Model Selection and Averaging of Health Costs in Episode Treatment Groups

Shujuan Huang[1]
Liberty Mutual

Brian Hartman[2]
Brigham Young University

Vytaras Brazauskas[3]
University of Wisconsin-Milwaukee

To appear in *ASTIN Bulletin*
(*Submitted*: January 27, 2015 *Revised*: January 11, 2016; July 21, 2016 *Accepted*: August 7, 2016)

Abstract. Episode Treatment Groups (ETGs) classify related services into medically relevant and distinct units describing an episode of care. Proper model selection for those ETG based costs is essential to adequately price and manage health insurance risks. The optimal claim cost model (or model probabilities) can vary depending on the disease. We compare four potential models (lognormal, gamma, log-skew-t, and Lomax) using four different model selection methods (AIC and BIC weights, Random Forest feature classification, and Bayesian model averaging) on 320 episode treatment groups. Using the data from a major health insurer, which consists of more than 33 million observations from 9 million claimants, we compare the various methods on both speed and precision, and also examine the wide range of selected models for the different ETGs. Several case studies are provided for illustration. It is found that Random Forest feature selection is computationally efficient and sufficiently accurate, hence being preferred in this large data set. When feasible (on smaller data sets), Bayesian model averaging is preferred because of the posterior model probabilities.

Keywords and phrases: Akaike Weights; Bayesian Model Selection; Model Averaging; Random Forest.

---

[1] Shujuan Huang, Ph.D., is a Senior Analyst in the Advanced Analytics Group at Liberty Mutual, Seattle, WA 98154. *e-mail:* hshujuan@gmail.com Much of this work was completed while the author was a Ph.D. student in the Department of Mathematics at the University of Connecticut.

[2] Corresponding author: Brian Hartman, Ph.D., ASA, is an Assistant Professor and Actuarial Program Director in the Department of Statistics, Brigham Young University, Provo, UT 84602. *e-mail*: hartman@stat.byu.edu Much of this work was completed while the author was an Assistant Professor in the Department of Mathematics at the University of Connecticut.

[3] Vytaras Brazauskas, Ph.D., ASA, is a Professor and Actuarial Science Program Co-Director in the Department of Mathematical Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI 53201. *e-mail*: vytaras@uwm.edu

1 Introduction

Predictive analytics in healthcare has been gaining popularity as the power of data are more extensively harnessed and revealed in practice. Through the development of lab and diagnostic tests, our healthcare related data has multiplied to the point that we have terabytes of information to be evaluated. From the health plan insurer's point of view, predictive modeling can dramatically help with cost control, pricing, reserving, risk management and marketing. Duncan (2011) comprehensively reviewed healthcare risk adjustment and predictive modeling using models for predicting health costs such as the generalized linear model, tree-based models and artificial neural networks with applications in Medicaid/Medicare risk adjustment and other areas. Dove *et al.* (2003) described the development and validation of a predictive model designed to identify and target HMO members who are likely to incur high costs. Frees *et al.* (2011) model total health expenditures through multiple events using two-part models.

Symmetry Episode Treatment Groups (ETGs) were introduced and patented by OPTUM as an episode grouper for medical and pharmacy claims. They combine related services into a distinct medically relevant unit describing a complete episode of care, thus applying to diverse groups such as healthcare providers, researchers and administrators. ETGs have been used to look at the quality of care and efficiency of outcomes for specific illnesses (see, for example, Leary *et al.*, 1997, and Forthman *et al.*, 2000, 2005, 2010). Health insurers are interested in better understanding the potential future costs of their book of business. With ETGs, we can see how much each patient spent on any disease in a year. Then we can incorporate information about the disease profile of the book of business going forward to better estimate future claim costs. Symmetry ETGs are currently used by more than 300 healthcare plans and their providers in the United States and similar groupings are used globally.

When modeling the annual costs for a single ETG across the book of business, the choice of model is important. The statistical cost distribution of lower cost, more common diseases will have a very different shape than that of rare, high cost diseases. As a set of possibles models, we chose four different distributions (gamma, lognormal, Lomax, and log-skew-t) with varying tail thickness and skewness. In this paper, we compare three different methods to find the optimal model for each ETG. We explore the relationship between speed and accuracy among the methods. In Section 2, we describe

our dataset. In Section 3, we describe the four candidate distributions and the three different model selection techniques to choose among them. In Section 4, we perform a simulation study to compare how well the three techniques work. In Section 5, we apply the methods to our actual claims data. In Section 6, we conclude and give suggestions for implementation.

2 Data

We are using ETG cost data from a major national health insurer. It has 33 million sample observations from 9 million claimants. Each row in our dataset contains the total cost to the insurer from July 2011 through June 2012 on claims associated with a given ETG. For example, imagine policyholder John had both iron deficiency anemia (ETG 2082) and personality disorder (ETG 2394) between July 2011 and June 2012. He did not use his health insurance for any other reason than to treat those two diseases, but cost the insurance company $1450 treating his anemia and another $2500 treating his personality disorder. His rows in our dataset would be:

TABLE 2.1: Example data rows for John.

