SOCIETY OF ACTUARIES

Article from:

North American Actuarial Journal

January 2008 – Vol.12 No.1

Predictive Modeling of Costs for a Chronic Disease with Acute High-Cost Episodes

Marjorie A. Rosenberg* and Philip M. Farrell†

Abstract

Chronic diseases account for 75% of U.S. national health care expenditures as estimated by the Centers for Disease Control. Many chronic diseases are punctuated by acute episodes of illnesses that occur randomly and create cost spikes in utilization from one year to the next. Modeling to account for these random events provides better estimates of (1) future costs and (2) their variability.

A Bayesian statistical model is used to predict the incidence and cost of hospitalizations for one chronic disease. A two-part statistical model is described that separately models the utilization and cost of hospitalization. Individual demographic characteristics are included as well as a simple biological classification system to adjust for the severity of disease among individuals.

Results by child, as well as by calendar year, are presented. Using a simple approach to incorporate severity, the model produces reasonable estimates of the number of hospitalizations and cost of hospitalization for the group in total, as well as for a separate group of High Utilizers.

The study reflects real-world experiences of persons entering and leaving a group. Modeling at an individual level provides a way to adjust for individual-level severity. The ability to model uneven and unpredictable occurrence of utilization, and potential cost, would be beneficial in the design of insurance programs or for disease management programs.

1. Introduction

The most prevalent chronic diseases in the United States are heart disease, cancer, and diabetes mellitus (Centers for Disease Control and Prevention 2004). The predicted aging of the U.S. population will contribute to a rise in the incidence of chronic diseases (Bonow et al. 2002). Approximately 34% of the U.S. population (70,000,000 people) have one or more types of cardiovascular disease, which constitute 38% of all deaths (American Heart Association 2005). Many chronic diseases are associated with acute episodes of illnesses that require extensive therapeutic interventions. Thus, although a stable pattern of the incidence of costs over time may appear for a large population in the aggregate, closer analysis of the costs at the individual level may reveal a different pattern. For instance, there could be some level of baseline costs incurred, with frequent occurrence of acute care episodes. The ability to forecast these potentially uneven and random occurrences of utilization, and their cost, would be beneficial in the design of insurance programs or for disease management programs.

The purpose of this article is to demonstrate the usefulness of a predictive model that estimates the incidence and costs of hospitalization over time for a heterogeneous group of individuals with a chronic disease. Predictive modeling involves the use of data to forecast future events (Rosenberg et al. 2007). Application areas of predictive modeling in health care include underwriting and pricing of groups, stratifying risks for clinical management, and program evaluation of clinical and financial outcomes (Rosenberg and Johnson 2007).

* Marjorie A. Rosenberg, PhD, FSA, is Associate Professor in the Wisconsin School of Business and the School of Medicine and Public Health at the University of Wisconsin–Madison. mrosenberg@bus.wisc.edu.

† Phillip M. Farrell, MD, PhD, is Professor in the School of Medicine and Public Health at the University of Wisconsin–Madison. pmfarrell@wisc.edu.

The number of hospitalizations vary by individual by year, where the observed number of hospitalizations are random variables from an underlying probability distribution. Similarly the cost per hospitalization is also a random variable from some underlying probability distribution. Both the number and cost random variables have probability distributions that are long-tailed, with large values having a positive, although small, probability of occurring. To model these quantities for individuals with chronic diseases requires a way to include the individual-level severity of the disease, as well as other individual-level covariates such as age and sex. Using individual-level data allows the model to reflect differences in numbers of hospitalizations and cost per hospitalization by individuals.

Actuaries generally model health care costs using trend factors. The advantage of using trend factors is the simplicity of their application to predict the next year's aggregate claims. The disadvantage of this method is that it does not incorporate the movement of individuals in and out of a group; depending on their level of utilization this movement would impact the projections.

A Bayesian two-part model is used to predict the frequency of hospitalization by year of age in the first part, and conditional on usage, the cost per hospitalization in the second part. Two-part models are useful for health care data that often feature a large proportion of zeros as well as incorporate long-tailed distributions in both the first and second parts. Actuaries commonly use two-part models as in Bowers et al. (1997, Chap. 2, p. 28).

The next section provides a short description of the data and the model, and demonstrates how the predictions are calculated. The results depict the viability of the statistical model, and the article concludes with a discussion of the findings and other application areas.

2. METHODS

2.1 Data

The study uses hospital utilization and cost data for children with cystic fibrosis (CF), a genetically inherited disease that affects the pulmonary and gastrointestinal systems and nutritional status of the patient. Individuals with CF have potential compromises in their digestive, respiratory, and endocrine systems (Marshall 2004; Fitzsimmons 1995). Because of the variability of symptoms among patients, the disease can be difficult to diagnose, and treatment is often delayed unless newborn screening programs exist (Fost and Farrell 1989; Farrell and Mischler 1992; Farrell et al. 2000). Hospitalizations are a major cost factor for children with CF. Silber, Gleeson, and Zhao (1999) showed that children with major organ disease (like CF) showed a 54% increase in length of hospital stays and 79% increase in hospital charges as compared with children without chronic disease.

The data were collected from a randomized clinical trial in Wisconsin. The overall purpose of the trial was to address the hypothesis that early diagnosis of cystic fibrosis through neonatal screening would be medically beneficial without major risks (Farrell and Mischler 1992). Wisconsin babies with CF born from April 1985 to June 1994 were screened for CF at birth and then randomized to either a control group or a screened group. Newborns with CF in the control group were diagnosed through traditional means, such as signs and symptoms, family history, or when the results from the newborn screening were disclosed after the child reached four years of age in accordance with an ethically sound design and avoidance of selection bias (Fost and Farrell 1989).

Electronic data for 77 children were available from July 1989 to June 2003 leading to both truncated and censored observations. Truncated observations arose as the data collection period began after the birth of some children who may have incurred costs before the data were available. Censored observations occurred when the data collection phase ended and children were still alive and at risk or when children discontinued participation in the study.

Table 1 summarizes the data for 1990–2002, those calendar years with a full year of data.[1] The number of children, exposure, number of hospitalizations, and cost of hospitalizations are shown for each calendar year. In 1990 there were 45.5 children-years of exposure, growing to a maximum of 67.6 in 1994. Exposure declined in

[1] All years given are calendar years unless otherwise noted.

Table 1
Number of Children, Exposure, Number of Hospitalizations, and Facility Cost for 1990–2002 in Total, High Utilizers, and Remainder of Children

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
No. children													
All	46	48	57	66	69	65	61	59	59	57	55	55	53
High Utilizers	8	8	8	8	8	7	7	7	7	7	6	6	5
Remainder	38	40	49	58	61	58	54	52	52	50	49	49	48
Exposure													
All	45.5	47.1	56.6	65.3	67.6	62.8	60.0	59.0	57.9	56.3	55.0	54.4	53.0
High Utilizers	8.0	8.0	8.0	8.0	7.5	7.0	7.0	7.0	7.0	6.5	6.0	5.8	5.0
Remainder	37.5	39.1	48.6	57.3	60.1	55.8	53.0	52.0	50.9	49.8	49.0	48.6	48.0
No. hospitalizations													
All	37	40	22	19	21	14	23	15	15	11	6	9	12
High Utilizers	26	31	12	11	10	4	9	7	10	9	4	4	7
Remainder	11	9	10	8	11	10	14	8	5	2	2	5	5
Cost of hospitalizations													
All	$572,107	$443,747	$209,944	$178,336	$209,029	$74,975	$233,279	$129,896	$182,461	$133,178	$84,254	$164,587	$230,930
High Utilizers	406,171	371,737	88,720	82,136	130,592	21,382	77,277	77,880	147,077	125,561	70,287	75,510	87,625
Remainder	165,936	72,010	121,224	96,200	78,437	53,593	156,002	52,016	35,384	7,617	13,967	89,077	143,305

1995–2002 from 62.8 to 53.0. Table 1 reflects a total of 244 hospitalizations over the 13 years. A subset of eight children, labeled ''High Utilizers,'' had a higher number of hospitalizations during the study period with a total of 144 hospitalizations, or 59% of the hospitalizations overall. One child had 38 hospitalizations. The rate of hospitalization per year was consistently higher for the High Utilizers as compared to the remainder of the children.

Figure 1 shows the distribution of the rate of hospitalization per year per child, which is adjusted for the exposure of each child in the study. The *y*-axis shows the number of children, and the *x*-axis shows the rate per year. The right-skewed distribution shows a point mass at zero reflecting the 30 children who had no hospitalizations during the study period. The dark shaded boxes are the High Utilizers, who are generally shown at the high end of the utilization.

All costs were adjusted to 2001 dollars by adjusting costs on a monthly basis using the medical care component of the Consumer Price Index (http://data.bls.gov). For this study we used cost related to the care inside the hospital such as nursing, pharmacy, and laboratory, which was collected from a detailed cost-accounting process at the University of Wisconsin–Madison Hospitals and Clinics. Table 1 shows the cost of the hospitalizations for 1990–2002 for the group, High Utilizers, and the Remainder. Similar to the numbers of hospitalization, the High Utilizers had 61% of the total cost.

