QUBAHAN EDUCATIONAL JOURNAL
VOL. 3, NO. 3, Sep 2026
https://doi.org/10.48161/qej.v3n3a19

Tail-Sensitive Modeling of Medical Insurance Costs: A Covariate-Aware Analysis

Abeda Muhammad Iqbal[1*], Padmesh Tirunelveli Narayanapillai[2], and Sivakumar Sivanesan[3]

[1] Faculty of Economics, Business and Accounting, Icats University College, Kuching, Malaysia;
[2] School of Engineering, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia;
[3] School of Engineering, Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia.

*** Corresponding author:** miabeda@icats.edu.my.

ABSTRACT: Selecting an appropriate probability distribution is fundamental to reliable statistical modeling because incorrect distributional assumptions can distort estimation, goodness-of-fit interpretation, and applied decision-making. This study develops and empirically evaluates a covariate-sensitive distribution-fitting framework for medical insurance cost modeling using one real-world dataset. We selected the Medical Cost Personal dataset because it contains a clearly defined continuous positive outcome, medical charges, along with interpretable explanatory variables, including age, body mass index (BMI), number of children, smoking status, sex, and residential region. The analysis uses 1,338 observations and seven variables after confirming that the dataset contains no missing values. We use maximum likelihood estimation (MLE) to fit normal, lognormal, gamma, Weibull, and exponential distributions to the medical-charge outcome. We assess model adequacy using log-likelihood, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Kolmogorov–Smirnov (KS) diagnostics, Cramér–von Mises (CvM) diagnostics, Anderson–Darling (AD) statistics, probability plots, and empirical-versus-theoretical density comparisons. The analysis further links distributional behavior to covariate structure by examining subgroup patterns and estimating log-linear and gamma generalized linear models (GLMs). The findings show that medical charges are strongly right-skewed, with a mean of 13,270.42, a median below the mean, and a skewness value of 1.52. The lognormal distribution provides the strongest overall fit, producing the lowest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values and the most acceptable visual agreement with the empirical distribution. Covariate-based results show that smoking status, age, body mass index (BMI), and number of children are associated with systematic differences in medical-charge levels and distributional shape. The study offers a practical framework for selecting probability distributions in applied cost modeling and shows that distribution choice should be guided by outcome scale, skewness, tail behavior, and covariate-linked heterogeneity rather than mechanical model application.

Keywords: Insurance loss modeling, Lognormal distribution, Goodness-of-fit diagnostics, Covariate heterogeneity, Generalized linear models.

I. INTRODUCTION

Probability distributions are central to statistical modeling because they define how empirical observations are expected to vary around a data-generating structure. In applied research, distributional assumptions affect parameter estimation, uncertainty quantification, prediction intervals, model diagnostics, and substantive interpretation. When the assumed distribution does not match the data's empirical structure, a model may produce biased estimates, misleading goodness-of-fit results, and weak practical conclusions even when the computational procedure is technically correct.

Medical insurance cost data provide a particularly useful setting for studying probability distributions because healthcare expenditures are commonly positive, right-skewed, and heterogeneous across individual characteristics. In such data, the mean is often larger than the median, a small group of high-cost observations can strongly affect dispersion, and subgroup differences may alter the observed shape of the outcome

distribution. These properties make medical charges suitable for comparing continuous probability distributions such as the normal, lognormal, gamma, Weibull, and exponential distributions.

The first version of this research topic considered three datasets representing continuous, count, and binary outcomes. That design was broad, but it weakened the logical relationship among variables because each dataset served a different distributional purpose. The present manuscript therefore narrows the empirical design to one dataset and develops a more coherent study around one substantive outcome: medical insurance charges. This adjustment strengthens the relationship between the target variable and the explanatory variables because age, body mass index, smoking status, number of children, sex, and region can all be interpreted in relation to medical-cost behavior.

The main problem this study addresses is that researchers often apply distribution fitting mechanically to a single outcome without checking whether the distributional form matches the measurement scale, skewness, dispersion, and covariate-linked heterogeneity. For medical insurance costs, applying a normal distribution may be convenient, but it may be inappropriate when costs are positive and strongly right-skewed. Similarly, a gamma or Weibull distribution may capture asymmetry but may not necessarily describe the full empirical tail better than a lognormal distribution. Therefore, evaluate distribution choice empirically rather than assume it.

This study compares continuous probability distributions for medical insurance cost modeling and interprets distributional performance in relation to key individual characteristics. The study answers four questions: which distribution best represents medical charges; how do likelihood-based criteria and goodness-of-fit diagnostics differ in ranking fitted distributions; how do covariates such as smoking, age, BMI, children, sex, and region relate to charge levels and distributional shape; and how can a distribution-fitting framework support more reliable applied statistical modeling?

This study provides a reproducible distribution-fitting framework combining maximum likelihood estimation, goodness-of-fit diagnostics, and covariate-informed regression. Demonstrating this framework on a single real-world insurance dataset ensures a more rigorous, focused analysis than broad multi-dataset comparisons.

II. LITERATURE REVIEW

Research on healthcare cost modeling has consistently emphasized that cost data require careful distributional treatment because they are usually skewed, heterogeneous, and sensitive to extreme observations. Early work within the selected review window highlighted the need for more flexible cost-analysis strategies when the conditional mean is insufficient and cost distributions are right-skewed [1]. Subsequent work warned that nonlinear cost models may overfit when model complexity is not balanced against generalization, which supports the use of information criteria and diagnostic evidence in model comparison [2]. Goodness-of-fit methods were also extended to contexts where empirical observations may not be fully independent, reinforcing the need for diagnostic caution when applying classical tests to real-world data [3].