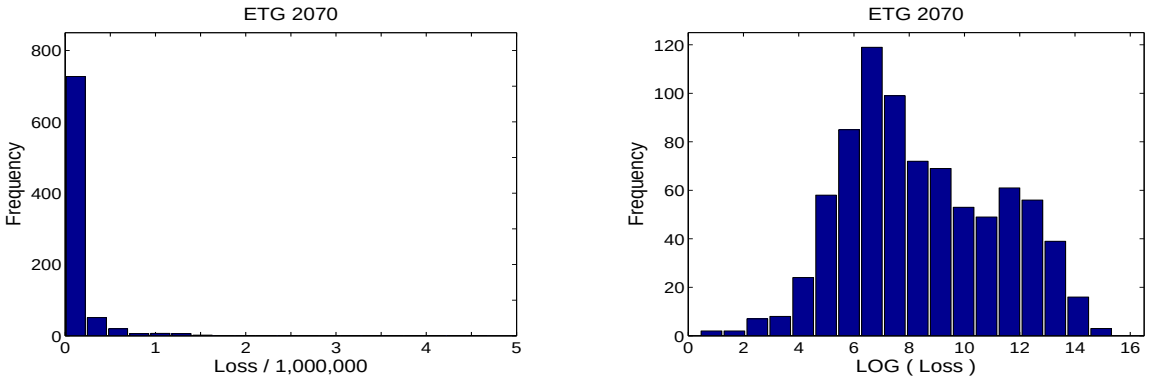
Policy ID	ETG	Cost
123	2082	1450
123	2394	2500

For those policyholders without claim cost on certain ETGs, there is no zero record for them in the data set. There are 347 ETGs in all, including 320 non-routine ETGs, such as AIDS, hemophilia, and personality disorder. We only consider those non-routine ETGs in this paper because the routine ETGs (e.g., physical exams, standard checkups, immunizations) are rather consistent from year to year. They are not worth the effort to model. Basic summary statistics for a range of non-routine ETGs are shown in Table 2.2 for illustration. The ETGs were chosen to exemplify the differences in shape and scale among the costs for the different diseases.

TABLE 2.2: Dictionary and summary statistics for selected non-routine ETGs.

ETG Code	*Number of Policies with Non-Zero Costs*	*ETG Description*	*Mean*	*Standard Deviation*
1301	13,534	AIDS	15,570	25,246
1635	2,679	Hyper-functioning adrenal gland	2,035	8,963
1640	1,162	Hypo-functioning parathyroid gland	1,704	6,314
2068	16,554	Agranulocytosis	4,677	17,923
2070	822	Hemophilia	94,343	303,552
2080	944	Anemia of chronic diseases	2,434	10,943
2082	49,409	Iron deficiency anemia	1,772	5,208
2394	1,550	Personality disorder	1,718	5,263
3868	42,401	Congestive heart failure	10,870	56,777
4370	50	Lung transplant	461,226	338,683
4744	4,162	Trauma of stomach or esophagus	6,562	10,994
7112	1,668	Juvenile rheumatoid arthritis	7,193	27,441

The histograms of these costs both on the original and log scale give insight into the skewness and tail thickness of the data. Although the ETGs show similar shape with a heavy tail and right skewness on the original scale, the histograms for those costs on the log scale vary among different ETGs. The histograms for three selected ETGs are shown in Figure 2.1. The costs for each ETG vary greatly in both shape and scale.