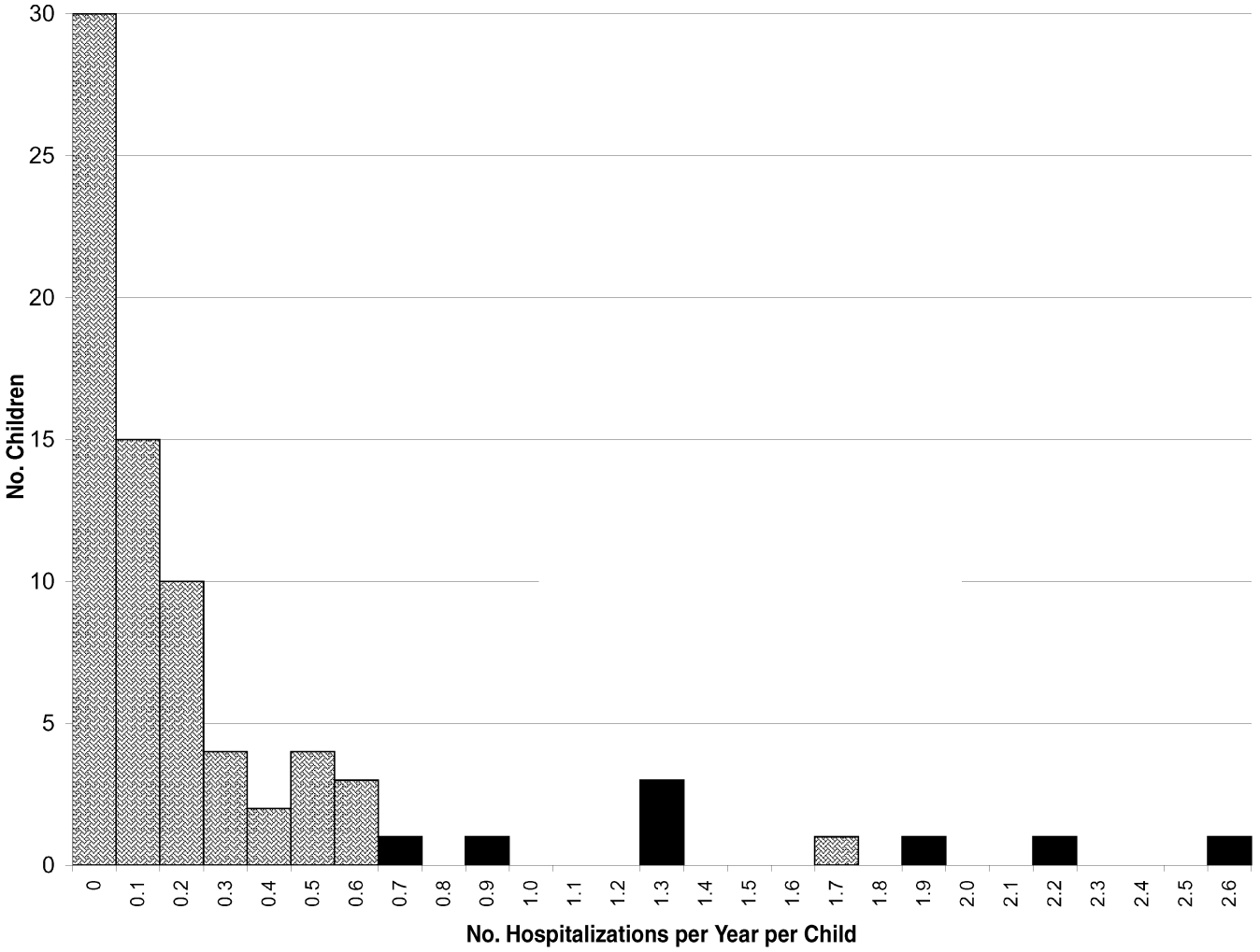
Figure 2 illustrates the wide variation of the cost per hospitalization by child for those with at least one hospitalization. The *y*-axis shows the proportion of children in each category adjusted for the length of the amount interval so that the area of each bar equals the proportion of the total cost that is attributed to that interval. The *x*-axis is the average cost per hospitalization per child. Forty-seven children had at least one hospitalization; the cost per hospitalization per child ranged from $1,000 to $47,100, with the majority of the costs between $5,000 and $20,000 per visit. The distribution of the cost per hospitalization is also right-skewed. The eight High Utilizers are shown above the bars, indicating that they are dispersed in this distribution and not all to the right.

2.2 Model

Details of the Bayesian two-part model and inference of the parameters are found in Rosenberg and Farrell (2007). We modeled by year of age to allow for the left-truncation and right-censoring of the data (Lin et al. 1997). We used WinBUGS version 1.4 for this analysis, which is a specialized software program designed to implement Bayesian models. For a tutorial on WinBUGS see Scollnik (2001) or Fryback, Stout, and Rosenberg (2001), or for further reading on understanding Bayesian models see Gelman et al. (2004) and Gilks, Richardson, and Spiegelhalter (1996).

The Markov Chain Monte Carlo (MCMC) technique was used to estimate the parameters.

Figure 1
Distribution of Number of Hospitalizations per Year by Child Adjusted for Exposure with High Utilizers Shown as Dark-Shaded Cells



MCMC is a simulation technique (the Monte Carlo part) and combined with Markov chain concepts creates simulated draws from the posterior distribution of the parameters. The set of all draws form the joint posterior distribution of the parameters. A description of MCMC is provided in the Appendix.

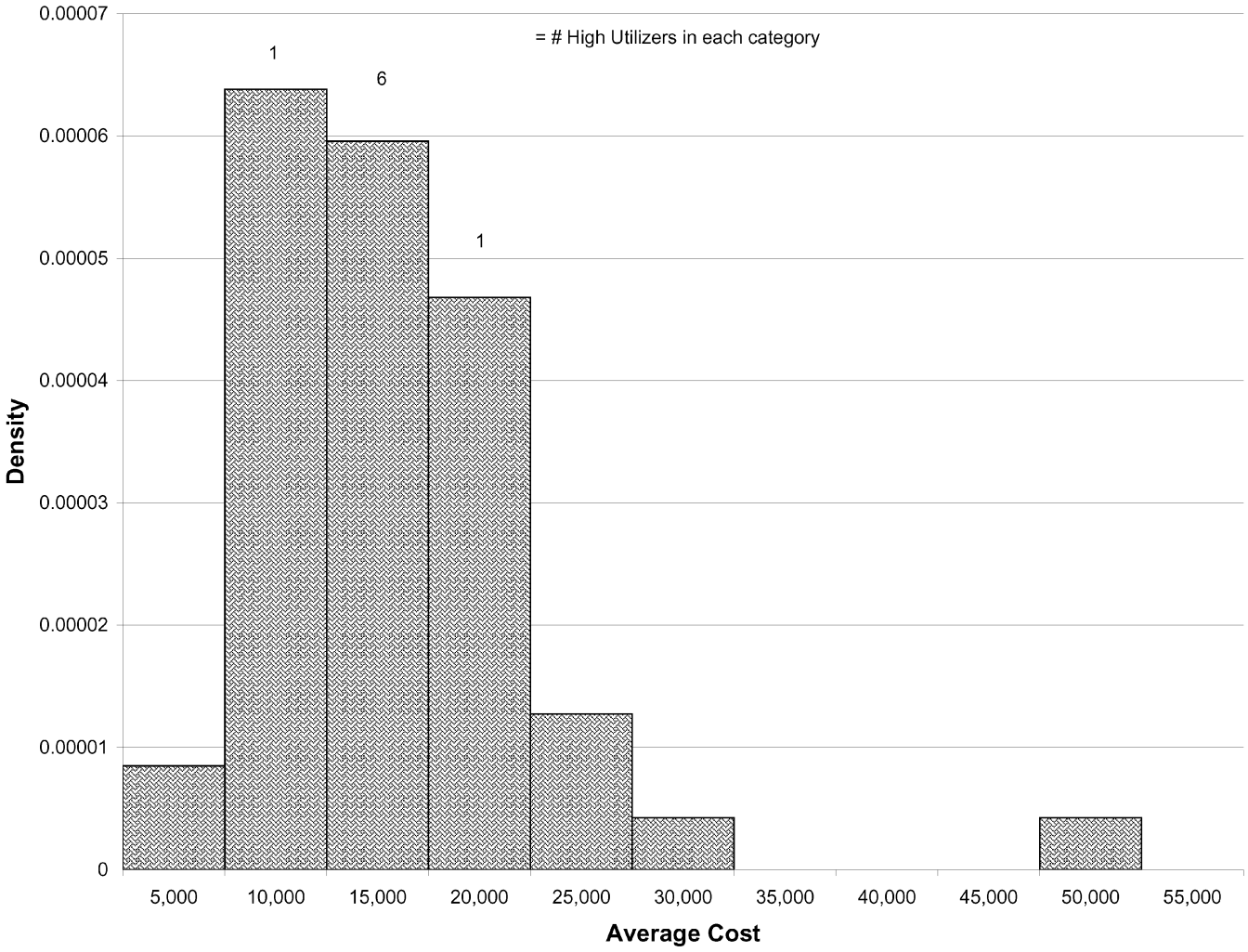
Although a majority of children had no hospitalizations, some children had multiple hospitalizations within a year of age. We modeled the number of hospitalizations for a child i at a particular age j as a Poisson distribution with parameter $e_{ij} \cdot \mu_{ij}$, where e_{ij} represents the exposure for the child within the year of age, and μ_{ij} represents the average number of hospitalizations within the year of age for the child. The logarithm of the mean number of hospitalizations at each year of age for each child, $\log(\mu_{i,j})$, was modeled as a linear combination of unknown parameters and explanatory variables, plus a random effects term that incorporated variation at the child level.[2] The logarithm was used to ensure that the estimate of the mean was greater than zero. We assumed noninformative priors for the unknown

[2] In Klugman, Panjer, and Willmot (2004) the Negative Binomial distribution was motivated using a Poisson distribution with the mean having a Gamma distribution. Here we are using a Poisson distribution, with an assumption that the mean has a Lognormal distribution.

parameters that were normally distributed with a zero mean and very large variance. The priors for the individual random effects terms were assumed to have a normal distribution with mean of zero and a common variance across children. The prior for the common variance term was assumed to have a noninformative Inverse Gamma distribution.