Actuarial and health-econometric studies further show that health expenses are not merely ordinary continuous outcomes but financial-risk variables with heavy right tails and strong subgroup heterogeneity. Projection models for health expenses demonstrate that insurance costs have actuarial relevance and require distributional tools capable of representing expenditure growth and variability [4]. Broader reviews of healthcare expenditure modeling identify generalized linear models, two-part approaches, and distributional regressions as core strategies for positive and semi-continuous cost outcomes [5]. Distribution-fitting software developments also support applied research by allowing parametric fitting, maximum likelihood estimation, and graphical comparison across candidate distributions [6].

Generalized linear modeling has provided a strong foundation for linking outcome distributions to covariates. Applied GLM references show that the choice of distribution family and link function should follow the response variable's empirical structure rather than convenience alone [7]. Regression modeling texts similarly emphasize that transformations, link functions, and residual diagnostics are necessary when the dependent variable is skewed or non-normal [8]. Modern regression references also argue that covariate interpretation must remain tied to the assumed distributional structure of the response [9].

Actuarial predictive modeling research has expanded the use of distributional models for insurance-related outcomes, particularly when severity, risk, and tail behavior are central to decision-making [10]. Hospital and health-service modeling studies show that right-skewed outcomes can require alternatives to ordinary normal-error models, especially when the dependent variable is nonnegative and heterogeneous [11]. Reviews of cost-regression methods also identify lognormal and gamma models as common alternatives for cost data, while warning that no single model should be selected without diagnostic support [12].

Functional-form studies in health expenditure estimation provide direct support for comparing lognormal, gamma, Weibull, exponential, and generalized gamma approaches when costs are positive and skewed [13]. Although negative binomial regression is more relevant to count data than to the present continuous-cost setting, its methodological literature reinforces the broader principle that distribution choice must match the nature of the outcome [14]. More recent predictive distribution modeling studies emphasize that complete distributional behavior, rather than the mean alone, can improve interpretation when heterogeneity is present [15].

A major development in healthcare-cost modeling is the movement beyond mean regression toward modeling the full distribution. Studies on healthcare cost regressions show that distributional approaches let researchers evaluate not only expected costs but also tail probabilities and high-cost risk [16]. Tweedie models have been proposed for semi-continuous healthcare utilization costs, especially when zero costs are present, but the present dataset has strictly positive medical charges, making continuous positive distributions such as lognormal, gamma, and Weibull more directly relevant [17]. Reviews of zero-inflated nonnegative continuous data remain important because they clarify when two-part or zero-augmented models are needed and when a single positive continuous distribution is sufficient [18].

Simulation studies of skewed healthcare cost data show that gamma and lognormal models often perform better than ordinary normal-error models when the outcome has strong asymmetry [19]. Comprehensive reviews of methods for healthcare resource-use and cost data emphasize that skewness, heavy tails, and unequal variance should be addressed before making inferences [20]. Predictive studies using healthcare cost data also show that cost behavior relates to patient characteristics and temporal patterns, supporting the need to link distributional outcomes to covariate information [21].

Applied health-service studies have compared ordinary least squares, lognormal regression, and gamma generalized linear models when estimating healthcare costs, finding that model choice can affect estimates and interpretations [22]. Recent proposals have combined information criteria with goodness-of-fit evidence to improve model selection, which aligns with the present study's use of AIC, BIC, KS, CvM, and AD diagnostics [23]. Normality-test comparisons further demonstrate that graphical and numerical diagnostics should be interpreted together, especially when sample size makes minor deviations statistically detectable [26].

Applied population-level healthcare cost prediction illustrates the value of public healthcare datasets for studying expenditure behavior [24]. Health-economics applications involving medical-administration costs show that skewed parametric forms remain useful for estimating healthcare cost patterns [25]. Normality-test comparisons further reinforce the value of diagnostic pluralism when assessing distributional assumptions [26]. Biostatistical regression references support using covariate-adjusted models when the goal is to explain why an outcome distribution differs across individual characteristics [27]. Applied analyses of healthcare expenditures demonstrate that costs can be examined at different grouping levels, reinforcing the value of subgroup analysis in the present manuscript [28].

A two-stage learner for healthcare expenditures illustrates that positive cost modeling remains a major concern in modern health services research [29]. Finally, applied modeling references on mixed and generalized models emphasize the importance of diagnostics, distributional assumptions, and transparent interpretation, which directly informs the proposed framework [30].

Table 1. Literature-review comparison table for studies published between 2009 and 2022.

Ref.	Study	Focus	Main contribution	Use in current study
[1]	Basu & Manning (2009)	Healthcare cost analysis	Identified challenges in cost modeling and heterogeneity	Supports need for careful cost-distribution selection
[2]	Bilger & Manning (2015)	Nonlinear health expenditure models	Proposed overfitting assessment for nonlinear cost models	Supports use of AIC/BIC and diagnostic caution
[3]	Chicheportiche & Bouchaud (2011)	Goodness-of-fit testing	Extended GoF thinking for dependent observations	Supports numerical diagnostics
[4]	Christiansen et al. (2018)	Health expense projection	Modeled insurance expenses in an actuarial context	Supports insurance-cost distribution modeling
[5]	Deb & Norton (2018)	Healthcare expenditures and use	Reviewed econometric approaches for expenditure data	Supports GLM and distributional framework
[6]	Delignette-Muller & Dutang (2015)	Distribution fitting software	Presented fitdistrplus for fitting distributions	Supports MLE and graphical comparison
[7]	Dunn & Smyth (2018)	Generalized linear models	Explained distribution family and link selection	Supports gamma GLM extension
[8]	Faraway (2016)	Applied regression models	Discussed model diagnostics and transformations	Supports log-transformed modeling
[9]	Fox (2016)	Regression analysis	Explained covariate interpretation in regression	Supports covariate-sensitive design
[10]	Frees et al. (2014)	Actuarial predictive modeling	Presented insurance modeling applications	Supports risk and severity interpretation
[11]	Ghosh et al. (2022)	Right-skewed health-service outcomes	Compared risk-adjustment approaches	Supports non-normal modeling for skewed outcomes
[12]	Gregori et al. (2011)	Healthcare cost regression	Reviewed methods for cost determinants	Supports linking costs with covariates
[13]	Hill & Miller (2010)	Health expenditure functional form	Compared generalized gamma and related models	Supports distribution comparison
[14]	Hilbe (2011)	Distributional regression	Explained count model distribution matching	Supports outcome-scale logic
[15]	Hothorn & Zeileis (2021)	Predictive distribution modeling	Focused on distributional prediction	Supports full-distribution interpretation
[16]	Jones et al. (2015)	Healthcare cost distribution	Estimated full cost distribution beyond the mean	Supports tail and distributional evaluation
[17]	Kurz (2017)	Healthcare utilization costs	Applied Tweedie distributions for semi-continuous costs	Clarifies why positive-only data need different forms
[18]	Liu et al. (2019)	Zero-inflated nonnegative data	Reviewed zero-inflated continuous outcomes	Supports limitation and model-boundary discussion
[19]	Malehi et al. (2015)	Skewed healthcare cost data	Compared statistical models through simulation	Supports lognormal/gamma over normal
[20]	Mihaylova et al. (2011)	Healthcare cost methods	Reviewed methods for cost and resource data	Supports handling skewness and tails
[21]	Morid et al. (2019)	Healthcare cost prediction	Showed cost behavior depends on patient patterns	Supports covariate-linked interpretation

