Bayesian GLM Analysis of Medical Insurance Charges

Han Liu **Junlin Huo** **Haoyue Zhang**
Department of Applied Mathematics and Statistics
Johns Hopkins University
hliu192@jh.edu jhuo5@jh.edu hzhan294@jh.edu

Abstract

Medical insurance charges vary widely across individuals, driven by differences in demographic and behavioral risk factors. While prior studies have established that characteristics such as age, BMI, smoking behavior, and geographic region meaningfully affect healthcare expenditures, most analyses rely on classical regression techniques that provide limited uncertainty quantification and struggle with the heavy-tailed nature of cost data. In this work, we re-examine these relationships using a Bayesian generalized linear modeling (GLM) framework applied to a publicly available dataset of 1,338 insured individuals in the United States. The Bayesian approach yields full posterior uncertainty quantification and probabilistic comparisons across subgroups, offering a more interpretable and robust modeling strategy for heterogeneous medical cost data. Our framework also provides a flexible foundation for future extensions such as hierarchical or mixture models.

1 Purpose of Project

Medical insurance costs vary widely across individuals, even among those enrolled in similar plans. This variability is closely linked to demographic, socioeconomic, and behavioral factors such as age, obesity, smoking status, and family structure [Borah et al., 2016, Ward et al., 2021, Xu et al., 2015]. Understanding these determinants is essential for accurate premium pricing, risk adjustment, and evidence-based health policy.

A substantial literature documents the influence of individual-level characteristics. For instance, Valero-Elizondo et al. [2018] show that socioeconomic disparities relate to cardiovascular risk factors—including obesity, hypertension, and smoking—that increase healthcare utilization. Mayfield et al. [2021] find that demographic and social factors such as age, sex, race/ethnicity, and insurance type predict high emergency department charges. International evidence reinforces these patterns: Adjei-Mantey and Horioka [2023] identify socioeconomic status, education, and health behaviors as major determinants of medical expenditures in Ghana. These findings motivate studying age, BMI, smoking status, number of dependents, and region in the U.S. dataset used here.

Many prior analyses rely on classical frequentist models such as linear, logistic, or quantile regression [Morid et al., 2018, Barber and Thompson, 2004, Jones et al., 2015]. While effective for point estimation, these approaches provide limited uncertainty quantification and may perform poorly with the heavy-tailed, heterogeneous distributions typical of medical cost data.

Bayesian modeling yields full posterior distributions, enabling richer uncertainty quantification and interpretable probabilistic comparisons—for example, assessing the probability that smokers incur higher charges or that BMI positively affects expenditures.

In this project, we apply Bayesian generalized linear models (GLMs) [Gelman et al., 1995] to a medical insurance cost dataset [mosap abdelghany, 2025]. Our primary research question is how demographic and behavioral factors—such as smoking, BMI, and age—affect medical insurance expenditures, and how Bayesian GLMs can quantify these effects and their associated uncertainty.

Preprint.

By combining empirical insights with flexible Bayesian modeling, we aim to provide a clearer characterization of individual-level healthcare expenditures. We fit several Bayesian GLMs tailored to the data and compare them with alternative models, demonstrating that the Bayesian specifications best capture the distributional characteristics of the dataset.

2 Data

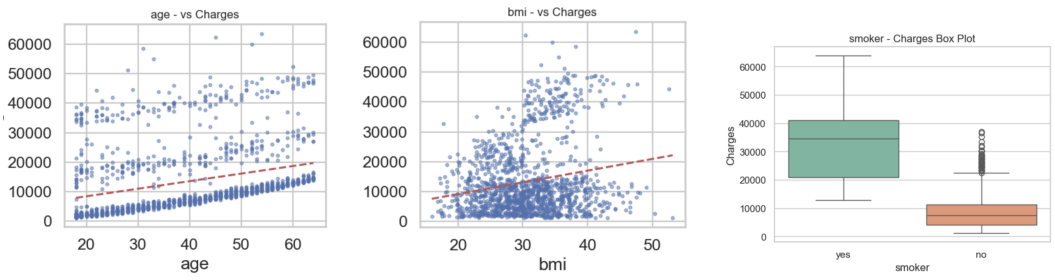
The dataset used in this project is titled *Medical Insurance Cost Dataset*[mosap abdelghany, 2025], and it is sourced from the public machine learning platform Kaggle. This is a well-cleaned dataset containing 1,338 insured individuals in the United States, recording each person's medical insurance charges along with a set of related variables, including: **age**, **sex**, body mass index (**BMI**), number of children covered by the insurance plan (**children**), smoking status (**smoker**), and residential region within the United States (**region**). The target variable is **charges**, which represents the annual medical insurance cost billed to each individual.