The second part modeled costs for the *j*th hospitalization for the *i*th child as Gamma random variables. Gamma random variables have two parameters, α and θ, where α is the shape parameter and θ is the scale parameter. The parameterization as defined by Klugman, Panjer, and Willmot (2004) was used, where the mean was equal to $\alpha \cdot \theta$, so that an increasing scale parameter increased average costs. The Gamma distribution was chosen based on the shape of the data in Figure 2. We allowed the shape and scale parameter to vary by individual. The shape parameter was assumed to be distributed as a Gamma random variable, with the prior parameters set at (5.0, 2.0) so that it was a proper distribution but with wide variance. Sensitivity analyses were completed using other prior distributions, and the final choice of the parameters of the prior distribution did not impact the ultimate results. The scale parameter for the *i*th child was a function of the demographic and severity information. As in the first part of the model, noninformative priors were assumed for the regression parameters.

Figure 2
Distribution of Cost per Hospitalization by Child



The following explanatory variables were included in both parts of the model. A simple biological severity system was defined that grouped children with meconium ileus, an obstruction of the intestine (ileus), in one category (21% of children), children with a severe form of the disease (as measured by genotype) in a second category (51% of the children), and the remainder in a third category (28% of the children). Other explanatory variables included an indicator of female gender, age at hospitalization, indicator of being assigned to the screened group, and age at diagnosis. The first part also included attained age as an explanatory variable.

The Bayesian MCMC process was verified for convergence. Alternative statistical models were explored, varying the assumed distributions and the priors, as well as ways to include the explanatory variables. The statistical importance of the explanatory variables, as well as the fit of the posterior expected number and costs of hospitalizations by year of age by child, was summarized in Rosenberg and Farrell (2007).

2.3 Prediction

Parameters were estimated for both parts of the model that covered ages 0–18. In this article, we focused on the posterior distribution of numbers of hospitalizations and costs of hospitalization in a year of age for two particular children and for the entire group of active children for a particular calendar year. The posterior distribution of the parameters by child were used as inputs to a simulation of the numbers of hospitalization by year of age and the cost per hospitalization by child to determine the aggregate costs by child and by calendar year. Estimates of total costs for 77 children in a particular calendar year were based on age and exposure of children in that calendar year.

To simulate the frequency of hospitalizations, we used the draws for $\mu_{i,j}$ from the MCMC simulation, as this parameter represented the expected number of hospitalizations by year of age for child i. Poisson random variables were simulated for each child with an adjustment of the known exposure made for comparisons to the observed numbers.

For each of the predicted hospitalizations, we simulated the cost per hospitalization. Different methods of prediction were required for those children who had incurred at least one hospitalization during the study period as well as for those children who had no hospitalizations. For children without observed costs, the simulation may have generated some hospitalizations; therefore some method of generating costs was needed.

For those 47 children with cost data we used the simulated values for (α_i, θ_i) for each child as generated by our model, and simulated a Gamma random variable with those parameters for the number of hospitalizations required.

For each of the 30 children without cost data we needed (α_i, θ_i) to simulate costs. A simulated value of $\log(\theta_i)$ was found based on individual characteristics and posterior distributions of cost regression parameters. For these children their α_i is a random variable from the posterior distribution based on the values of all α_i from all children with costs. Once the (α_i, θ_i) were simulated, then the costs were simulated as for those children who had hospitalizations.

Bayesian models assume that parameters are random variables and not point estimates. Using the parameters generated by the MCMC process produces a distribution of either the number of hospitalizations or cost per hospitalization. For example, each $\mu_{i,j}$ was used to simulate the number of hospitalizations for child i at age j, and for each hospitalization each (α_i, θ_i) was used to simulate the cost per hospitalization. The mean, variance, percentiles, or other function is found by using the simulated sample.

Ignoring the randomness of the parameter estimates underestimates the variance of any predictions that were calculated. For example, let $S_{i,\text{Age}}$ be the random variable for the total costs, $N_{i,\text{Age}}$ be the random variable for the number of hospitalizations, and X_i be the random variable for the cost per hospitalization for child i at a particular age. Using conditional expectation, then $\text{E}[S_{i,\text{Age}}] = \text{E}[N_{i,\text{Age}}] \cdot \text{E}[X_i]$, $\text{E}[N_{i,\text{Age}}] = \text{E}[\mu_{i,\text{Age}}]$, and $\text{E}[X_i] = \text{E}[\alpha_i \cdot \theta_i]$. Note that the expectations of $\mu_{i,\text{Age}}$ and $\alpha_i \cdot \theta_i$, the mean of a Poisson and Gamma random variable, respectively, are over their posterior distribution. The variance of total costs per child per age was found by $\text{Var}[S_{i,\text{Age}}] = \text{Var}[N_{i,\text{Age}}] \cdot (\text{E}[X_i])^2 + \text{E}[N_{i,\text{Age}}] \cdot \text{Var}[X_i]$, where $\text{Var}[N_{i,\text{Age}}] = \text{E}[\mu_{i,\text{Age}}] + \text{Var}[\mu_{i,\text{Age}}]$ and $\text{Var}[X_i] = \text{E}[\alpha_i \cdot \theta_i^2] + \text{Var}[\alpha_i \cdot \theta_i]$. In non-Bayesian studies the variance of the number of hospitalizations by year of age would just be the

first component; similarly for the variance of the cost per hospitalization. In the Bayesian analysis the variances of $N_{i,\text{Age}}$ and X_i each have an added component to the variance that adjusts for the extra variation for the parameters. Note also that using a point estimate for the mean in a Monte Carlo simulation will understate the variance regardless of the number of trials used for the simulation.

There are other ways of reflecting the variability of the parameter estimates without using a Bayesian model. However, the Bayesian model incorporates this variability more naturally.

3. RESULTS

The usefulness of the model is illustrated by showing the results for two children in the study. Child B is a high utilizer, while child A is not. We show results for four years—1990, 1992, 1996, and 1999—in aggregate, for the group of High Utilizers, and for the remainder.

Figures 3 and 4 show the posterior distributions of the *average number* of hospitalizations for child A and B for age 0. The graphs show that the average is not a fixed-point estimate, but rather a distribution of values. The overall average for the posterior distribution for the mean number of hospitalizations at age 0 for child A is 0.72, while the overall average for child B is 5.76. The distribution for the High Utilizer is shifted to the right of child A with a much higher variance.

Figures 5 and 6 show a similar pattern for the *average cost* per hospitalization for child A and child B, respectively. These are the posterior distributions of the averages, which are also both skewed to the right. The average cost per hospitalization is higher for this High Utilizer ($8,100 vs. $3,100), and the skewness is approximately 1.4.

Using the distributions for the average number of hospitalizations, we simulate the total number of hospitalizations for each child at age 0. These distributions are shown in Figures 7 and 8. The outcomes are integer-valued to depict the number of hospitalizations. Child A has a 53% probability of not having a hospitalization at age 0, while child B has a 1% chance of not being hospitalized. Not surprisingly, the overall average of the count distributions for each child is the same as the average of the averages discussed above. These distributions are what could be outcomes at each year of age. The figures show how widespread the data are; the summary statistics provide a more concise way of describing the data. Percentiles are easy statistics to calculate from the distribution and are informative about the spread of the data for long-tailed distributions, perhaps more so than the standard deviation.

Similarly Figures 9 and 10 show the predictive distributions of the aggregate costs for children A and B, respectively. This distribution is the combined result of the simulation of the numbers of hospitalizations and the cost per hospitalization. For child A the average is $2,218 with a standard deviation of $3,330; however, the simulation resulted in one outcome of $38,123. For child B the average cost is $46,404 with a standard deviation of $27,871. Here the 95th percentile is $98,460, but one trial of the simulation was $241,568. These results show the impact of the two long-tailed distributions used in the first and second parts of the model, where costs are widely spread.

Figures 11–14 show the predictive distributions of the number of hospitalizations, and Figures 15–18 show the aggregate costs for the entire group of children for 1990, 1992, 1996, and 1999. The horizontal scales are identical for the four years so that changes from one year to the next are more readily apparent. Also included on the graphs are the exposure, the actual number or cost of hospitalizations for the calendar year, and some summary statistics detailing the simulated average, standard deviation, and 5th through 95th percentiles.

For these four years the observed number of hospitalizations is within one standard deviation of the expected. Although the exposure increases until 1996, the number and cost of hospitalizations decrease over time. The observed values fall in different percentiles over time, from the 75th percentile in 1990, 5th percentile in 1992, 83rd percentile in 1996, and 50th percentile in 1999. The costs generally followed the same percentile, except in 1996, where costs were in the 59th percentile. These figures show the long-tailed nature of the data and that an observed value in the tail of the distribution does not necessarily make it wrong but does indicate it is rare.

Figures 19, 20, and 21 illustrate the posterior distributions of the number of hospitalizations in aggregate, for the High Utilizers, and for other than these High Utilizers in 1996. In aggregate the exposure was 60, and the exposure for the High Utilizers was 7. Interestingly, the distribution of counts in 1996 for the High Utilizer is very similar to the rest of the group, even though there was a vast difference in exposure. Figures 22, 23, and 24 show the predictive distributions for the cost in aggregate, for the High Utilizers, and for other than these High Utilizers. Here the distribution for cost shows a longer tail for other than High Utilizers.

All of these statistics, and others such as the coefficient of variation, skewness, minimum, and maximum, are shown in Tables 2–5 by calendar year, separately for the entire group, High Utilizers, and all other.