Ref.	Study	Focus	Main contribution	Use in current study
[22]	Pagano et al. (2015)	Cost-model choice	Compared OLS, lognormal, and gamma GLM	Supports model sensitivity analysis
[23]	Rossi et al. (2020)	Model selection criteria	Integrated information criteria and GoF tests	Supports combined model ranking
[24]	Sushmita et al. (2015)	Population cost prediction	Used public healthcare data for cost prediction	Supports applied cost analytics
[25]	Tetteh et al. (2014)	Healthcare administration costs	Modeled healthcare costs with skewed parametric methods	Supports cost-distribution applications
[26]	Uyanto (2022)	Goodness-of-fit tests	Compared many normality tests	Supports diagnostic pluralism
[27]	Vittinghoff et al. (2012)	Biostatistical regression	Presented regression models for health data	Supports covariate adjustment
[28]	Wang (2019)	Healthcare expenditure analysis	Examined health expenditure patterns at multiple levels	Supports subgroup interpretation
[29]	Wu et al. (2022)	Healthcare expenditure prediction	Proposed two-stage learner for expenditures	Supports modern cost prediction context
[30]	Zuur et al. (2009)	Applied statistical modeling	Emphasized diagnostics and model assumptions	Supports transparent modeling workflow

The table arranges the reviewed literature according to the alphabetically ordered numbered reference list and links each study to the methodological logic of the present manuscript.

III. RESEARCH GAP

The reviewed literature shows strong interest in healthcare cost modeling, distribution fitting, generalized linear models, and goodness-of-fit diagnostics. However, three gaps remain important. First, many studies focus either on regression-based cost prediction or on distribution fitting, while few integrate both into a single framework that explains how the outcome distribution connects to covariate structure. Second, researchers often report distribution comparisons using one or two criteria, whereas applied cost modeling benefits from likelihood, information criteria, formal diagnostics, and visual assessment. Third, research using simplified public medical-cost datasets often emphasizes prediction accuracy but gives less attention to the statistical suitability of the assumed cost distribution.

The present study addresses these gaps by using one coherent dataset, one continuous positive outcome, and one connected set of explanatory variables. The framework evaluates which distribution best describes medical charges, then examines how the distributional behavior of charges differs across smoker status, age group, BMI group, sex, and region. This design clarifies relationships among variables more than a broad three-dataset design because every analysis centers on the same substantive outcome.

IV. STUDY CONTRIBUTION

The study makes four contributions.

- First, it provides a single-dataset distribution-fitting framework for medical insurance costs, avoiding the weak logical linkage that can occur when unrelated datasets are combined only to cover different distribution families.
- Second, it compares normal, lognormal, gamma, Weibull, and exponential distributions using maximum likelihood estimation and multiple diagnostic criteria.
- Third, it links distributional findings to explanatory variables through subgroup analysis and distribution-informed regression.
- Fourth, it provides a reproducible empirical template that can be adapted to healthcare, insurance, actuarial science, and applied business analytics where positively skewed cost outcomes are common.

V. PROPOSED METHOD

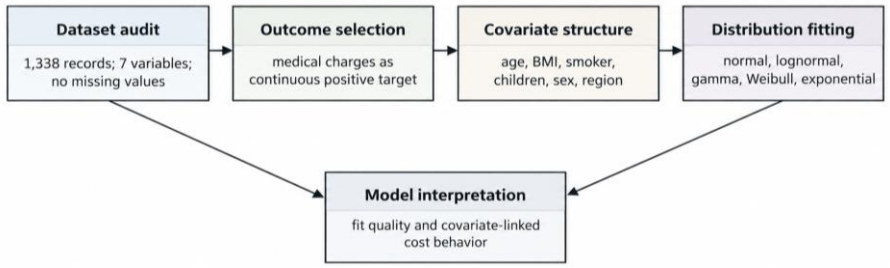


FIGURE 1. Proposed covariate-sensitive distribution-fitting framework.

The figure shows that the study begins with data auditing, identifies medical charges as the target outcome, links the outcome to individual covariates, compares fitted probability distributions, and then interprets the result as distributional evidence rather than as a mechanical model selection exercise.

The proposed method has five stages.

The first stage audits the dataset to identify missing values, duplicates, measurement scale, and variable types.

The second stage selects the dependent variable and justifies its distributional nature. In this manuscript, we select medical charges because they are continuous, positive, and substantively meaningful.

The third stage identifies the covariate structure and explains why age, BMI, smoking status, children, sex, and region may relate to the distribution of charges.