These variables cover key factors commonly considered in actuarial science and health economics. Therefore, the dataset is appropriate for analyzing the relationship between individual health characteristics and medical insurance expenditures.

2.1 Descriptive statistics & Explanatory Data Analysis (EDA)

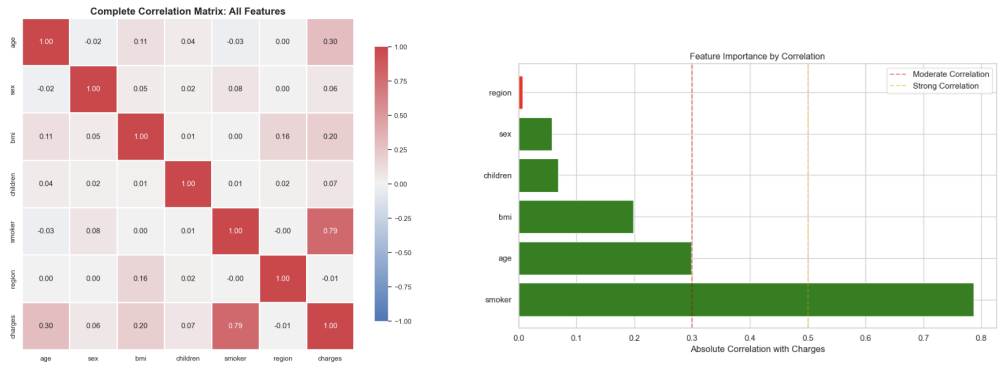
Variable	Type	Notes / Basic Statistics
age	Numerical	Mean $\approx$ 39; range 18–64; increases costs with age
bmi	Numerical	Mean $\approx$ 30; indicator of obesity; mildly right-skewed
children	Numerical (count)	0–5 dependent children; weak effect on charges
charges	Numerical (target)	Highly right-skewed (mean = 13.3k, max $>$ 63k); log-transformed
sex	Categorical	Male / Female; minimal effect on charges
smoker	Categorical	Yes / No; strongest cost driver with large separation
region	Categorical	NE / NW / SE / SW; little regional variation

Table 1: Summary of numerical and categorical variables in the insurance dataset.



(a) Age vs Charges (b) BMI vs Charges (c) Smoker vs Charges

Figure 1: Relationships between selected predictors and insurance charges. Smoker status shows the strongest separation, while age and BMI exhibit moderate positive associations.



(a) Correlation Matrix (b) Correlation-based Feature Importance

Figure 2: Multivariate correlation analysis of predictors and insurance charges.

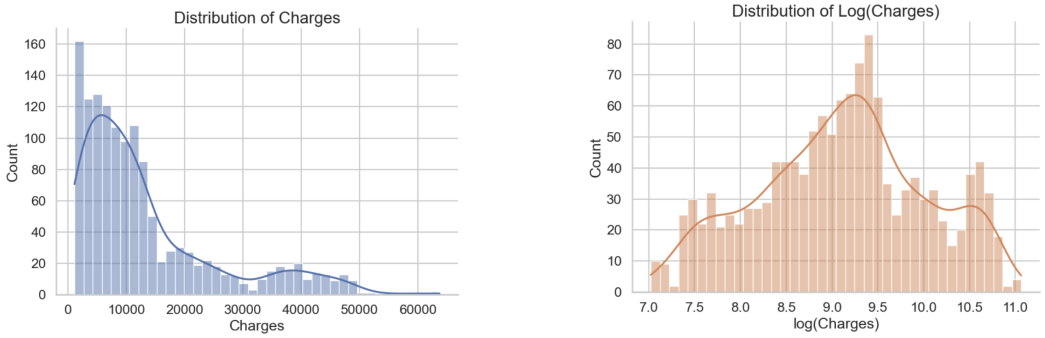
When examining the numerical predictors, both age and BMI show clear positive associations with medical insurance costs: older individuals tend to incur higher expenditures, and higher BMI levels are similarly linked to increased charges. In contrast, the number of children displays almost no meaningful correlation with charges.

