead, the insurance company ought to invest more in understanding hospital protocols, care plans, and treatment timelines, and recognize how standardized workflows and admission types shape patient treatment trajectories. Knowing these will allow for more refined reimbursement strategies and plans that both preserve the financial well-being of the clients, while also adhering to the financial interests of the company.