Figure 4
Posterior Distribution of Average Number of Hospitalizations: Child B, Age 0

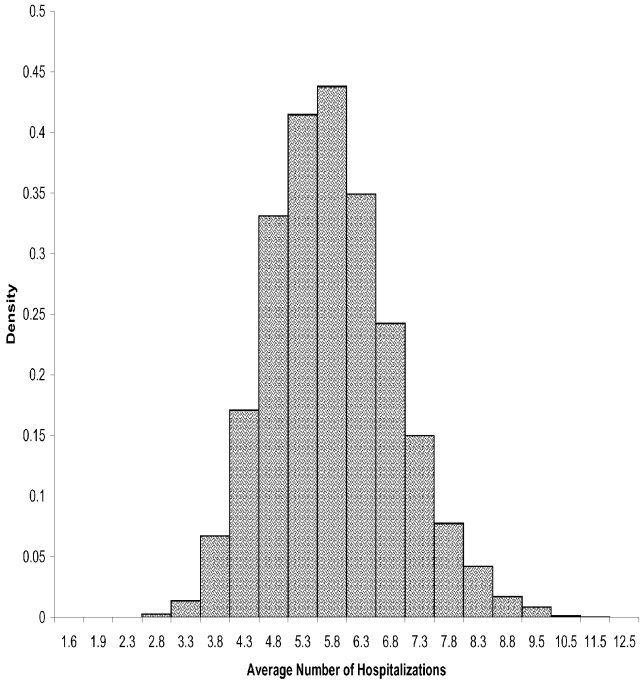


Figure 3
Posterior Distribution of Average Number of Hospitalizations: Child A, Age 0