For the categorical predictors, smokers consistently incur substantially higher charges than non-smokers. By comparison, region and sex contribute minimally to cost variation.

Correlation analysis reinforces these findings. Smoking status demonstrates by far the strongest correlation with charges (0.79), while age (0.30) and BMI (0.20) show moderate positive correlations. Other variables—including children, sex, and region—have correlations near zero. Overall, both the heatmap and the feature-importance plot confirm that smoking, age, and BMI are the primary drivers of medical insurance costs.

These findings motivate the use of a Bayesian GLM focusing on these key predictors.

2.2 Data Cleaning & Data Transformation



(a) Distribution of Charges (b) Distribution of log(Charges)

Figure 3: Comparison of raw and log-transformed insurance charges.

A log transformation is applied to the target variable *charges* to reduce right-skewness and produce a more symmetric distribution for modeling. Since the dataset is already well-cleaned as described above, no additional data-cleaning procedures are required.

3 Literature Review

A substantial body of empirical work has examined the individual-level determinants of healthcare and medical insurance costs. Demographic, socioeconomic, and behavioral variables—such as age, BMI, smoking status, and geographic region—are consistently shown to influence expenditures across diverse populations [Borah et al., 2016, Ward et al., 2021, Xu et al., 2015, Valero-Elizondo et al., 2018, Mayfield et al., 2021, Adjei-Mantey and Horioka, 2023]. These findings motivate the inclusion of similar variables in our analysis.

In addition to this empirical literature, several studies have explored Bayesian methods for modeling healthcare expenditures. Much of this work focuses on handling the heavy-tailed, skewed, and heterogeneous nature of medical cost data. For example, Basu et al. [2011] and Deb et al. [2006] discuss Bayesian two-part and semi-parametric models for cost distributions, demonstrating superior flexibility over classical approaches. Bayesian mixture and latent class models have also been applied to healthcare utilization and expenditures, allowing for unobserved heterogeneity in patient populations [Deb and Trivedi, 1997, Burgette and Reiter, 2013]. Other studies have applied Bayesian generalized linear frameworks—such as gamma or log-normal regression—to cost estimation problems, highlighting improved inference and uncertainty quantification in the presence of skewness and outliers [Venturini et al., 2008, Geweke, 2010].

Although Bayesian models have been successfully used in specialized econometric and health-services contexts, they remain less commonly applied in standard insurance-pricing datasets. Positioned within this landscape, our study contributes by applying Bayesian generalized linear models (GLMs) to a widely used medical insurance dataset. In contrast to existing Bayesian work that often focuses on mixture structures or treatment effects, we illustrate how a straightforward Bayesian GLM can provide clear uncertainty quantification and interpretable probabilistic statements in a typical cost-prediction setting.

4 Methods

4.1 Bayesian Model & Relation to Research Question

Our research objective is to model and predict individual-level medical insurance charges and to quantify how demographic and behavioral covariates (age, BMI, number of children, sex, region, and smoking status) influence cost. Because medical expenditures are strongly right-skewed, we model the log-transformed response,

$$y_i = \log(\text{charges}_i),$$

and standardize all continuous predictors. The general sampling model for all Gaussian-based specifications is

$$y_i \mid \mu_i, \sigma^2 \sim N(\mu_i, \sigma^2),$$

with μ_i determined by each model's regression structure. For the heavy-tailed smoker model, we introduce a latent robustness weight λ_i so that

$$y_i \mid \mu_i, \sigma^2, \lambda_i \sim N\big(\mu_i,\, \sigma^2/\lambda_i\big)\,, \qquad \lambda_i \sim \Gamma\Big(\frac{\nu}{2}, \frac{\nu}{2}\Big)\,,$$

which produces a Student-t marginal likelihood.