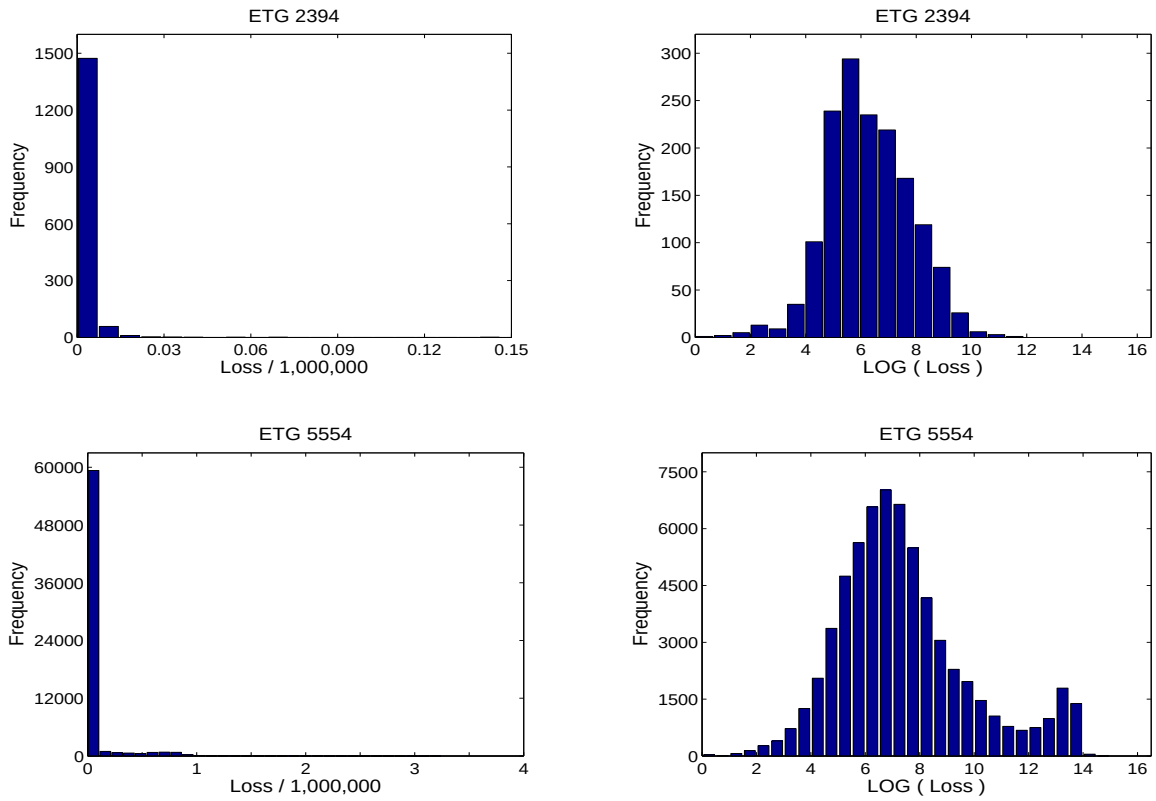


FIGURE 2.1: Histograms of cost (left panel) and *log*-cost (right panel) for three ETGs: 2070, 2394, 5554.

3 Methods

Our methods take a set of candidate distributions, and determine which distribution (or weighted average of distributions) best fits a set of data. In our application, we chose a set of plausible candidate distributions based on our exploration of the data and recognizing that our claim cost data is constrained to be positive. We chose to consider the lognormal, gamma, Lomax, and log-skew-t distributions in this paper. Lognormal and gamma distributions are widely used in numerous fields (see, e.g., Kleiber and Kotz, 2003). The Lomax distribution is essentially a Pareto distribution that has been shifted so that its support begins at zero (see, Klugman *et al.*, 2012). The log-skew-t distribution is a continuous probability distribution of a random variable whose logarithm is skew-t distributed. The skew-t distribution generalizes the t distribution to allow for non-zero skewness. The skew-t distribution is extensively investigated as a promising candidate for both theoretical and empirical work in actuarial science (see, e.g., Ferreira and Steel, 2007; Jones and Faddy, 2003; Eling, 2012). The

density functions of the four distributions have different shapes and tail thicknesses, but all have been used in business, economics, and actuarial modeling.

For each ETG, we will determine the optimal distribution using three different model selection techniques: AIC and BIC weights, parallel model selection, and random forest feature classification.

3.1 AIC and BIC Weights

AIC and BIC are measures of model fit. In both cases, the statistic is penalized by the complexity of the model so that the chosen model is only as complex as necessary. When comparing multiple potential distributions, the AIC (or BIC) can be computed for each distribution and the one with the smallest AIC is chosen as optimal. However, many studies, such as Shtatland *et al* (2000), show that choosing a single optimal model from AIC or BIC can be computationally expensive (depending on the likelihood to maximize) and have other disadvantages including, as Kuha (2004) shows, that while both AIC and BIC are generally good approximations of their own theoretical target quantities, they can still fail in some very simple examples. Rather than simply looking for an AIC-optimal or BIC-optimal model, we calculate the AIC and BIC weights. These weights can be easily calculated from the raw AIC/BIC values, and provide an approximation as the probabilities of each model being the best model in an AIC or BIC sense. Burnham and Anderson (2002) presented a way to approximate the probability that a chosen distribution is optimal. These approximations are known as AIC (or BIC) weights and are computed as follows,

$$w_i^{\text{AIC}} = \frac{\exp(-\Delta_i^{\text{AIC}}/2)}{\sum_{k=1}^{K}\exp(-\Delta_k^{\text{AIC}}/2)} \quad \text{with} \quad \Delta_i^{\text{AIC}} = \text{AIC}_i - \min\left\{\text{AIC}_1, \ldots, \text{AIC}_K\right\},$$

$$w_i^{\text{BIC}} = \frac{\exp(-\Delta_i^{\text{BIC}}/2)}{\sum_{k=1}^{K}\exp(-\Delta_k^{\text{BIC}}/2)} \quad \text{with} \quad \Delta_i^{\text{BIC}} = \text{BIC}_i - \min\left\{\text{BIC}_1, \ldots, \text{BIC}_K\right\},$$

where K denotes the number of candidate models.

3.2 Bayesian Inference and Parallel Model Selection

Parallel model selection (Congdon, 2006) is a Bayesian method which estimates the posterior probabilities of each distribution being the best, enabling model averaging and providing deeper insights into the relationships between the distributions. The uncertainty in the model-selection process can also be explicitly assessed. We selected the priors of the parameters in the various candidate models to

be semi-informative. Using fully non-informative priors overly penalizes complex distributions (those with a large number of parameters). The priors are defined in Table 3.2.