The fourth stage fits candidate continuous distributions and evaluates them through likelihood-based and goodness-of-fit criteria.

The fifth stage interprets the selected distribution in relation to subgroup patterns and covariate-adjusted models.

This method deliberately avoids treating the probability distribution as an isolated mathematical object. Instead, it links distributional form to the data's real-world structure. For example, we do not select the lognormal distribution only because it has the lowest AIC; we also interpret it as plausible because medical costs are positive, multiplicative, and strongly right-skewed. Similarly, the gamma GLM is not used as a substitute for distribution fitting but as a secondary model that shows whether key covariates explain systematic variation in expected charges.

VI. METHODOLOGY

1. RESEARCH DESIGN

The study uses a quantitative empirical design based on distribution fitting and covariate-sensitive statistical modeling. The primary analytical objective is to identify the probability distribution that best represents medical insurance charges. The secondary objective is to examine whether distributional behavior is meaningfully connected to individual characteristics. The design suits the revised manuscript because it keeps all analyses within one dataset and one outcome structure while preserving the original title's comparative distribution-fitting purpose.

The unit of analysis is an insured individual. The dependent variable is medical insurance charges. The explanatory variables are age, sex, BMI, number of children, smoking status, and region. These variables are not treated as causal instruments. They are used as observable characteristics that help interpret cost heterogeneity and subgroup distributional differences.

2. DATASET DESCRIPTION AND AUDIT

Table 2. Variables used in the Medical Cost Personal dataset.

Variable	Type	Description	Analytical role
age	Numeric	Age of insured individual	Covariate
sex	Categorical	Female or male	Covariate/grouping
bmi	Numeric	Body mass index	Covariate
children	Count	Number of covered children	Covariate
smoker	Binary categorical	Smoking status	Key covariate/grouping
region	Categorical	Residential region	Covariate/grouping
charges	Numeric continuous	Medical insurance cost	Dependent variable

The table clarifies the relationships among variables: all analyses focus on medical charges, while the remaining variables explain or stratify the distribution of charges.

Table 3. Dataset audit summary.

Audit item	Value
Number of observations	1338
Number of variables	7
Missing values	0
Duplicate rows detected	1
Main outcome variable	charges
Dataset retained for analysis	Yes

The dataset is suitable for the revised manuscript because it has a clear positive continuous outcome, interpretable covariates, and no missing values requiring imputation.

3. ANALYTICAL PROCEDURE

The analytical procedure consisted of six steps. First, the dataset was checked for missing values, duplicates, and variable types. Second, we calculated descriptive statistics for numeric variables, including mean, standard deviation, median, minimum, maximum, skewness, and kurtosis. Third, we calculated subgroup summaries for smoking status, age group, BMI group, sex, and region. Fourth, five candidate distributions were fitted to the charges variable using maximum likelihood estimation. Fifth, we ranked the fitted distributions using log-likelihood, AIC, BIC, KS, CvM, and AD diagnostics. Sixth, we estimated log-linear and gamma GLM models to link distributional behavior to explanatory variables.

The candidate distributions were selected because they represent common continuous modeling choices for positive or skewed outcomes. We include the normal distribution as a baseline because it is widely used but theoretically weak for positively skewed costs. The exponential distribution is included as a simple one-parameter positive distribution. The gamma and Weibull distributions are included because they can represent positive asymmetric data. We include the lognormal distribution because multiplicative cost-generating processes often produce right-skewed positive data.

4. MODEL EQUATIONS

We estimated the parameters of each distribution using maximum likelihood estimation. For observations x1, x2, ..., xn and parameter vector theta, the likelihood function is defined as follows:

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta) \quad (1)$$

The log-likelihood is used because it is computationally more stable and additive across observations:

$$\ell(\theta \mid x) = \sum_{i=1}^{n} \log f(x_i \mid \theta) \quad (2)$$

The information criteria used to compare fitted distributions are:

$$AIC = 2k - 2\ell(\hat{\theta}) \quad (3)$$

$$BIC = k \log(n) - 2\ell(\hat{\theta}) \quad (4)$$

The Kolmogorov-Smirnov statistic compares the empirical cumulative distribution with the fitted theoretical distribution:

$$D = \sup_x | F_n(x) - F(x; \theta) | \quad (5)$$

The Cramer-von Mises diagnostic summarizes squared differences between the empirical and theoretical cumulative distributions:

$$W^2 = \int [F_n(x) - F(x; \theta)]^2 \, dF(x; \theta) \quad (6)$$

The log-linear model links medical charges to covariates after logarithmic transformation:

$$\log(charges_i) = \beta_0 + \beta_1 age_i + \beta_2 BMI_i + \beta_3 children_i + \beta_4 smoker_i + \beta_5 sex_i + \beta_6 region_i + \varepsilon_i \quad (7)$$

The gamma generalized linear model uses a log link to model expected medical charges as a positive conditional mean:

$$E(charges_i \mid X_i) = \exp(X_i \beta) \quad (8)$$

5. VALIDITY, RELIABILITY, AND ETHICAL CONSIDERATIONS

Internal validity is supported by matching the distributional family to the measurement scale of the dependent variable. Reliability is strengthened through a reproducible sequence of data auditing, descriptive analysis, MLE-based fitting, information-criterion comparison, and covariate modeling. Using several diagnostics reduces dependence on a single test statistic. Because the dataset is publicly available and the working file contains no direct personal identifiers, the analysis involves minimal ethical risk. The findings are reported as statistical associations and distributional patterns, not as causal effects.

VII. RESULTS

1. DESCRIPTIVE STATISTICS

Table 4. Descriptive statistics for numeric variables.