4.2 Model A: Baseline Linear Effects

Model A specifies a homoscedastic Gaussian linear regression with only main effects:

$$X_A = \big[1,\ \text{age\_std},\ \text{bmi\_std},\ \text{children\_std},\ X_{\text{cat}}\big],$$

where X_{cat} contains one-hot encoded sex, smoker, and region indicators (with baseline levels omitted). The regression mean is

$$\mu_i = x_{A,i}^{\top}\beta_A.$$

Model A provides a simple, interpretable baseline linking each covariate linearly to expected log-costs.

4.3 Model C: Nonlinear and Smoker-Interaction Regression

Model C extends Model A by adding nonlinear age and BMI effects and smoker-specific deviations. In addition to standardized predictors, we form quadratic terms:

$$\text{age\_std\_sq} = \text{age\_std}^2, \qquad \text{bmi\_std\_sq} = \text{bmi\_std}^2,$$

and smoker interactions:

$$\begin{aligned} \text{age\_x\_smoker} &= \text{age\_std} \cdot \text{smoker\_yes}, \\ \text{bmi\_x\_smoker} &= \text{bmi\_std} \cdot \text{smoker\_yes}, \\ \text{age\_sq\_x\_smoker} &= \text{age\_std\_sq} \cdot \text{smoker\_yes}. \end{aligned}$$

The Model C design matrix is

$$\begin{aligned} X_C = \big[&1, \text{age\_std}, \ \text{age\_std\_sq}, \ \text{bmi\_std}, \ \text{bmi\_std\_sq}, \ \text{children\_std}, \\ &X_{\text{cat}}, \ \text{age\_x\_smoker}, \ \text{bmi\_x\_smoker}, \ \text{age\_sq\_x\_smoker} \big]. \end{aligned}$$

The regression mean is $\mu_i = x_{C,i}^\top \beta_C$. Model C captures nonlinear curvature and smoker–nonsmoker heterogeneity while maintaining interpretability.

4.4 Model S_non: Non-Smokers Under Model C Structure

Exploratory analysis revealed that smokers exhibit heavier tails and greater variability than non-smokers. This motivates the use of subgroup-specific models. Accordingly, the dataset was partitioned into smoker (n=274) and non-smoker (n=1064) subsets based on the binary smoking indicator.

Non-smokers exhibit smooth cost patterns with no extreme outliers. For non-smokers, we apply the Model C structure but drop any predictors that are structurally zero (smoker indicators and smoker interactions). The mean structure is

$$\mu_i = \beta_0 + \beta_1 \text{age\_std}_i + \beta_2 \text{age\_std\_sq}_i + \beta_3 \text{bmi\_std}_i + \beta_4 \text{bmi\_std\_sq}_i + \beta_5 \text{children\_std}_i + \delta^\top \text{sex/region}_i.$$

Because the noise distribution for non-smokers is approximately Gaussian, no robustness weights are required ($\lambda_i \equiv 1$).

4.5 Model S_smoker_plus: Heavy Tails, Splines, and Interactions

Smokers display substantially heavier tails and more complex nonlinear patterns. To model these features, we introduce spline expansions for age and BMI using truncated power bases with knots at empirical quartiles:

$$f_{\text{age}}(x) = \beta_1 x + \beta_2 x^2 + \sum_{j=1}^{J} \beta_{a,j} (x - k_{a,j})_+, \qquad f_{\text{bmi}}(x) = \beta_3 x + \beta_4 x^2 + \sum_{j=1}^{J} \beta_{b,j} (x - k_{b,j})_+.$$

The regression mean is

$$\mu_i = \beta_0 + f_{\text{age}}(\text{age\_std}_i) + f_{\text{bmi}}(\text{bmi\_std}_i) + \beta_{\text{int}}(\text{age\_std}_i \cdot \text{bmi\_std}_i) + \beta_5 \text{children\_std}_i + \delta^\top \text{sex/region}_i.$$

To accommodate extreme smoker observations, we use the Student-t likelihood described earlier via the latent weights λ_i, enabling adaptive downweighting of large residuals and improved robustness.