TABLE 3.2: Prior distribution settings.

Candidate Model (*Parameters*)	*Prior Distributions*	*Number of Thinned Samples Per Chain*	*Number of Burn-in Samples Per Chain*
lognormal (μ, τ)	$\mu \sim \text{normal}\,(6,\ 5)$ $\tau \sim \text{gamma}\,(4,\ 4.5)$	30,000	20,000
gamma (τ, ν)	$\tau \sim \text{gamma}\,(2,\ 3)$ $\nu \,\vert\, \omega \sim \text{exponential}\,(\omega)$ $\omega \sim \text{uniform}\,(0.01,\ 10)$	50,000	35,000
log-skew-t $(\alpha, \xi, \nu, \Omega)$	$\alpha \sim \text{normal}\,(50,\ 4)$ $\xi \,\vert\, \theta \sim \text{normal}\,(\theta,\ 7)$ $\nu \sim \text{exponential}\,(0.25)$ $\Omega \sim \text{inverse gamma}\,(6,\ 1)$ $\theta \sim \text{normal}\,(0,\ 5)$	300,000	260,000
Lomax (λ, α)	$\lambda \sim \text{gamma}\,(2,\ 3)$ $\alpha \,\vert\, \omega \sim \text{exponential}\,(\omega)$ $\omega \sim \text{uniform}\,(0.01,\ 10)$	300,000	20,000

We used the LaplaceDemon package in R to perform parallel MCMC algorithms. Several algorithms were tried and compared, such as Hit-and-Run Metropolis (Chen and Schmeiser, 1993), No-U-Turn Sampler (Hoffman and Gelman, 2014; Bai, 2009), and Hamiltonian Monte Carlo (Neal, 2011). We ran three chains in most cases, each in parallel, where a sequence $x_1, x_2, \ldots$ of random elements of some set is a Markov chain if the conditional distribution of x_{n+1} given $x_1, \ldots, x_n$ depends on x_n only. The three MCMC chains initialized with different starting values. The other two important settings are burn-in size and thinned sample size. Burn-in sample size refers to the number of samples discarded from the initial portion of a Markov chain so that the effect of initial values on the posterior inference is minimized. Thinning is used to reduce sample autocorrelation by keeping every k^{th} simulated draw from the sequence. In our application, we thinned our samples to every 10^{th} observation. This reduced the autocorrelation to an acceptable level, less than 0.1.

3.3 Random Forest Feature Classification

Computational speed is always an issue in large-scale analytics. Therefore, it is desirable to find a faster approach for large data sets which does not sacrifice too much accuracy. To this end, we can

think about the model selection process in a new way. We have four (in our case) possible groups that each set of ETG observations can be classified into, those best described by each distribution. Then the model selection problem becomes a classification problem where the explanatory variables are the features of the data and the response variable is the chosen distribution. Random forests can select a model for ETG-based groups of losses by constructing many decision trees during the training phase and allowing the trees to each choose an optimal model. The model selected is then the model selected by the most trees. For more information about random forests, see Breiman (2001) or Hastie *et al.* (2009).

To classify the datasets, we need to find some set of features to compare. Using all the observations will ensure that the maximum amount of information is used, but is also the most computationally expensive. We have experimented with two different sets of potential features, moment-based and percentile-based.

- Moment-based characteristics (e.g., mean, standard deviation, coefficient of variation, skewness, and kurtosis) for raw data and the same measures for log-data.
- Percentile-based characteristics (e.g., 10th, 25th, 50th, 75th, 90th percentiles, median absolute deviation, and interquartile range) for raw data and the same measures for log-data.

We then compare the features of the individual ETG data to the features of data simulated from the candidate distribution. For example, we simulate many datasets from a gamma distribution with a range of reasonable parameters. We then calculate the summary statistics for each of those datasets. Since we know those statistics came from a gamma distribution, we use them to train the random forests. The random forest then looks at our data and decides which known distribution the data most resembles.

To determine which set of features we are going to use in our model, we compared the out-of-bag error rate (similar to leave-one-out cross validation) when we use each set of summary statistics (and both). The results are presented in Table 3.4.

TABLE 3.4: Performance of moment-based features versus percentile-based features in the simulated data.

Candidate Models Used	*Feature Selection*	*Out-of-Bag Error Rate*
lognormal, gamma, Lomax	Moment-based features only	0.25%
	Percentile-based features only	1.00%
	Both types of features	0.08%
lognormal, gamma, Lomax, log-skew-t	Moment-based features only	3.53%
	Percentile-based features only	13.63%
	Both types of features	2.01%

The performance of RF also depends on the difficulty of the tasks. If the set of possible distributions have obvious distinguishable features (the lognormal, gamma and Lomax distributions are very similar), RF would recognize that and the misclassification rate would be very low. But if the distributions are quite similar, then it is more difficult to distinguish the models. The more candidate distributions with similar characteristics, the worse the random forest performs.