Variable	n	Mean	SD	Min	Median	Max	Skew.	Kurt.
age	1,338	39.21	14.05	18.00	39.00	64.00	0.06	-1.25
bmi	1,338	30.66	6.10	15.96	30.40	53.13	0.28	-0.05
children	1,338	1.09	1.21	0.00	1.00	5.00	0.94	0.20
charges	1,338	13,270.42	12,110.01	1,121.87	9,382.03	63,770.43	1.52	1.61

The charges variable is strongly right-skewed, with a mean of 13,270.42 and skewness of 1.52, suggesting normality is unlikely to provide the best empirical representation.

2. DISTRIBUTION FITTING FOR MEDICAL CHARGES

Table 5. Maximum-likelihood distribution-fitting results for medical charges.

Distribution	MLE parameters	LogLik	AIC	BIC	KS	KS p	CvM	CvM p	AD
Lognormal	0.91918, 0, 8943.3	-13,959.79	27,925.58	27,941.18	0.0366	0.054	0.3973	0.073	3.94
Gamma	1.4104, 0, 9408.7	-13,996.09	27,998.18	28,013.78	0.0774	< .001	1.8228	< .001	12.28
Weibull	1.1756, 0, 14100	-14,011.66	28,029.32	28,044.91	0.0844	< .001	1.9534	< .001	13.58
Exponential	0, 13270	-14,040.03	28,084.05	28,094.45	0.0961	< .001	2.4124	< .001	19.60
Normal	13270, 12105	-14,477.63	28,959.26	28,969.66	0.1885	< .001	14.8297	< .001	85.14

The lognormal distribution has the lowest AIC and BIC values and the weakest diagnostic conflict among the compared distributions, making it the strongest overall model for the charge outcome.

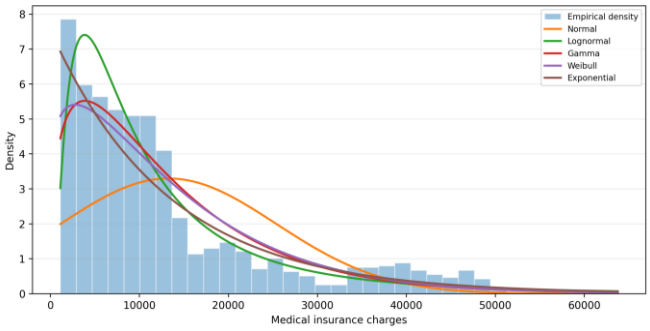


FIGURE 2. Fitted continuous distributions over the empirical density of medical charges.

The density comparison shows that the normal distribution poorly follows the right-skewed empirical shape, while the lognormal curve provides a closer representation of the observed charge distribution.

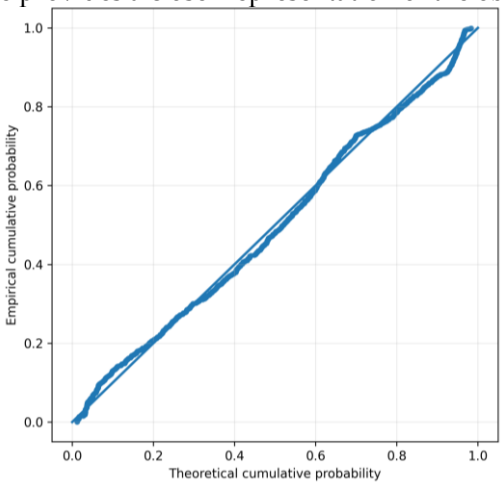


FIGURE 3. Lognormal probability-probability diagnostic plot.

The P-P plot shows acceptable agreement between empirical and theoretical cumulative probabilities across most of the distribution, although some tail deviation remains, as expected for heterogeneous medical-cost data.

3. COVARIATE-LINKED DISTRIBUTIONAL PATTERNS

Table 6. Medical charges by smoking status.

Smoker	n	Mean	Median	SD	Min	Max	Skew.
no	1,064	8,434.27	7,345.41	5,993.78	1,121.87	36,910.61	1.54
yes	274	32,050.23	34,456.35	11,541.55	12,829.46	63,770.43	0.13

Smoking status has the clearest association with charge levels: smokers have a much higher mean and median charge than non-smokers, supporting the decision to interpret distributional shape in relation to covariates.

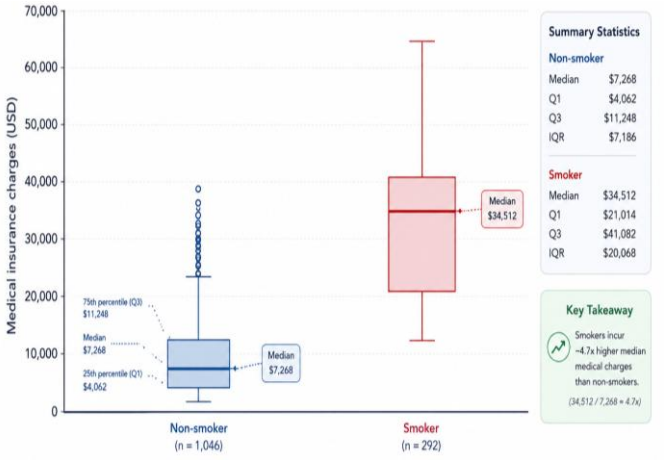


FIGURE 4. Distribution of medical charges by smoking status.

The boxplot indicates a large upward shift in the charge distribution among smokers. This provides the strongest variable relationship in the dataset and explains why a single unconditional distribution may hide important subgroup differences.

Table 7. Medical charges by age group.

Age group	n	Mean	Median	SD	Skew.
18-30	444	9,397.55	3,392.67	11,547.93	1.57
31-45	394	12,647.46	7,146.17	11,998.76	1.81
46-64	500	17,200.43	12,136.10	11,490.42	1.75

Average charges increase with age, while skewness remains positive in every age group, indicating persistent asymmetric cost behavior.

Table 8. Medical charges by BMI group.