4.6 Prior Specification

All models share the same weakly informative priors:

$$\begin{aligned} \beta &\sim N(0, \tau^2 I), \qquad \tau = 100, \\ \sigma^2 &\sim \text{Inv-Gamma}(a_0, b_0), \qquad a_0 = 2, \ b_0 = 1. \end{aligned}$$

For the heavy-tailed smoker model S_smoker_plus we additionally specify

$$\lambda_i \sim \Gamma\Big(\frac{\nu}{2}, \frac{\nu}{2}\Big), \qquad \nu = 3,$$

producing a Student-t marginal error.

4.7 MCMC Algorithm

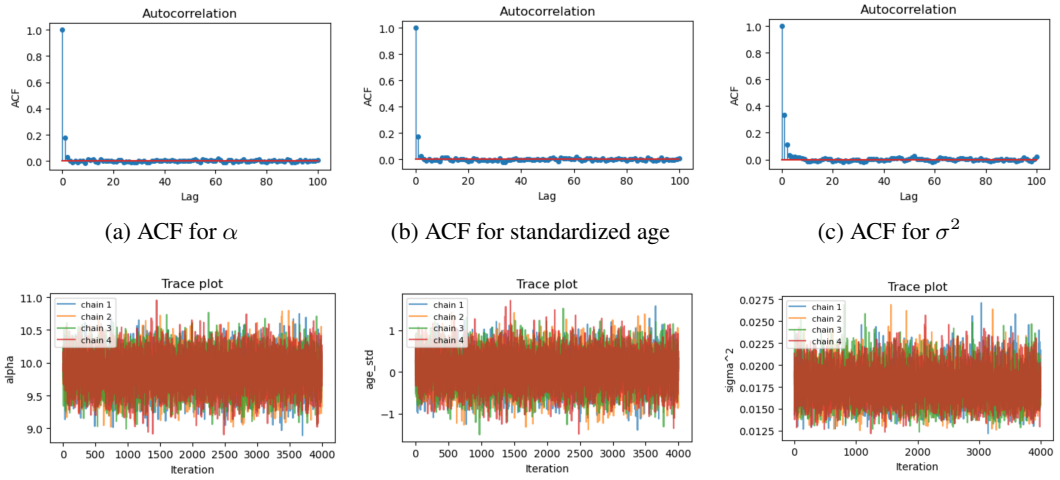
All models are estimated using Gibbs sampling with four parallel chains. Gaussian models (A, C, S_non) alternate between:

$$\beta \mid \sigma^2, y \sim N(m_n, V_n), \qquad \sigma^2 \mid \beta, y \sim \text{Inv-Gamma}\big(a_0 + \tfrac{N}{2},\ b_0 + \tfrac{1}{2}(y - X\beta)^\top (y - X\beta)\big)\,.$$

For S_smoker_plus, the latent weights are updated via:

$$\lambda_i \mid \beta, \sigma^2, y_i \sim \Gamma\left(\frac{\nu+1}{2},\ \frac{\nu + (y_i - \mu_i)^2/\sigma^2}{2}\right).$$

To sample from the distribution, we run 4 chains in paralle. Each chain runs for 6000 iterations with 2000 burn-in, yielding 16,000 retained draws. Trace plots, ACFs, ESS, and $\widehat{R}$ diagnostics confirm excellent mixing, negligible autocorrelation, and reliable posterior convergence. We show the ESS and $\widehat{R}$ for Model C as well as Model Smoker Plus in Table 2, and we show the Trace plots and ACFs in Figure 4 for Model Smoker Plus for reference.