Table 3.5 shows the RF classification results on simulated in-sample data and Table 3.6 shows the results on the simulated but out-of-sample data.

TABLE 3.5: Random forest classification results on *in-sample* data.

Candidate Models Used	*Number of Trees in Random Forest*	*Number of Vars Used at Each Split*	*Out-of-Bag Error Rate*
lognormal, gamma, Lomax, log-skew-t	4,000	6	0.25%
lognormal, gamma, Lomax	4,000	6	0.00%

TABLE 3.6: Random forest classification results on *out-of-sample* data.

Candidate Models Used	*Misclassification Rate*
lognormal, gamma, Lomax, log-skew-t	23.8%
lognormal, gamma, Lomax	1.2%

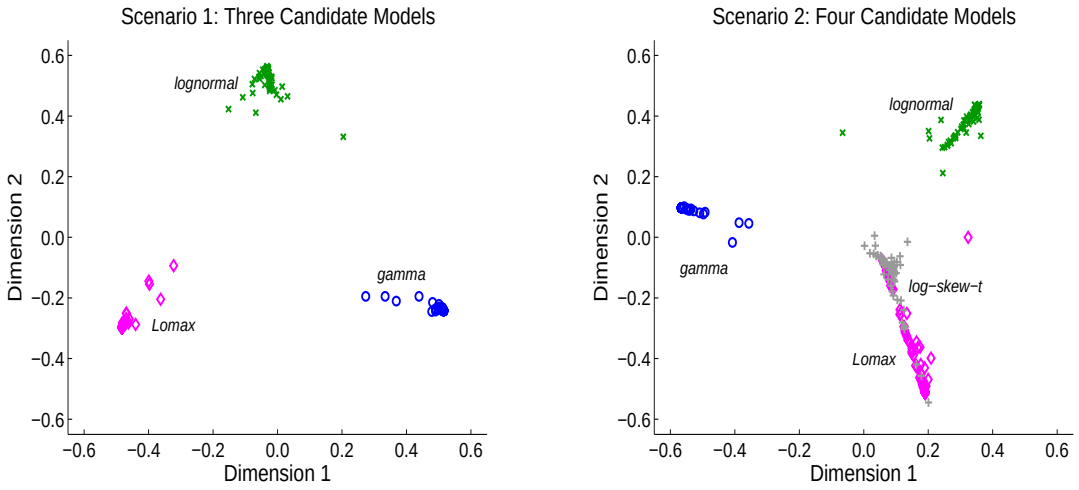


FIGURE 3.1: Multidimensional scaling plots of proximity matrix for two scenarios.

Multidimensional scaling is an ordination technique to visualize the level of similarity between individual cases in a data set. It aims to place each object in n-dimensional space such that the between-object distances are preserved as far as possible. In Figure 3.1, the statistical features of each data set are represented by a point in a two dimensional space. The points are arranged in this space so that the distances between pairs of points relates to the similarities among the pairs of objects. That is, two similar objects are represented by two points that are close together, and two dissimilar objects are represented by two points that are far apart. Tables 3.5 and 3.6 tell us that if only three distributions (gamma, lognormal, Lomax) are considered, they are easily distinguishable. When the log-skew-t distribution is added to the mix, the plot shows that it is very similar to the Lomax distribution. Thus, it is clear that the most difficult task is classification between all four distributions (lognormal, gamma, log-skew-t, Lomax) because the points from different distributions cannot be easily distinguished.

4 Simulation Study

To further explore the differences between the three model selection techniques, we set up a simulation study. (For the simulation study and the analyses in the next section we used R.) First, we use the MLE approach to fit the four distributions on the same real ETG data. And then we use these MLE-fitted models to simulate four random samples with 600 observations each that follows one of the

lognormal, gamma, log-skew-t, and Lomax distributions. Then, we apply the three model selection methodologies (AIC weights, RF, Bayesian) to the simulated data sets and check how accurately each approach identifies the true model. Our findings are summarized in Table 4.1.

TABLE 4.1: Model selection accuracy: AIC weights, Random Forest, Bayesian.

Model Selection Methodology	*Selected Distribution*	*Distribution Used to Simulate Data*			
		lognormal	gamma	log-skew-t	Lomax
AIC weights	lognormal	75.81%	0.00%	24.19%	0.00%
	gamma	0.00%	94.42%	5.58%	0.00%
	log-skew-t	0.00%	0.00%	93.91%	6.09%
	Lomax	0.00%	0.00%	27.81%	72.19%
Random Forest	lognormal	99.70%	0.00%	0.10%	0.20%
	gamma	11.30%	62.75%	15.00%	10.95%
	log-skew-t	0.08%	0.03%	67.58%	32.33%
	Lomax	0.03%	0.00%	43.98%	56.00%
Bayesian	lognormal	100.00%	0.00%	0.00%	0.00%
	gamma	1.90%	93.90%	3.14%	1.06%
	log-skew-t	0.00%	0.00%	100.00%	0.00%
	Lomax	0.23%	0.00%	38.54%	61.23%