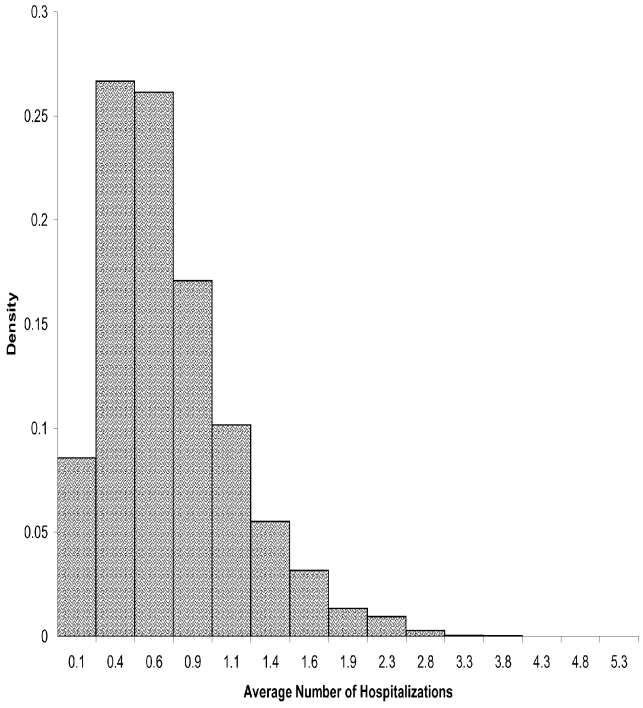


Figure 5
Posterior Distribution of Average Cost per Hospitalization: Child A

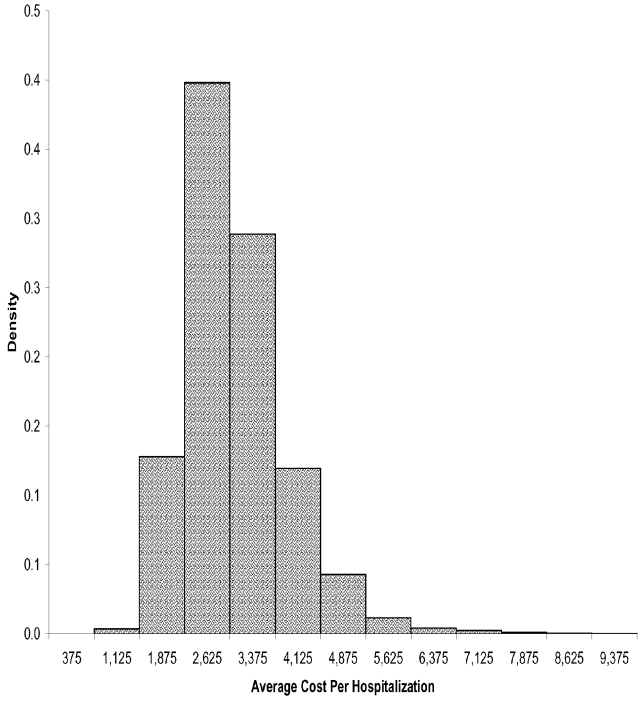


Figure 6

Posterior Distribution of Average Cost per Hospitalization: Child B

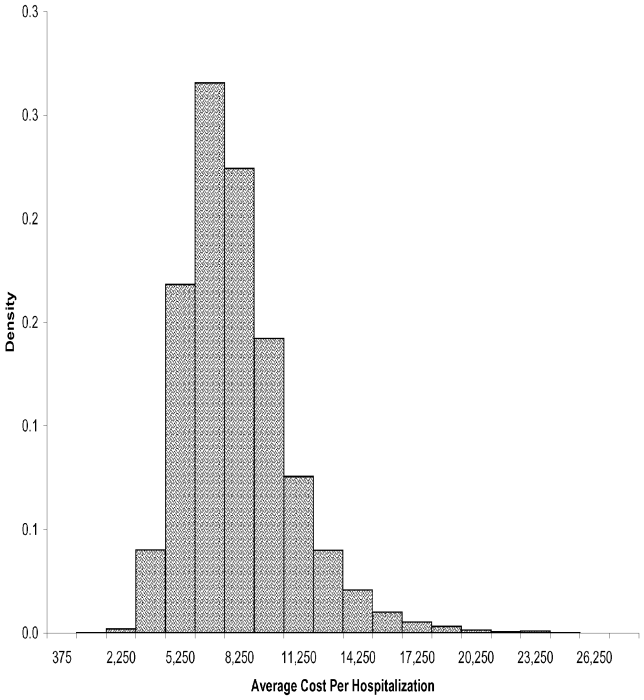


Figure 7

Predictive Distribution of Number of Hospitalizations: Child A, Age 0

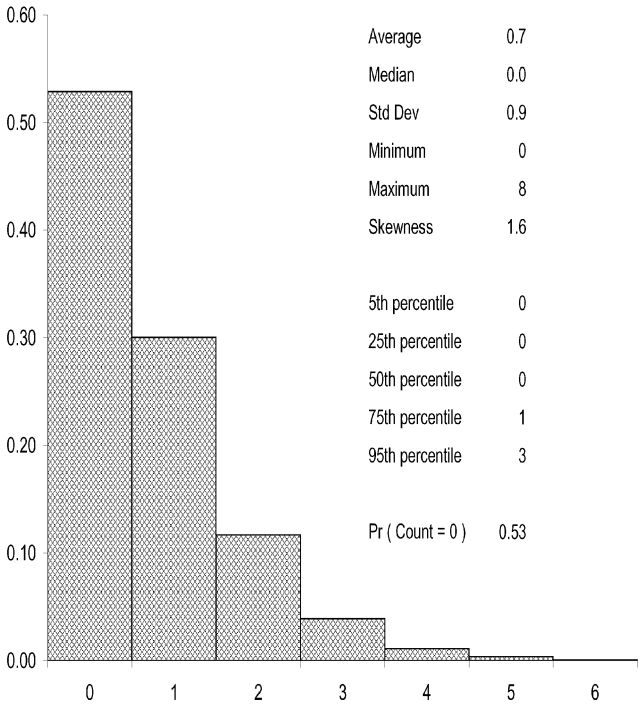


Figure 8

Predictive Distribution of Number of Hospitalizations: Child B, Age 0

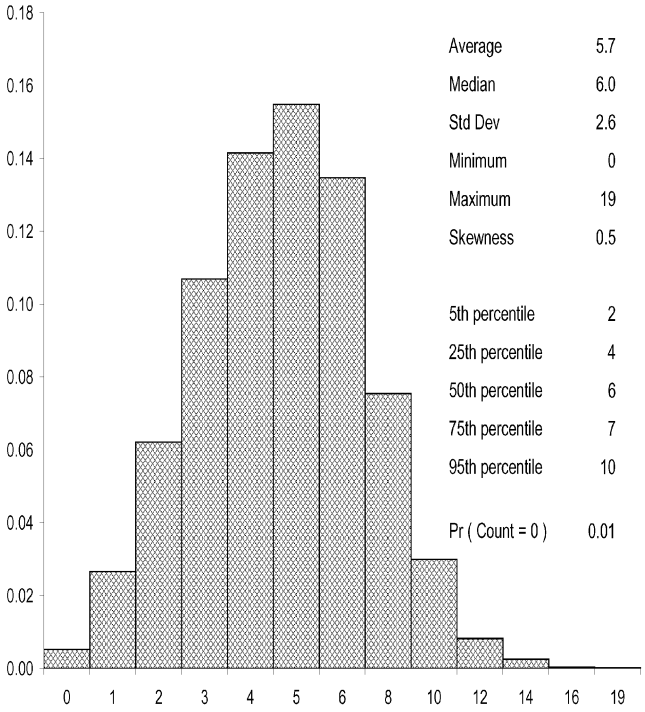


Figure 9

Predictive Distribution of Costs of Hospitalization: Child A, Age 0

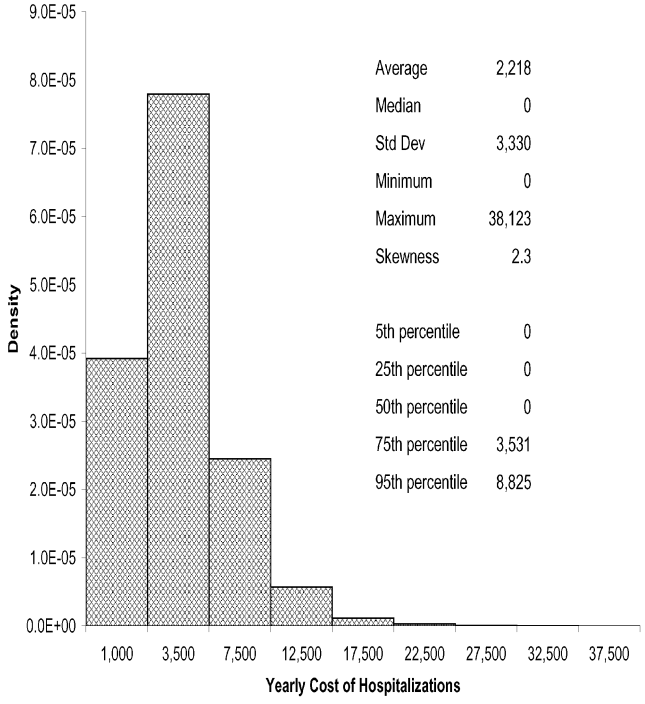


Figure 10
Predictive Distribution of Costs of Hospitalization: Child B, Age 0

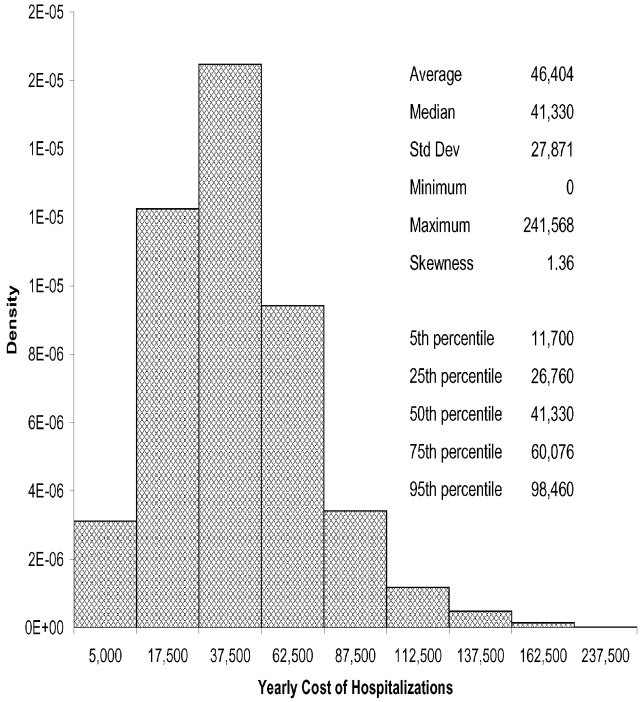


Figure 11
Predictive Distribution of Number of Hospitalizations in 1990

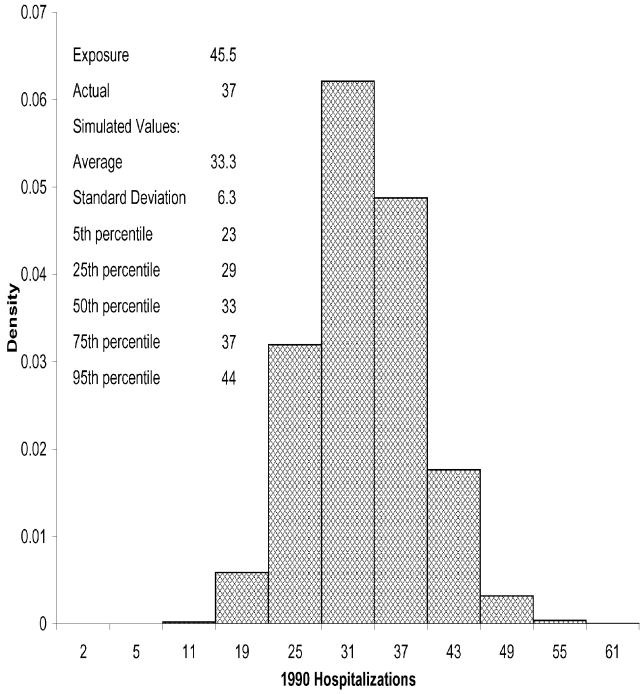


Figure 12
Predictive Distribution of Number of Hospitalizations in 1992

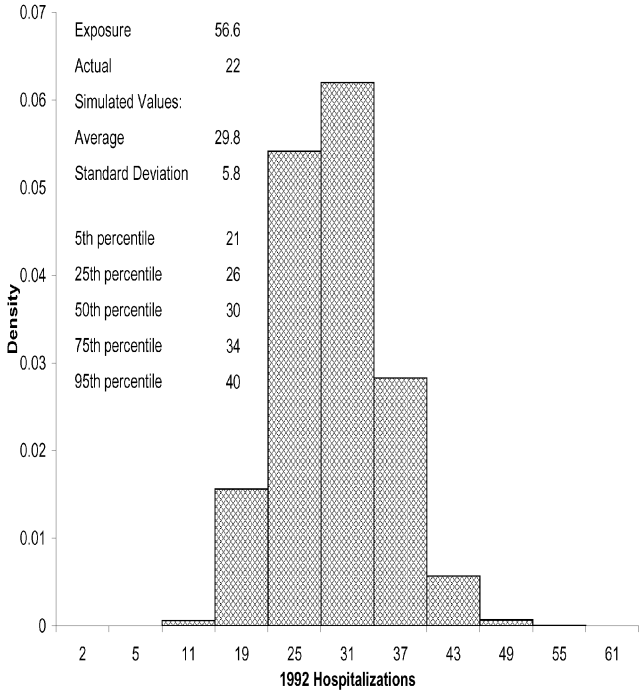


Figure 13
Predictive Distribution of Number of Hospitalizations in 1996

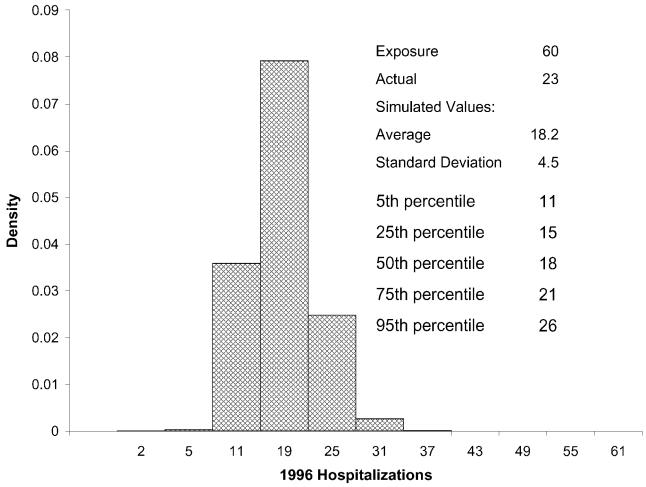


Figure 14
Predictive Distribution of Number of Hospitalizations in 1999

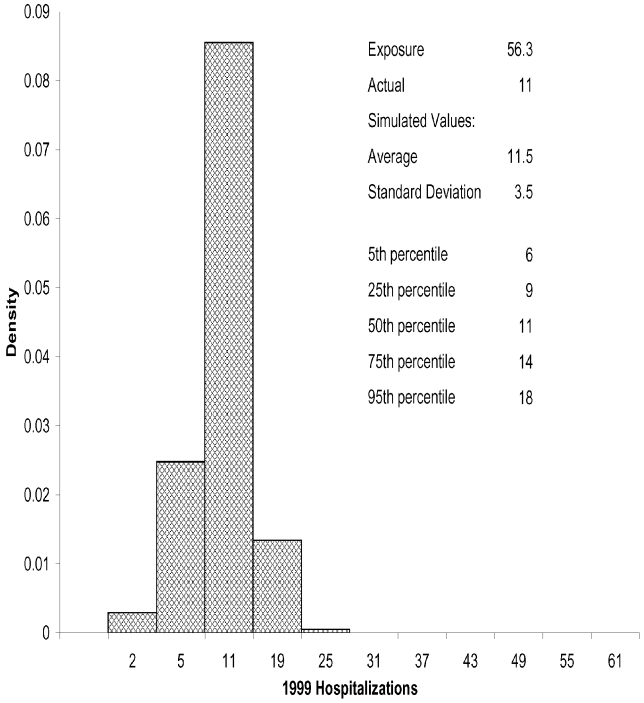