(a) ACF for α (b) ACF for standardized age (c) ACF for σ^2

(d) Trace Plot for α (e) Trace Plot for standardized age (f) Trace Plot for σ^2

Figure 4: ACF and Trace Plot for **Model S_smoker_plus** for several key parameters. The ACF plots for α, standardized age, and σ^2 all exhibit an immediate drop to near zero after lag 1, indicating minimal serial dependence and efficient mixing across iterations. This rapid decay suggests that the Gibbs sampler produces nearly uncorrelated draws for these parameters. The corresponding trace plots show all four chains fluctuating tightly around a stable mean with no visible drift, separation, or long-term trends. The dense interweaving of chains demonstrates that each parameter has reached its stationary distribution and that the sampler explores the posterior space effectively for Model S_smoker_plus.

Effective Sample Size (ESS) measures how many "independent" draws the Markov chain effectively contains. Because MCMC draws are correlated, the true amount of information is smaller than the raw number of samples. A large ESS (typically above 1,000) indicates good mixing and low autocorrelation. In all models, ESS values range from approximately 6,000 to 11,000, demonstrating that the Gibbs sampler produces highly efficient updates.

R hat (Gelman–Rubin statistic) compares between-chain and within-chain variance to assess convergence. Values close to 1.00 indicate that all chains have converged to the same posterior distribution. The parameters for all models achieve $\widehat{R} \approx 1.00$ (within the acceptable range of $[0.99, 1.01]$), confirming that the chains are well-mixed and that posterior summaries are reliable.

Table 2: Predictive performance of candidate models.

Model	RMSE	95% Coverage
Model A (main effects)	8,365	93.6%
Model C (poly + interactions)	5,207	94.7%
Model S_non	4,641	93.7%
Model S_smoker_plus (t-reg + splines)	**4,507**	**97.4%**

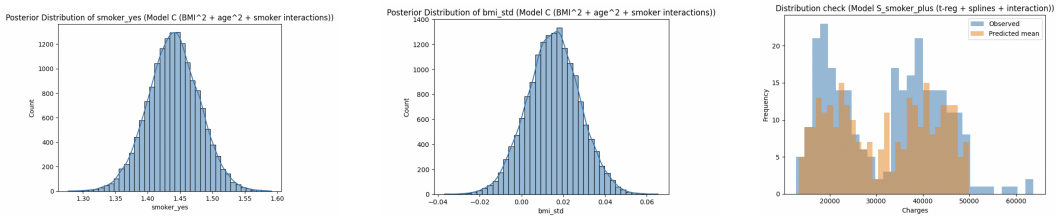
5 Results

5.1 Predictive Performance Overview

We compare four models with increasing complexity: the baseline linear model (Model A), a model with polynomial terms and interactions (Model C), a version of Model C fitted only to non-smokers (Model S_non), and a Student-t spline model for smokers (Model S_smoker_plus). Their RMSE and 95% interval coverage are shown in Table 2. Model A has the largest error (RMSE = 8,365). After adding nonlinear age/BMI terms and smoker interactions, Model C reduces the RMSE to 5,207, which shows that these additional terms help the model capture more structure in the data.

The two subgroup models perform better. Model S_non improves to 4,641, while the Student-t spline model for smokers (Model S_smoker_plus) achieves the lowest error (4,507) and the highest coverage. This suggests that smokers and non-smokers follow different patterns, and that heavy-tailed errors are useful for handling large or irregular values in the smoker group.

Figure 5 displays the observed vs. predicted plots for each model. As complexity increases, the points move closer to the 45-degree line. The improvement is especially visible for observations with high charges, where the baseline model shows clear underestimation but the later models fit more closely.



(a) Smoker coefficient (Model C) (b) BMI squared term (Model C) (c) Model Smoker Plus

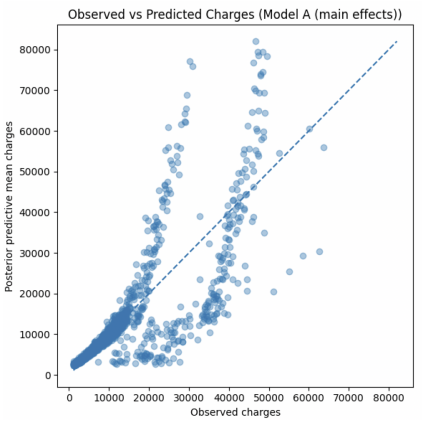
Figure 6: Posterior inference for smoker effect (a), BMI nonlinear term (b), and subgroup predictive distribution for Model Smoker Plus (c). From figure (c), we observe a nearly mixture of 2 Normal Distributions for people who smoke, indicating a potential subgroup among people who smoke.