In each 4×4 matrix in Table 4.1, if the probabilities on the diagonals are close to 100%, the method accurately selects the true model. From the results, we can observe and compare level of the model uncertainty and prediction power over different methods. Though the most computationally intense of the three methods, on an average sense, Bayesian performs best because it exactly identifies lognormal and log-skew-t distributions and it is slightly less certain about gamma and Lomax compared to AIC weights. AIC weights did a good job on average. Random Forest performs slightly more poorly than the other two, but it still can almost surely identify the model with the best fit. Especially when we need to deal with big data sets, its efficiency is valuable without losing much accuracy.

5 Results

The weights w_i^{AIC} are known as *AIC weights* or *Akaike weights*. Similarly, the weights w_i^{BIC} are called the *BIC weights*. For illustrative purposes, the AIC values and Akaike weights on four models for selected ETGs (see Table 2.2) are provided in Table 5.1.

TABLE 5.1: Akaike weights and AIC values for the four candidate models and selected ETGs.

ETG	*Akaike weights*				*AIC values*			
Code	lognormal	gamma	log-skew-t	Lomax	lognormal	gamma	log-skew-t	Lomax
1301	0.000	0.000	1.000	0.000	288,909	289,613	286,796	287,556
1635	0.000	0.000	1.000	0.000	44,022	46,907	43,765	43,808
1640	0.000	0.000	1.000	0.000	18,640	19,920	18,567	18,617
2068	0.000	0.000	1.000	0.000	286,108	299,983	285,891	285,954
2070	0.882	0.000	0.118	0.000	17,897	18,309	17,901	17,930
2080	0.000	0.000	0.998	0.002	14,755	15,835	14,684	14,697
2082	0.000	0.000	1.000	0.000	725,294	760,699	724,756	726,749
2394	0.000	0.000	1.000	0.000	25,175	26,374	25,144	25,182
3144	0.001	0.000	0.990	0.009	328	344	315	324
3169	0.000	0.000	1.000	0.000	2,508,992	2,606,074	2,508,562	2,511,985
3868	0.000	0.000	1.000	0.000	797,694	837,377	797,623	799,257
4370	0.002	0.087	0.816	0.095	1,416	1,408	1,404	1,408
4744	0.000	0.000	1.000	0.000	80,732	81,476	80,539	80,580
7112	0.000	0.000	0.973	0.027	30,786	32,166	30,766	30,773

For those randomly selected ETGs, the distributions for some ETGs such as ETG-1301 and ETG-3868 are immediately apparent. The log-skew-t distribution is also dominant for ETG-2080 and ETG-3144. It indicates that AIC values and Akaike weights have a strong preference for the log-skew-t distribution for most of these data sets. However, there are exceptions. For ETG-2070, the probability spreads between two distributions: 0.882 probability to lognormal model and 0.118 probability to the log-skew-t. And for ETG-4370, the probability spreads among all four distributions: 0.002 probability for the lognormal model, 0.087 probability for the gamma distribution, 0.816 probability for the log-skew-t distribution and 0.095 probability for the Lomax distribution.

Next, we applied parallel model selection to the same randomly selected ETGs; the posterior model probabilities are given in Table 5.2. The optimal distributions for some ETGs such as hemophilia, AIDS, and agranulocytosis are immediately apparent. The lognormal distribution is also dominant for lung transplants and many others. For personality disorder, the probability spreads between two distributions: 0.783 probability to lognormal model and 0.217 probability to the log-skew-t.

In addition to improved understanding of the data, these probabilities can be used for model averaging. When one model is dramatically better than the others, only knowing the best model will be sufficient. When the potential models are very similar in their fit for some data sets, a simulation

should account for that model uncertainty by drawing a proportion of the simulations from each of the models that fit the data well. For example, to simulate future ETG cost streams for personality disorder, 78.3% samples can be drawn from lognormal distribution, and 21.7% of the samples drawn from log-skew-t. Under the standard methods, the proper model proportions are unknown.

TABLE 5.2: Posterior model probabilities using parallel model selection for selected ETGs.

ETG Code	*ETG Description*	lognormal	gamma	log-skew-t	Lomax
1301	AIDS	0.000	0.000	1.000	0.000
1635	Hyper-functioning adrenal gland	0.000	0.000	1.000	0.000
1640	Hypo-functioning parathyroid gland	0.000	0.000	1.000	0.000
2068	Agranulocytosis	0.000	0.000	1.000	0.000
2070	Hemophilia	1.000	0.000	0.000	0.000
2080	Anemia of chronic diseases	0.000	0.000	1.000	0.000
2082	Iron deficiency anemia	0.000	0.000	1.000	0.000
2394	Personality disorder	0.783	0.000	0.217	0.000
3868	Congestive heart failure	0.450	0.000	0.550	0.000
4370	Lung transplant	0.999	0.000	0.001	0.000
4744	Trauma of stomach or esophagus	0.000	0.000	1.000	0.000
7112	Juvenile rheumatoid arthritis	0.999	0.000	0.001	0.000

The three methodologies vary widely in computational burden. For our entire dataset, the random forest methodology was very fast (2 minutes), the AIC/BIC weights were somewhat slower (4 hours), and the parallel model averaging was the slowest (4 weeks). Please note that all these model selections had to be made in sequence because we were limited to the laptop for computation. If the MCMC chains were run in parallel, the computational time would likely be reduced by about an order of 300 (4 weeks reduced to a few hours).