Figure 15
Predictive Distribution of Cost of Hospitalizations in 1990

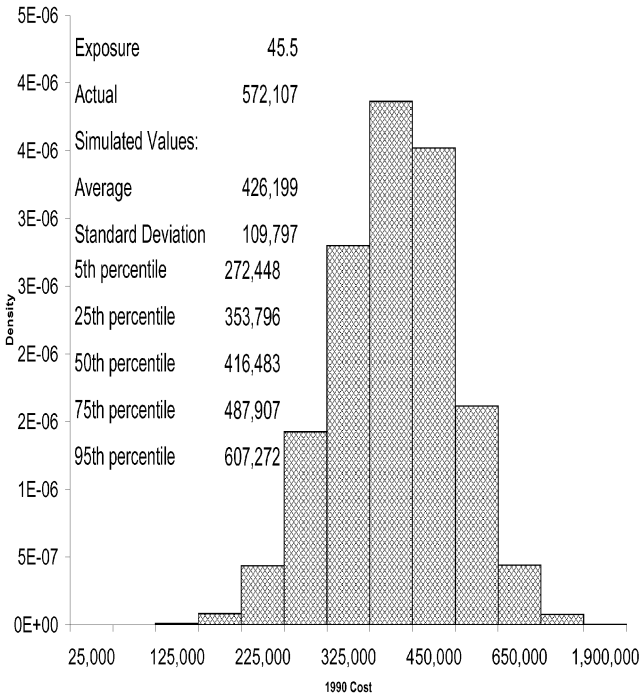


Figure 16
Predictive Distribution of Cost of Hospitalizations in 1992

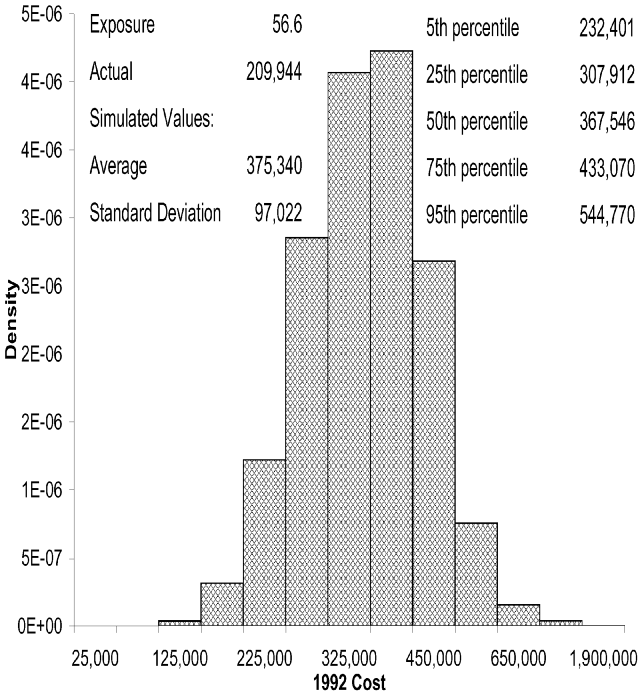


Figure 17
Predictive Distribution of Cost of Hospitalizations in 1996

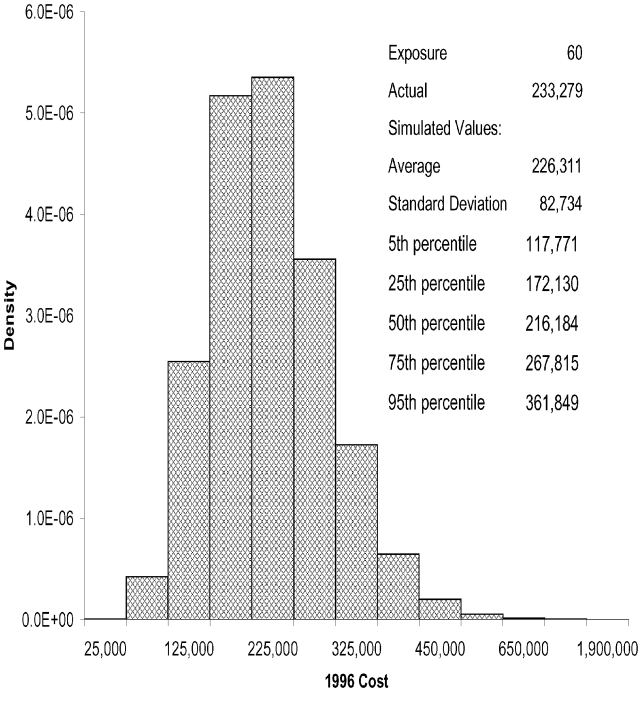


Figure 18
Predictive Distribution of Cost of Hospitalizations in 1999

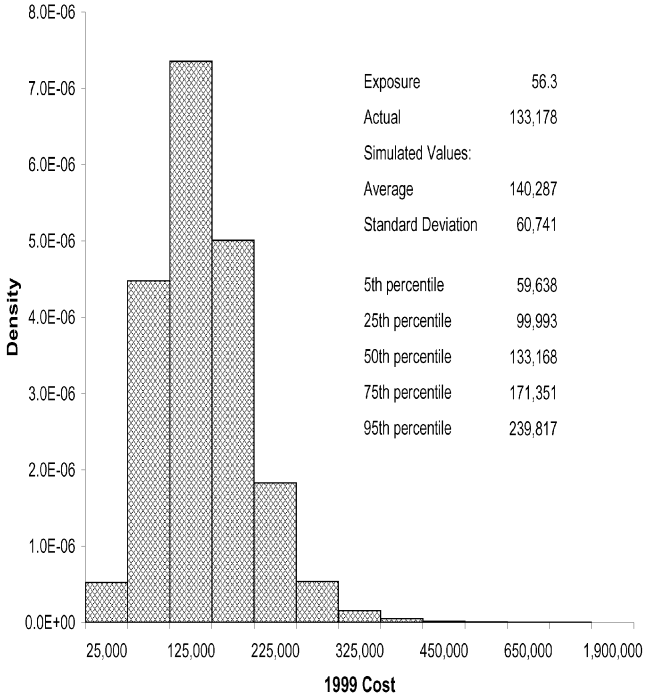


Figure 19
Predictive Distribution of Number of Hospitalizations in 1996: All Children

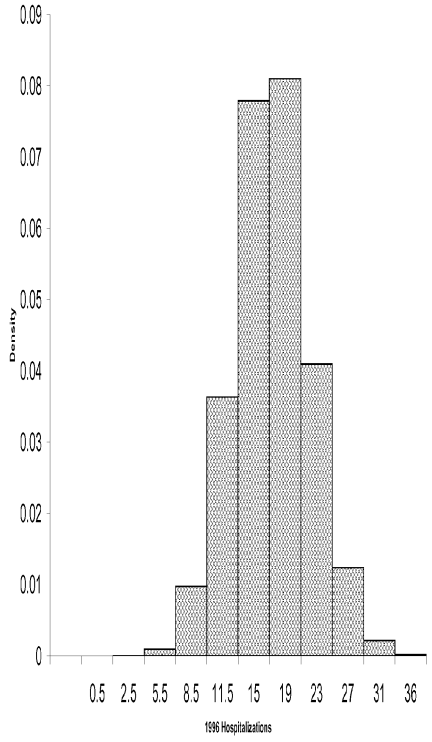


Figure 20
Predictive Distribution of Number of Hospitalizations in 1996: High Utilizers

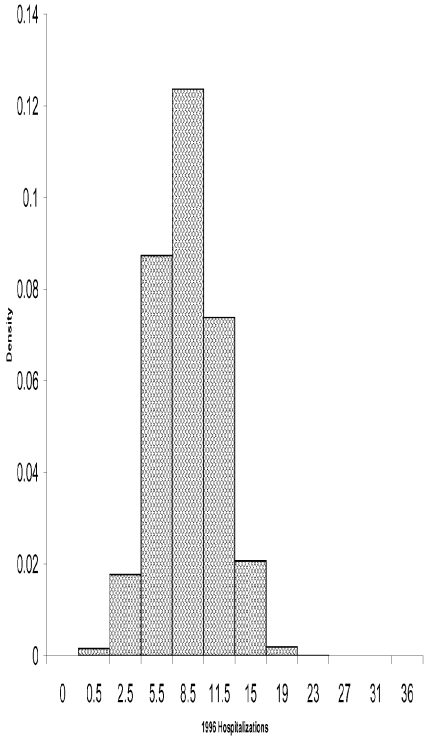


Figure 21
Predictive Distribution of Number of Hospitalizations in 1996: Other than High Utilizers

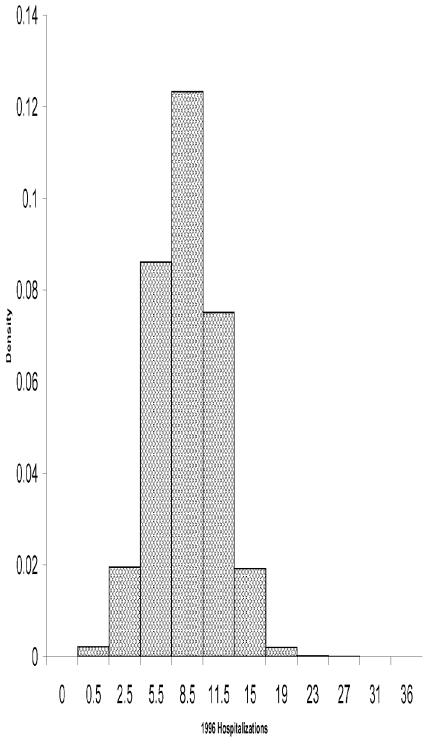


Figure 22

Predictive Distribution of Cost of Hospitalizations in 1996: All Children

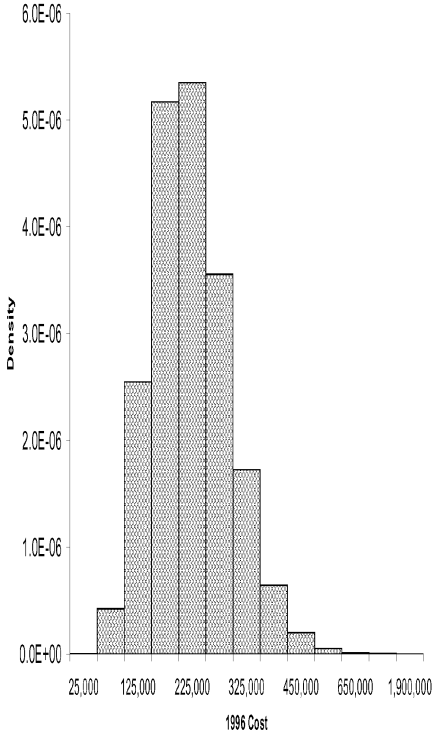


Figure 23

Predictive Distribution of Cost of Hospitalizations in 1996: High Utilizers

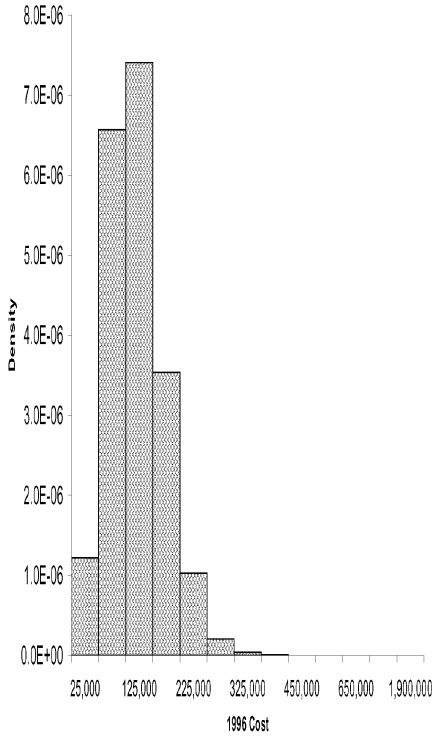


Figure 24

Predictive Distribution of Number of Hospitalizations in 1996: Other than High Utilizers