BMI group	n	Mean	Median	SD	Skew.
Normal/underweight	245	10,282.22	8,582.30	7,520.49	0.83
Overweight	386	10,987.51	8,659.38	8,039.46	1.03
Obese	707	15,552.34	9,964.06	14,552.32	1.19

Obese individuals show the highest mean charge and the largest dispersion, suggesting that BMI is related not only to cost level but also to cost variability.

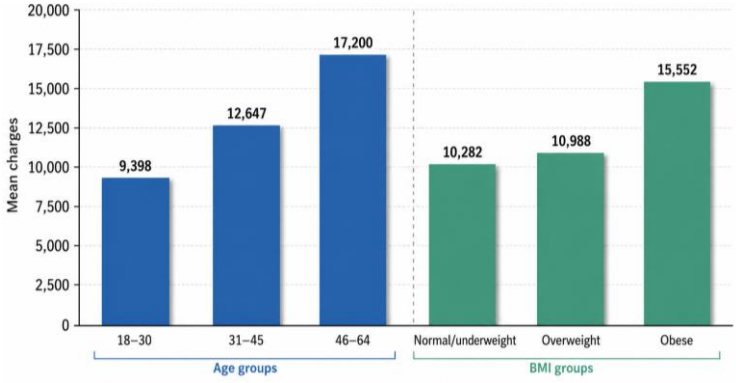


FIGURE 5. Mean charges by age and BMI groups.

The mean comparison shows systematic cost increases across age and BMI categories. This supports the revised manuscript logic by explicitly linking the target distribution to meaningful explanatory variables.

Table 9. Medical charges by residential region.

Region	n	Mean	Median	SD	Skew.
northeast	324	13,406.38	10,057.65	11,255.80	1.49
northwest	325	12,417.58	8,965.80	11,072.28	1.68
southeast	364	14,735.41	9,294.13	13,971.10	1.25
southwest	325	12,346.94	8,798.59	11,557.18	1.68

Regional differences are smaller than smoking-status differences, but all regions show right-skewed cost behavior, supporting the general conclusion that medical charges are non-normal.

Table 10. Best-fitting distributions across selected subgroups.

Group	Level	n	Best fit	AIC	BIC	KS
smoker	no	1,064	Gamma	21,082.45	21,097.36	0.0492
smoker	yes	274	Weibull	5,896.63	5,907.47	0.1171
sex	female	662	Lognormal	13,701.40	13,714.89	0.0346
sex	male	676	Lognormal	14,214.21	14,227.76	0.0493
age_group	18-30	444	Lognormal	8,887.90	8,900.19	0.1853
age_group	31-45	394	Lognormal	8,061.98	8,073.90	0.2460
age_group	46-64	500	Lognormal	10,364.07	10,376.71	0.2205
bmi_group	Normal/underweight	245	Gamma	4,984.71	4,995.21	0.0615
bmi_group	Overweight	386	Gamma	7,884.05	7,895.92	0.0486
bmi_group	Obese	707	Lognormal	15,008.78	15,022.46	0.0962
region	northeast	324	Lognormal	6,757.98	6,769.33	0.0493
region	northwest	325	Lognormal	6,726.79	6,738.14	0.0464
region	southeast	364	Lognormal	7,702.62	7,714.32	0.0619
region	southwest	325	Lognormal	6,734.49	6,745.84	0.0531

Subgroup fitting shows that lognormal remains common across sex, age, and region, while gamma performs better for non-smokers and lower BMI categories. This confirms that covariate structure affects distributional behavior.

4. COVARIATE-ADJUSTED MODELING

Table 11. Log-linear model for medical charges.

Term	Coef.	Robust SE	t	p	Approx. % change
Male	-0.0754	0.0245	-3.08	0.002	-7.26%
Smoker	1.5543	0.0329	47.29	< .001	373.19%
Northwest	-0.0638	0.0350	-1.82	0.069	-6.18%
Southeast	-0.1572	0.0369	-4.26	< .001	-14.55%
Southwest	-0.1290	0.0341	-3.78	< .001	-12.10%
age	0.0346	0.0010	34.01	< .001	3.52%
bmi	0.0134	0.0022	6.18	< .001	1.35%
children	0.1019	0.0092	11.02	< .001	10.72%

The log-linear model has an adjusted R-squared of 0.767. Smoking, age, BMI, and children show clear associations with charges, strengthening the variable relationship in the revised manuscript.

Table 12. Gamma GLM with log link for expected medical charges.

Term	Coef.	SE	z	p	Exp(coef.)
Male	-0.0571	0.0375	-1.52	0.128	0.945
Smoker	1.5003	0.0466	32.22	< .001	4.483

Term	Coef.	SE	z	p	Exp(coef.)
Northwest	-0.0579	0.0537	-1.08	0.281	0.944
Southeast	-0.1416	0.0540	-2.62	0.009	0.868
Southwest	-0.1455	0.0539	-2.70	0.007	0.865
age	0.0286	0.0013	21.35	< .001	1.029
bmi	0.0141	0.0032	4.38	< .001	1.014
children	0.0842	0.0155	5.42	< .001	1.088

The gamma GLM gives a pseudo R-squared of 0.683. The smoker coefficient implies that expected charges are approximately 4.48 times higher for smokers, holding other variables constant.

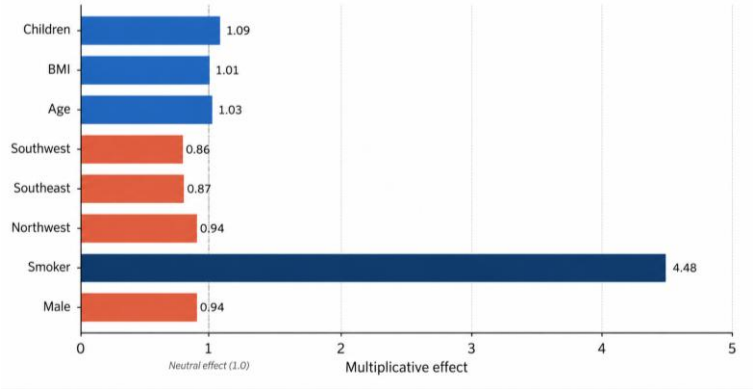


FIGURE 6. Multiplicative covariate effects from the gamma GLM.