Now we explore how consistent the RF and AIC methodologies are in selecting the same model (for all 320 ETGs). First, in Table 5.3, we only use three distributions (lognormal, gamma, Lomax) as candidates for model selection. Those three distributions have obvious distinguishable features. In the 3×3 matrix, RF and AIC agree on all the 197 ETGs model selections on the diagonal. For some ETGs, compared to RF, AIC prefers lognormal to Lomax.

TABLE 5.3: Comparison of model assignments by RF and AIC for all 320 ETGs (when *three* candidate models used).

Distribution Selected by RF	*Distribution Selected by AIC* lognormal	gamma	Lomax	*RF Total*
lognormal	100	11	19	130
gamma	1	5	3	9
Lomax	87	2	92	181
AIC Total	188	18	114	320

Next, in Table 5.4, we use four distributions (lognormal, gamma, Lomax, log-skew-t). AIC has an apparent preference for the log-skew-t distribution because it selects this model for 292 of 320 ETGs. Random forest also selects the log-skew-t distribution for most ETGs, but at the same time it assigns 131 ETGs to lognormal distribution. One common theme is that none of the methods select the gamma distribution for any ETG. That is understandable because compared to other distributions, gamma is relatively light tailed. Given the heavy tails for most ETG costs, once the log-skew-t distribution is one of the candidates, no method selected the gamma distribution as the best model.

TABLE 5.4: Comparison of model assignments by RF and AIC for all 320 ETGs (when *four* candidate models used).

Distribution Selected by RF	*Distribution Selected by AIC* lognormal	gamma	Lomax	log-skew-t	*RF Total*
lognormal	23	0	2	106	131
gamma	0	0	0	0	0
Lomax	1	0	1	25	27
log-skew-t	0	0	1	161	162
AIC Total	24	0	4	292	320

6 Conclusions

Predictive modeling has grown to be a powerful tool in healthcare in terms of cost control, pricing, reserving, marketing and risk management. ETGs (Episode Treatment Groups) were introduced for identifying and classifying an entire episode of care for evidence-based medicine and healthcare management reporting. In spite of ETGs wide use, how to effectively use ETGs for health plan risk management is still an outstanding and interesting issue from the insurers point of view. This research

aims to investigate the application of ETGs in health plan risk management, with a focus on model selection for those ETG-based costs. In this paper, we compared four potential models: lognormal, gamma, log-skew-t, and Lomax; where gamma is the default distribution for positive continuous explanatory variables in practice. None of the methods select the gamma distribution as the best model for any of the 320 different ETGs. Thus, one needs to be cautious when using a gamma model for heavy-tailed data.

In addition to model selection and averaging, this paper also contributes by recommending various model selection techniques for different data sizes and goals of the analyst. The four techniques considered in this paper are AIC weights, BIC weights, Bayesian parallel model selection and Random Forest feature classification. AIC and BIC are commonly used maximum likelihood driven information criteria, and try to balance good fit with parsimony. BIC generally penalizes free parameters more strongly than AIC, but in our experiments their results are quite similar. Parallel model selection can yield us the probabilities of each model being the best given the data among all models under consideration, enabling model averaging and providing deeper insights into the relationships between the models. Since we have 33 million ETG cost observations from 9 million claimants, we proposed random forest feature classification in order to achieve greater computational efficiency. Since the random forest model selection is based on summary statistics rather than the original big data sets, computing time is significantly reduced. Our results show that random forest only takes 2 minutes for the whole dataset, but AIC/BIC needs around 4 hours. Parallel model selection may need approximately 4 weeks with our computing constraints. With better computing resources, especially the ability to run more processes in parallel, can reduce the computing time for parallel model selection to a few hours. Furthermore, we compared the accuracy among the four methods. On average, the parallel model selection approach performs best because it exactly identifies lognormal and log-skew-t distribution, though is less certain about gamma and Lomax compared to AIC weights. AIC weights also did a good job on average. Random Forest performs a little bit worse than the other two, but it still can generally identify the model with the best fit. Especially when we need to deal with big data, its efficiency is valuable without losing much accuracy.

When looking to implement these methods, please note that model averaging has as one of its special cases using a single model to fit a dataset (i.e., all of the probability mass ends up with a single

model). Because our methods are finding the best model to fit (and predict from) our data, if a single model is best it will be selected. We actually saw that occur in a good number of our ETGs. Model averaging is only a superior approach when it beats using a single model, and it has to compete every time. There is also a non-zero structural cost to using a more complicated model. The modelers will have to be educated on the method and IT staff will have to implement it. The size of this cost varies greatly depending on the company and the application. Because of that variance, we chose to exclude that consideration from our model.