5.2 Posterior Inference

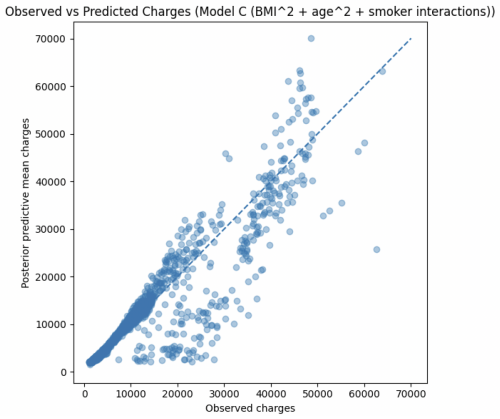
The posterior summaries from Model C reveal nonlinear patterns that are not captured by the baseline model. The negative coefficients for the squared age and squared BMI terms indicate diminishing marginal effects: the increase in medical charges slows as age or BMI grows, consistent with the curvature evident in the data.

Smoking remains one of the most influential predictors across all models. Compared with the baseline, Model C includes interactions between smoking and both age and BMI, and these terms appear significant. This suggests that the relationship between these covariates and medical charges differs structurally for smokers. Figures 6a and 6b show the posterior distributions of the smoker indicator and the BMI-squared term; both are sharply concentrated away from zero, confirming their strong and nonlinear contributions.

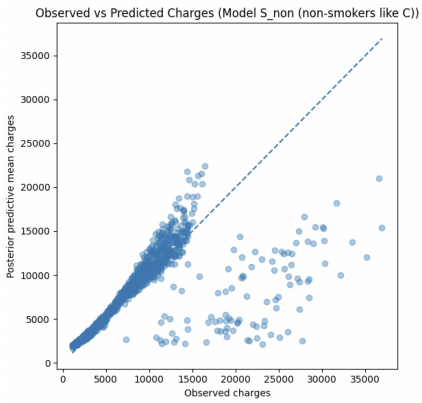
A more complex pattern arises in the posterior predictive distribution for smokers, which exhibits clear bimodality. As shown in Figure 6c, the predictive draws naturally separate into a moderate-cost mode and a high-severity mode. This reflects latent heterogeneity within the smoker subgroup: a dominant moderate-cost component and a smaller, high-cost component with heavier tails. Gaussian



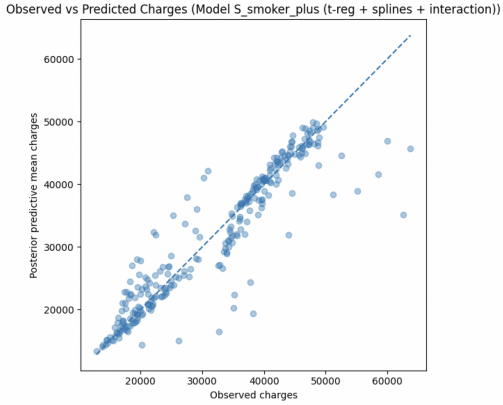
(a) Model A



(b) Model C



(c) Model S_non



(d) Model S_smoker_plus

Figure 5: Observed vs. predicted charges for the four models. As model complexity increases, the points move closer to the 45-degree line, especially for high charges.

models, constrained by unimodality and light-tailed errors, collapse these components toward a single mean and underestimate extreme outcomes. In contrast, the Student–t likelihood and spline-based mean function used in S_smoker_plus introduce the flexibility needed to approximate a mixture-like structure. As a result, posterior predictive draws reproduce the two-mode pattern, indicating that the model captures meaningful multi-modality inherent in the data rather than overfitting noise.

6 Discussion and Conclusion

Across all models, predictive accuracy improves substantially once nonlinear terms and subgroup structures are introduced. Linear models cannot capture the strong curvature and heterogeneity in medical cost data, whereas polynomial terms, spline components, and smoker-specific interactions provide clear gains in fit. Subgroup models—especially the Student–t spline model for smokers—yield the most reliable predictions.