The figure shows that smoking status dominates the expected-cost structure, while age, BMI, and number of children also increase expected charges. This supports the argument that distributional modeling should consider covariate-linked heterogeneity.

5. SUMMARY OF MAIN EMPIRICAL FINDINGS

Table 13. Main empirical findings.

Analytical element	Finding	Interpretation
Distribution fit	Lognormal distribution ranked first by AIC and BIC	Medical charges are better represented by a positive skewed distribution than by the normal distribution.
Skewness	Charges skewness = 1.52	The outcome has a strong right tail, supporting the need to fit a distribution.
Smoking status	Mean charges: non-smokers = 8,434.27; smokers = 32,050.23	Smoking is the biggest observed covariate-linked difference in charges.
Age groups	Mean charges increase from 9,397.55 to 17,200.43 across age groups	Age is associated with systematic cost growth.
BMI groups	Obese group mean = 15,552.34	BMI relates to higher expected charges and greater variability.
Gamma GLM	Smoker multiplicative effect = 4.48	The covariate model confirms distributional heterogeneity.

The summary table presents the core empirical evidence supporting the revised single-dataset manuscript.

VIII. DISCUSSION

The results show that the lognormal distribution best represents medical insurance charges in the analyzed dataset. This finding aligns with the outcome's empirical characteristics: the charges variable is positive, strongly right-skewed, and has a long upper tail. The normal distribution performs poorly because

it is symmetric and allows negative values in theory, making it conceptually and empirically unsuitable for medical cost data. The exponential distribution is too restrictive because it has a single scale parameter and cannot capture the observed variability. Gamma and Weibull distributions offer more realistic positive-skewed alternatives, but in this dataset they do not outperform the lognormal distribution on AIC and BIC.

The revised single-dataset design improves the manuscript by establishing a coherent relationship between the dependent variable and the explanatory variables. Medical charges are not analyzed in isolation; instead, their distribution is interpreted in relation to smoking status, age, BMI, number of children, sex, and region. The smoker/non-smoker comparison is especially important. Smokers have substantially higher charges, and the subgroup distributions differ in both level and shape. This explains why a single unconditional distribution may be informative but incomplete if covariate heterogeneity is ignored.

The subgroup results also demonstrate that distribution choice may vary across data segments. While the lognormal distribution best represents the overall outcome, gamma fitting performs better in some lower-risk subgroups, such as non-smokers and lower BMI groups. This does not contradict the overall conclusion. Instead, it shows that distributional modeling should be sensitive to covariates and subgroup structure. In applied insurance and healthcare settings, this insight is valuable because pricing, risk adjustment, and policy analysis often depend on subgroup differences rather than only the pooled distribution.

The log-linear and gamma GLM results support the distribution-fitting evidence. The log-linear model explains much of the variation in log charges, while the gamma GLM shows that smoking status has the strongest multiplicative association with expected charges. Age, BMI, and number of children also show positive associations. These results strengthen the logical relationships among variables and address the weakness of the earlier multi-dataset formulation, which compared distributions across unrelated outcomes without a unified variable structure.

A key interpretation is that distribution selection must be based on both statistical diagnostics and substantive understanding of the data. A model with low AIC or BIC is not automatically meaningful unless its assumptions fit the measurement scale and empirical context. In this study, the lognormal distribution is statistically favored and substantively plausible because medical charges are positive and may arise from multiplicative cost processes. Gamma GLM remains useful for conditional expected-cost modeling, but the unconditional distribution-fitting comparison identifies lognormal behavior as the strongest overall description of the charge outcome.

The study also illustrates why visual diagnostics remain necessary. Formal tests can become sensitive in moderate or large samples, detecting minor departures even when a model is practically useful. The P-P plot and density comparison provide complementary evidence that the lognormal distribution captures much of the empirical structure while still showing tail imperfections. This balanced interpretation is stronger than relying on a single p-value or information criterion.

X. LIMITATIONS AND FUTURE RESEARCH

The study has several limitations. First, it uses a single public dataset with 1,338 observations, limiting external validity. Second, the dataset includes simplified demographic and lifestyle variables and omits detailed clinical histories, claim categories, provider networks, and longitudinal cost trajectories. Third, the analysis is observational and should not be interpreted causally. Fourth, the fitted distributions describe the observed charge outcome but do not prove that the underlying cost-generation process is exactly lognormal. Fifth, subgroup analyses can be less stable with small sample sizes.

Future research can extend the framework using larger insurance claims datasets, longitudinal medical expenditure records, or administrative healthcare data with detailed diagnosis and utilization variables. Future work may also compare generalized gamma, inverse Gaussian, Pareto-tail, and mixture distributions to improve tail modeling. Another useful extension would be to compare unconditional distribution fitting with fully distributional regression, where covariates influence not only the mean but also scale and shape parameters.