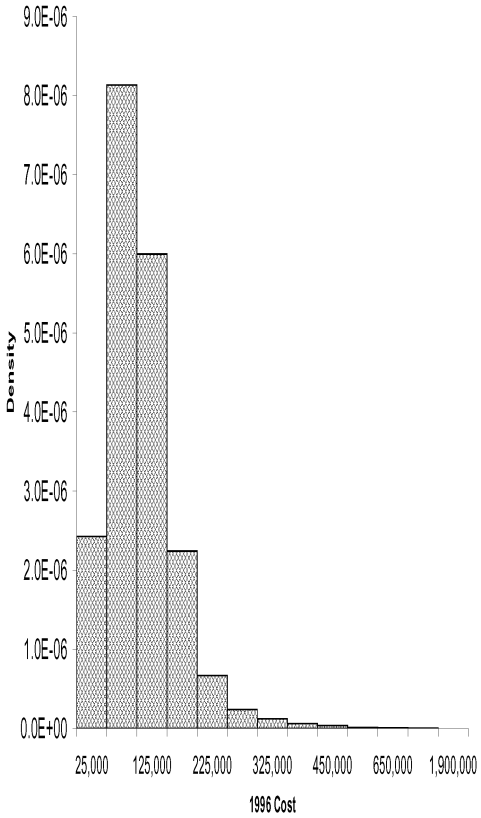


Table 2
1990 Hospitalization Summary for Entire Group, High Utilizers, and Remainder

Statistic	Number			Cost		
	Entire Group	High Utilizers Only	Remainder	Entire Group	High Utilizers Only	Remainder
Actual	37	26	11	572,107	406,171	165,936
Simulated values						
Average	33.3	22.9	10.3	426,199	290,403	135,796
Median	33.0	23.0	10.0	416,483	285,029	125,412
Standard Deviation	6.3	5.1	3.5	109,797	78,656	76,300
Coefficient of Variation	0.2	0.2	0.3	0.3	0.3	0.6
Minimum	13	6	0	114,517	73,449	0
Maximum	61	45	26	3,462,740	693,371	3,290,141
Skewness	0.2	0.2	0.3	2.6	0.4	7.8
Percentiles						
5th percentile	23	15	5	272,448	171,392	48,779
25th percentile	29	19	8	353,796	234,214	90,273
50th percentile	33	23	10	416,483	285,029	125,412
75th percentile	37	26	13	487,907	339,964	168,098
95th percentile	44	32	16	607,272	428,470	250,155
Pr(Simulated Total < Actual)	70%	70%	54%	92%	92%	74%

Table 3
1992 Hospitalization Summary for Entire Group, High Utilizers, and Remainder

Statistic	Number			Cost		
	Entire Group	High Utilizers Only	Remainder	Entire Group	High Utilizers Only	Remainder
Actual	22	12	10	209,944	88,720	121,224
Simulated values						
Average	29.8	18.3	11.5	375,340	232,796	142,543
Median	30.0	18.0	11.0	367,546	227,169	133,268
Standard Deviation	5.8	4.5	3.7	97,022	70,104	66,864
Coefficient of Variation	0.2	0.2	0.3	0.3	0.3	0.5
Minimum	11	3	1	118,167	40,467	785
Maximum	55	39	28	1,107,436	594,231	912,354
Skewness	0.2	0.2	0.4	0.7	0.4	1.7
Percentiles						
5th percentile	21	11	6	232,401	126,890	55,959
25th percentile	26	15	9	307,912	183,353	97,306
50th percentile	30	18	11	367,546	227,169	133,268
75th percentile	34	21	14	433,070	276,921	175,698
95th percentile	40	26	18	544,770	356,824	257,764
Pr(Simulated Total < Actual)	7%	6%	31%	2%	1%	42%

Table 4
1996 Hospitalization Summary for Entire Group, High Utilizers, and Remainder

Statistic	Number			Cost		
	Entire Group	High Utilizers Only	Remainder	Entire Group	High Utilizers Only	Remainder
Actual	23	9	14	233,279	77,277	156,002
Simulated values						
Average	18.2	9.1	9.1	226,311	118,526	107,785
Median	18.0	9.0	9.0	216,184	112,964	96,492
Standard Deviation	4.5	3.1	3.1	82,734	49,613	65,924
Coefficient of Variation	0.2	0.3	0.3	0.4	0.4	0.6
Minimum	4	0	0	32,773	0	0
Maximum	39	23	26	1,763,026	384,669	1,670,289
Skewness	0.3	0.4	0.4	2.1	0.6	3.9
Percentiles						
5th percentile	11	4	4	117,771	46,682	35,507
25th percentile	15	7	7	172,130	82,415	67,686
50th percentile	18	9	9	216,184	112,964	96,492
75th percentile	21	11	11	267,815	148,263	132,601
95th percentile	26	15	14	361,849	208,612	210,295
Pr(Simulated Total < Actual)	83%	45%	91%	59%	21%	85%

Table 5
1999 Hospitalization Summary for Entire Group, High Utilizers, and Remainder

Statistic	Number			Cost		
	Entire Group	High Utilizers Only	Remainder	Entire Group	High Utilizers Only	Remainder
Actual	11	9	2	133,178	125,561	7,617
Simulated values						
Average	11.5	5.6	5.9	140,287	72,075	68,212
Median	11.0	5.0	6.0	133,168	67,338	60,067
Standard Deviation	3.5	2.5	2.5	60,741	37,829	47,387
Coefficient of Variation	0.3	0.4	0.4	0.4	0.5	0.7
Minimum	1	0	0	1,175	0	0
Maximum	27	18	21	1,228,494	271,518	1,179,855
Skewness	0.3	0.5	0.5	2.3	0.8	4.5
Percentiles						
5th percentile	6	2	2	59,638	18,998	15,891
25th percentile	9	4	4	99,993	44,347	38,598
50th percentile	11	5	6	133,168	67,338	60,067
75th percentile	14	7	7	171,351	94,022	87,304
95th percentile	18	10	10	239,817	141,780	143,725
Pr(Simulated Total < Actual)	40%	87%	2%	50%	91%	2%

4. CONCLUSION

This Bayesian model provides an approach to predict hospitalizations and costs based on long-tailed, complex data. The model was able to consider whether data were censored or truncated and to distinguish by child. The Bayesian model produced posterior distributions of the parameters that enabled predictions of costs for a child, as well as in the aggregate. While averages and standard deviations are useful, the entire distribution shows a range of possible values that is helpful when the distribution is skewed. For long-tailed distributions like utilization and cost, outliers are not necessarily true outliers, but rather can be samples in the right tail of a distribution with a fat tail. These observations should not necessarily be trimmed or thrown out, especially for small groups. The predictive distribution allows one to compute the probability of having a cost as high as the one that was incurred. The built-in severity component enabled costs and utilization to be automatically adjusted for any individual movement in or out of the group. The model, while not designed for calendar time, adapted to changes in utilization over time through changes in exposure and the ages of the exposed children.

As a simple comparison, we used a sample from the 2003 Healthcare Cost and Utilization Project (HCUP) Nationwide Inpatient Sample sponsored by the Agency for Healthcare Research and Quality. The HCUP data are a nationwide representative sample of hospital discharges. We sampled 1,938 discharges where the records contained an ICD9 code of cystic fibrosis (277.00, 277.01, 277.02, 277.03, 277.09) in any of the diagnosis fields on the record. The All Patient Refined Diagnosis Related Group (APR-DRG) is a system of classifying discharges based on more clinically meaningful groups and includes another code for disease severity. Of these 1,938 discharges, 95 were missing the APR-DRG code and severity. Of the remaining 1,843 discharges, 1,270 had an APR-DRG of cystic fibrosis, and the remaining 573 (31%) had an APR-DRG of something other than cystic fibrosis. In fact, there were 106 APR-DRGs for these 573 discharges. Sixteen hundred fifty-nine of the 1,938 (85%) were from one of three Major Diagnostic Groups (MDGs) (respiratory system, digestive system, or endocrine, nutritional, and metabolic system). The other 15% of discharges were from 16 other MDGs. Seventy-four percent were either level 2 or 3 for severity (moderate or major loss of functioning) and 64% for level 1 mortality (minor likelihood of dying). Although the data are discharge-based and not by person, the data showed the heterogeneity of CF hospitalization data with the variety of diagnostic groups represented. The Bayesian model designed for this study was simple from a variable inclusion perspective, but accurate in predicting costs and hospitalizations by calendar year.

While the prevalence of cystic fibrosis in the population is low as compared to the prevalence of heart disease, cancer, and other leading

chronic diseases, analysis of utilization of health care services for children with CF provides a useful approach to modeling costs from these other diseases. Extending the lifetime of those with CF will increase the prevalence of those with the disease and increase the overall costs of the disease over time. For instance, those living longer with CF can live long enough to develop chronic lung disease or diabetes mellitus. Children with chronic disease with increased life expectancy grow to be adults with chronic disease. Smoking status, weight, and age are examples of risk factors that can be included in individual-level modeling for heart disease or cancer. Prediction of cost outcomes for use in disease management programs is easily completed for comparing one subgroup with another as shown with the comparison of some High Utilizers with the rest of children.