A key finding is that the final model reproduces both the heavy right tail of medical expenditures and the bimodal pattern among smokers. This bimodality reflects latent heterogeneity, potentially separating a moderate-risk group without smoking-related conditions from a high-risk group with advanced comorbidities. This is a promising direction for further study.

Capturing this structure requires flexible error distributions and nonlinear mean functions that adapt across risk strata. Overall, these results show that Bayesian GLMs can recover underlying structures—such as subgroup-specific risk patterns and multimodal cost distributions—that classical methods often obscure, yielding more accurate and interpretable predictions.

References

Kwame Adjei-Mantey and Charles Yuji Horioka. Determinants of health insurance enrollment and health expenditure in ghana: an empirical analysis. *Review of Economics of the Household*, 21(4): 1269–1288, 2023.

Julie Barber and Simon Thompson. Multiple regression of cost data: use of generalised linear models. *Journal of health services research & policy*, 9(4):197–204, 2004.

Anirban Basu, Daniel Polsky, and Willard G Manning. Estimating treatment effects on healthcare costs under exogeneity: is there a 'magic bullet'? *Health Services and Outcomes Research Methodology*, 11(1):1–26, 2011.

Bijan Borah, James Naessens, Kerry Olsen, and Nilay Shah. Explaining obesity-and smoking-related healthcare costs through unconditional quantile regression. *Journal of Health Economics and Outcomes Research*, 1(1):23, 2016.

Lane F Burgette and Jerome P Reiter. Multiple-shrinkage multinomial probit models with applications to simulating geographies in public use data. *Bayesian analysis (Online)*, 8(2):10–1214, 2013.

Partha Deb and Pravin K Trivedi. Demand for medical care by the elderly: a finite mixture approach. *Journal of applied Econometrics*, 12(3):313–336, 1997.

Partha Deb, Murat K Munkin, and Pravin K Trivedi. Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. *Journal of Applied Econometrics*, 21(7): 1081–1099, 2006.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

John Geweke. *Complete and incomplete econometric models*. Princeton University Press, 2010.

Andrew M Jones, James Lomas, and Nigel Rice. Healthcare cost regressions: going beyond the mean to estimate the full distribution. *Health economics*, 24(9):1192–1212, 2015.

Carlene A Mayfield, Marco Geraci, Michael Dulin, Jan M Eberth, and Anwar T Merchant. Social and demographic characteristics of frequent or high-charge emergency department users: A quantile regression application. *Journal of Evaluation in Clinical Practice*, 27(6):1271–1280, 2021.

Mohammad Amin Morid, Kensaku Kawamoto, Travis Ault, Josette Dorius, and Samir Abdelrahman. Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation. In *AMIA annual symposium proceedings*, volume 2017, page 1312, 2018.

mosap abdelghany. Medical insurance cost dataset, 2025. URL `https://www.kaggle.com/dsv/12853160`.

Javier Valero-Elizondo, Jonathan C Hong, Erica S Spatz, Joseph A Salami, Nihar R Desai, Jamal S Rana, Rohan Khera, Salim S Virani, Ron Blankstein, Michael J Blaha, et al. Persistent socioeconomic disparities in cardiovascular risk factors and health in the united states: Medical expenditure panel survey 2002–2013. *Atherosclerosis*, 269:301–305, 2018.

Sergio Venturini, Francesca Dominici, and Giovanni Parmigiani. Gamma shape mixtures for heavy-tailed distributions. *arXiv preprint arXiv:0807.4663*, 2008.

Zachary J Ward, Sara N Bleich, Michael W Long, and Steven L Gortmaker. Association of body mass index with health care expenditures in the united states by age and sex. *PloS one*, 16(3): e0247307, 2021.

Xin Xu, Ellen E Bishop, Sara M Kennedy, Sean A Simpson, and Terry F Pechacek. Annual healthcare spending attributable to cigarette smoking: an update. *American journal of preventive medicine*, 48(3):326–333, 2015.