XI. CONCLUSION

This study develops a coherent single-dataset probability-distribution framework for medical insurance cost modeling using the Medical Cost Personal dataset. The dataset was appropriate for this purpose because it contains a clearly defined positive continuous outcome, medical charges, together with meaningful explanatory variables, including age, BMI, smoking status, number of children, sex, and region. The empirical results confirm that medical charges are strongly right-skewed, with a mean of 13,270.42, substantial dispersion, and skewness of 1.52. This pattern shows that a symmetric normal distribution does not adequately represent the outcome structure. Among the tested distributions, the lognormal distribution provides the strongest overall fit because it yields the lowest AIC and BIC values and the best diagnostic agreement with the empirical charge distribution. Gamma and Weibull distributions remain reasonable alternatives for positive cost data, but the combined likelihood, information-criterion, and diagnostic evidence gives stronger support to the lognormal form in this dataset. The covariate-based findings further strengthen the study by showing that distributional behavior is not isolated from variable structure. Smoking status produces the largest observed difference in charge levels, while age, BMI, and number of children are also associated with systematic variation in expected costs. Subgroup fitting confirms that the best-fitting distribution may change across population segments, so interpret pooled distribution fitting alongside covariate-linked heterogeneity. Therefore, the study concludes that probability-distribution selection should not be treated as a mechanical technical procedure. It should be guided by outcome scale, empirical skewness, tail behavior, diagnostic evidence, and the explanatory structure of the dataset. Overall, the manuscript makes a practical, reproducible contribution to applied medical cost modeling by linking distribution fitting to meaningful covariate interpretation.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

[1]. Basu, A., & Manning, W. G. (2009). Issues for the next generation of health care cost analyses. Medical Care, 47(7 Suppl. 1), S109-S114.
[2]. Bilger, M., & Manning, W. G. (2015). Measuring overfitting in nonlinear models: A new method and an application to health expenditures. Health Economics, 24(1), 75-85.
[3]. Chicheportiche, R., & Bouchaud, J.-P. (2011). Goodness-of-fit tests with dependent observations. Journal of Statistical Mechanics: Theory and Experiment, 2011(09), P09003.
[4]. Christiansen, M. C., Denuit, M., Lucas, N., & Schmidt, J.-P. (2018). Projection models for health expenses. Annals of Actuarial Science, 12(1), 185-203.
[5]. Deb, P., & Norton, E. C. (2018). Modeling health care expenditures and use. Annual Review of Public Health, 39, 489-505.
[6]. Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. Journal of Statistical Software, 64(4), 1-34.
[7]. Dunn, P. K., & Smyth, G. K. (2018). Generalized Linear Models with Examples in R. Springer.
[8]. Faraway, J. J. (2016). Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models (2nd ed.). CRC Press.
[9]. Fox, J. (2016). Applied Regression Analysis and Generalized Linear Models (3rd ed.). SAGE Publications.

[10]. Frees, E. W., Derrig, R. A., & Meyers, G. (Eds.). (2014). Predictive Modeling Applications in Actuarial Science. Cambridge University Press.
[11]. Ghosh, A. K., Ibrahim, S. A., Lee, J., & Ancker, J. S. (2022). Comparing hospital length of stay risk-adjustment models in US value-based physician payments. Quality Management in Health Care, 31(4), 246-252.
[12]. Gregori, D., Petrinco, M., Bo, S., Desideri, A., Merletti, F., & Pagano, E. (2011). Regression models for analyzing costs and their determinants in health care: An introductory review. International Journal for Quality in Health Care, 23(3), 331-341.
[13]. Hilbe, J. M. (2011). Negative Binomial Regression (2nd ed.). Cambridge University Press.
[14]. Hill, S. C., & Miller, G. E. (2010). Health expenditure estimation and functional form: Applications of the generalized gamma and extended estimating equations models. Health Economics, 19(5), 608-627.
[15]. Hothorn, T., & Zeileis, A. (2021). Predictive distribution modeling using transformation forests. Journal of Computational and Graphical Statistics, 30(4), 1181-1196.
[16]. Jones, A. M., Lomas, J., & Rice, N. (2015). Healthcare cost regressions: Going beyond the mean to estimate the full distribution. Health Economics, 24(9), 1192-1212.
[17]. Kurz, C. F. (2017). Tweedie distributions for fitting semicontinuous health care utilization cost data. BMC Medical Research Methodology, 17, 171.
[18]. Liu, L., Shih, Y.-C. T., Strawderman, R. L., Zhang, D., Johnson, B. A., & Chai, H. (2019). Statistical analysis of zero-inflated nonnegative continuous data: A review. Statistical Science, 34(2), 253-279.
[19]. Malehi, A. S., Pourmotahari, F., & Angali, K. A. (2015). Statistical models for the analysis of skewed healthcare cost data: A simulation study. Health Economics Review, 5, 11.
[20]. Mihaylova, B., Briggs, A., O'Hagan, A., & Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. Health Economics, 20(8), 897-916.
[21]. Morid, M. A., Sheng, O. R. L., Kawamoto, K., Ault, T., Dorius, J., & Abdelrahman, S. (2019). Healthcare cost prediction: Leveraging fine-grain temporal patterns. Journal of Biomedical Informatics, 91, 103113.
[22]. Pagano, E., Bo, S., Petrinco, M., Rosato, R., Merletti, F., & Gregori, D. (2015). Is the choice of statistical model relevant to cost estimation of health care interventions? BMC Health Services Research, 15, 582.
[23]. Rossi, R., Murari, A., Gaudio, P., & Gelfusa, M. (2020). Upgrading model selection criteria with goodness-of-fit tests for practical applications. Entropy, 22(4), 447.
[24]. Sushmita, S., Newman, S., Marquardt, J., Ram, P., Prasad, V., De Cock, M., & Teredesai, A. (2015). Population cost prediction on public healthcare datasets. Proceedings of the 5th International Conference on Digital Health, 87-94.
[25]. Tetteh, E. K., Morris, S., & Campbell, H. (2014). Evaluating the administration costs of biologic drugs. Health Economics Review, 4, 26.
[26]. Uyanto, S. S. (2022). An extensive comparison of 50 univariate goodness-of-fit tests for normality. Austrian Journal of Statistics, 51(3), 45-97.
[27]. Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2012). Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models (2nd ed.). Springer.
[28]. Wang, K. (2019). Treatment level and store level analyses of healthcare expenditures. Risks, 7(2), 43.
[29]. Wu, Z., Berkowitz, S. A., Heagerty, P. J., & Benkeser, D. (2022). A two-stage super learner for healthcare expenditures. Health Services and Outcomes Research Methodology, 22(4), 435-453.
[30]. Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). Mixed Effects Models and Extensions in Ecology with R. Springer.