Actuaries are often involved in the design and analysis of disease management and case management programs. In case management programs, health care professionals coordinate the care (Dove and Duncan 2004). For instance, Kretz and Pantos (1996) analyzed data for one female CF patient with a severe form of the disease, where case management decreased the cost of care by 33% through a reduction of hospitalizations by aggressively trying to improve lung function, having nursing care in a home setting, and improving the nutritional status of the patient. Disease management programs coordinate health care interventions (www.dmaa.org). These programs require some analysis to gauge their success. Models, such as the one presented in this article, would provide information to help measure the success of a program or provide input to pricing of insurance programs.

5. Acknowledgments

We would like to thank Anita Laxova, University of Wisconsin School of Medicine and Public Health, Mary Sue Nevel, University of Wisconsin Hospital and Clinics, and Danielle Rhiner, University of Wisconsin Medical Foundation, for their assistance in the data collection process. This work was supported by grants from the National Institutes of Health (DK 34108 and M01 RR03186 from the National Center for Research Resources to the University of Wisconsin Medical School) and the Cystic Fibrosis Foundation (A001-5-01).

Appendix

Markov Chain Monte Carlo Technique

In the early 1990s the Markov Chain Monte Carlo (MCMC) approach to conducting Bayesian analyses was introduced (Gelfand et al. 1990). The technique revolutionized the use of Bayesian models, as it was a simulation-based method that could generate posterior distributions of unknown parameters and functions of the parameters. With the increasing power of computers, today Bayesian models are increasingly seen in published papers. Using key words ''Bayesian health care'' in a Google search produced a listing of over 24,000 papers. Makov (2001), along with the accompanying discussion, summarized the use of Bayesian models in actuarial-related areas.

The original MCMC methods were the Metropolis-Hastings algorithm, with a special case called the Gibbs Sampler (Metropolis et al. 1953; Hastings 1970). The Gibbs Sampler is described here. The interested reader is referred to textbooks such as Gelman et al. (2004) and Gilks, Richardson, and Spiegelhalter (1996).

The key idea involves generating a sample from a distribution that cannot be simulated from directly. Using ideas from Markov chain theory, an equilibrium distribution is found from which simulated draws are taken. The equilibrium distribution in this case is the posterior distribution. Simulated values of the Markov chain are used to summarize the posterior distribution. The mean of the simulated values is an estimate of the mean of the posterior distribution, and similarly the variance of the simulated values is an estimate of the variance of the posterior distribution. A graph of all of the simulated values is a graphical estimate of the posterior density (Smith and Roberts 1993).

To create the posterior distribution, the full conditionals of each unknown parameter, given all the other parameters, is determined. These full conditionals are distributions that can be simulated. Initial values for each parameter are determined, and the full conditionals are sampled

in a predetermined cycle for a specified number of iterations.

The following details are modified from Rosenberg and Young (1999). Suppose we are given a joint density of three unknown parameters, considered as random variables. Let the joint density of U, V, and W be denoted by $[U, V, W]$, and suppose we are interested in obtaining the marginal density of U, denoted by $[U]$, and calculated analytically as $[U] = \int \int [U, V, W]\, dv\, dw$. This marginal distribution is generally difficult to calculate directly. The full conditionals are defined as $[U|V, W]$, $[V|U, W]$, and $[W|V, U]$ and provide a way to generate a sample from the marginal of U without calculating $[U]$ via integration.

We start with initial values of V and W, $V^{(0)}$ and $W^{(0)}$ to simulate $U^{(1)}$. Then using $U^{(1)}$ and $W^{(0)}$ we simulate a $V^{(1)}$. Finally, $U^{(1)}$ and $V^{(1)}$ are used to simulate $W^{(1)}$. The triplet $(U^{(1)}, V^{(1)}, W^{(1)})$ form one draw from the joint density of $[U, V, W]$. Under mild regularity conditions the distribution of $U^{(j)}$ converges to the marginal distribution of U as j gets large. For large m the observation $U^{(m)}$ can be treated as a realization of the random variable U. In practice one simulates a sequence with m large, and discards the first k values, called a burn-in, to eliminate the dependency of the sequence on the initial values. The remaining $m - k$ values are then used to estimate the density of U or other functions of U.

In this article, we use the joint draws from the posterior distributions of the parameters to simulate the number of hospitalizations per child for each year of age. For each hospitalization we simulate a cost per hospitalization. The predictive distributions of the total number of hospitalizations, and their respective costs, for each calendar year are sums of the number and costs of hospitalizations for the children in the study. Thus, here these predictive distributions are a function of the regression parameters used to simulate the number and cost of hospitalizations for each child.

REFERENCES

AMERICAN HEART ASSOCIATION. 2005. *Heart Disease and Stroke Statistics*. 2005 Update. Dallas: American Heart Association.

BONOW, R. O., L. A. SMAHA, S. C. SMITH, G. A. MENSAH, AND C. LENFANT. 2002. The International Burden of Cardiovascular Disease: Responding to the Emerging Global Epidemic. *Circulation* 106: 1602–5.

BOWERS, NEWTON L., JR., HANS U. GERBER, JAMES C. HICKMAN, DONALD A. JONES, AND CECIL J. NESBITT. 1997. *Actuarial Mathematics*. 2nd edition. Schaumburg, IL: Society of Actuaries.

CENTERS FOR DISEASE CONTROL AND PREVENTION. 2004. The Burden of Chronic Diseases and Their Risk Factors: National and State Perspectives 2004. www.cdc.gov/nccdphp/burdenbook2004.

FARRELL, P. M., AND E. H. MISCHLER. 1992. Newborn Screening for Cystic Fibrosis. *Advances in Pediatrics* 39: 35–70.

FARRELL, P. M., AND WISCONSIN CYSTIC FIBROSIS NEONATAL SCREENING STUDY GROUP. 2000. Improving the Health of Patients with Cystic Fibrosis through Newborn Screening. *Advances in Pediatrics* 47: 79–115.

FITZSIMMONS, S. 1995. Cystic Fibrosis: What's New. *Journal of Insurance Medicine* 27(2): 124–30.

FOST, N. C., AND P. M. FARRELL. 1989. A Prospective Randomized Trial of Early Diagnosis and Treatment of Cystic Fibrosis: A Unique Ethical Dilemma. *Clinical Research* 37: 495–500.

FRYBACK, D., N. STOUT, AND M. ROSENBERG. 2001. An Elementary Introduction to Bayesian Computing Using WinBUGS. *International Journal of Technology Assessment in Health Care* 77(1): 98–113.

GELFAND, A., S. HILLS, A. RACINE-POON, AND A. SMITH. 1990. Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association* 85: 972–85.

GELMAN, ANDREW, JOHN B. CARLIN, HAL S. STERN, AND DONALD B. RUBIN. 2004. *Bayesian Data Analysis*. 2nd edition. London: Chapman and Hall.

GILKS, W., S. RICHARDSON, AND DAVID SPIEGELHALTER. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

HASTINGS, W. K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* 57: 97–109.

KLUGMAN, STUART A., HARRY H. PANJER, AND GORDON E. WILLMOT. 2004. *Loss Models: From Data to Decisions*. 2nd edition. New York: John Wiley and Sons.

LIN, D. Y., E. J. FEUER, R. ETZIONI, AND Y. WAX. 1997. Estimating Medical Costs with Incomplete Follow-up Data. *Biometrics* 53: 419–34.

MAKOV, UDI E. 2001. Principal Applications of Bayesian Methods in Actuarial Science: A Perspective. *North American Actuarial Journal* 5(4): 53–73.

MARSHALL, B. C. 2004. Pulmonary Exacerbations in Cystic Fibrosis: It's Time to Be Explicit. *American Journal of Respiratory and Critical Care Medicine* 169: 781–82.

METROPOLIS, N., A. ROSENBLUTH, A. TELLER, AND E. TELLER. 1953. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* 21: 1087–92.

ROSENBERG, MARJORIE A., AND PHILIP M. FARRELL. 2007. Impact of a Newborn Screening Program on Inpatient Utilization for Children with Cystic Fibrosis. Working paper. http://research3.bus.wisc.edu/mrosenberg.

ROSENBERG, MARJORIE A., EDWARD W. FREES, JIAFENG SUN, PAUL H. JOHNSON JR., AND JAMES M. ROBINSON. 2007. Predictive Modeling with Longitudinal Data: A Case Study of Wisconsin Nursing Homes. *North American Actuarial Journal* 11(3): 54–69.

ROSENBERG, MARJORIE A., AND PAUL H. JOHNSON JR. 2007. Health Care Predictive Modeling Tools. *Health Watch* 54: 24–27.

ROSENBERG, MARJORIE A., AND VIRGINIA R. YOUNG. 1999. A Bayesian Approach to Understanding Time Series Data. *North American Actuarial Journal* 3(2): 130–43.

SCOLLNIK, D. P. M. 2001. Actuarial Modeling with MCMC and BUGS. *North American Actuarial Journal* 5(2): 96–124.

SILBER, J. H., S. P. GLEESON, AND H. M. ZHAO. 1999. The Influence of Chronic Disease on Resource Utilization in Common Acute Pediatric Conditions: Financial Concerns for Children's Hospitals. *Archives of Pediatric and Adolescent Medicine* 153(2): 169–79.

SMITH, A. F. M., AND G. O. ROBERTS. 1993. Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of Royal Statistical Society, Series B* 55: 3–23.

Discussions on this paper can be submitted until July 1, 2008. The authors reserve the right to reply to any discussion. Please see the Submission Guidelines for Authors on the inside back cover for instructions on the submission of discussions.