As a part of our future work, we plan to investigate the possible dependence among ETGs, and try to incorporate ETGs into risk assessment regression framework, as well as disease specific product design and pricing.

Acknowledgements

The authors are very appreciative of valuable insights and comments provided by four anonymous referees, which helped to improve the paper. This research is supported by a major national health insurer. The authors are grateful to the company for granting them the opportunity to work together on this paper and getting access to the data. Also, the third author gratefully acknowledges the hospitality of the Department of Mathematics at the University of Connecticut, where all three authors met for two weeks in the fall of 2013 and started working on this project.

References

[1] Akaike, H. (1978). On the likelihood of a time series model. *The Statistician*, **27**, 217–235.

[2] Bai, Y. (2009). Convergence of Adaptive Markov Chain Monte Carlo Methods. *Ph.D. dissertation*, Department of Statistics, University of Toronto.

[3] Breiman, L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32.

[4] Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

[5] Carlin, B.P. and Chib, S. (1995). Bayesian model choice via Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, **57**(3), 473–484.

[6] Chen, M.H. and Schmeiser, B. (1993). Performance of the Gibbs, Hit-and-Run, and Metropolis samplers. *Journal of Computational & Graphical Statistics*, **2**(3), 251–272.

[7] Congdon, P. (2006). Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Computational Statistics & Data Analysis*, **50**(2), 346–357.

[8] Dove, H.G., Duncan, I., and Robb, A. (2003). A prediction model for targeting low-cost, high-risk members of managed care organizations. *The American Journal of Managed Care*, **9**(5), 381–389.

[9] Duncan, I. (2011). *Healthcare Risk Adjustment and Predictive Modeling*. ACTEX Publications.

[10] Eling, M. (2012). Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? *Insurance: Mathematics & Economics*, **51**(2), 239–248.

[11] Ferreira, J. and Steel, M.F. (2007). A new class of skewed multivariate distributions with applications to regression analysis. *Statistica Sinica*, **17**(2), 505–529.

[12] Forthman, M.T., Dove, H.G., Wooster, L.D. (2000). Episode Treatment Groups (ETGs): A patient classification system for measuring outcomes performance by episode of illness. *Topics in Health Information Management*, **21**(2), 51–61.

[13] Forthman, M.T., Dove, H.G., Forthman, C.L., Henderson, R.D. (2005). Beyond severity of illness: Evaluating differences in patient intensity and complexity for valid assessment of medical practice pattern variation. *Managed Care Quarterly*, **13**(4), 9–17.

[14] Forthman, M.T., Gold, R.S., Dove, H.G., Henderson, R.D. (2010). Risk-adjusted indices for measuring the quality of inpatient care. *Quality Management in Health Care*, **19**(3), 265–277.

[15] Frees, E.W., Gao, J., and Rosenberg, M.A. (2011). Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal*, **15**(3), 377–392.

[16] Hartman, B.M. and Groendyke, C. (2013). Model selection and averaging in financial risk management. *North American Actuarial Journal*, **17**(3), 216–228.

[17] Hastie, T.J., Tibshirani, R.J., and Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

[18] Hoffman, M.D. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**(1), 1351–1381.

[19] Jones, M. and Faddy, M. (2003). A skew extension of the t-distribution, with applications. *Journal of the Royal Statistical Society: Series B*, **65**(1), 159–174.

[20] Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley.

[21] Klugman, S.A., Panjer, H.H., and Willmot, G.E. (2012). *Loss Models: From Data to Decisions*, 4th edition. Wiley.

[22] Kuha, J. (2004). AIC and BIC comparisons of assumptions and performance. *Sociological Methods & Research*, **33**(2), 188–229.

[23] Leary, R.S., Johantgen, M.E., Farley, D., Forthman, M.T., Wooster, L.D. (1997). All-payer severity-adjusted diagnosis-related groups: A uniform method to severity-adjust discharge data. *Topics in Health Information Management*, **17**(3), 60–71.

[24] Liaw, A. and Wiener, M. (2002). Classification and Regression by Random Forest. *R news*, **2**(3), 18–22.

[25] Neal, R. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G.L. Jones, X-L. Meng, eds.), 113–162. Chapman & Hall/CRC.

[26] Schwartz, E.M., Bradlow, E.T., and Fader, P.S. (2014). Model selection using database characteristics: Developing a classification tree for longitudinal incidence data. *Marketing Science*, **33**(2), 188–205.

[27] Shtatland, E.S., Moore, S., Dashevsky, I., Miroshnik, I., Cain, E., and Barton, M.B. (2000). How to be a Bayesian in SAS: Model selection uncertainty in PROC LOGISTIC and PROC GENMOD. In *Proceedings of the 13th Annual NorthEast SAS Users Group Conference*, 1–9.