



# **CareScience Length of Stay Risk Model**

Eugene Kroch, PhD  
Michael Duan, MS  
Emi Terasawa

(compiled by Emi Terasawa)

June 26, 2007

# Table of Contents

<b>I. SIGNIFICANCE</b> .....	<b>3</b>
<b>II. MEASUREMENT CHALLENGE</b> .....	<b>3</b>
<b>III. CARESCIENCE LENGTH OF STAY RISK MODEL</b> .....	<b>3</b>
3.1 DEFINING LENGTH OF STAY .....	3
3.2 RISK MODEL SPECIFICATION .....	4
3.3 INDEPENDENT VARIABLES .....	4
3.3.1 CACR Comorbidity Scores .....	5
3.3.2 Valid Procedures .....	5
3.3.3 Missing Independent Variables .....	6
3.4 SEMI-LOG MODEL .....	7
3.4.1 Geometric vs. Arithmetic Means .....	7
3.5 OUT OF RANGE PREDICTIONS .....	8
<b>IV. DATA SOURCE AND MODEL CALIBRATION</b> .....	<b>8</b>
4.1 MEDPAR DATA .....	8
4.2 ALL-PAYOR STATE DATA .....	8
4.3 PRIVATE CLIENT DATA .....	9
4.4 MODEL SELECTION FOR PRIVATE CLIENT DATA .....	9
4.5 MODEL SELECTION FOR PUBLIC DATA .....	9
<b>V. PERFORMANCE ASSESSMENT</b> .....	<b>9</b>
<b>APPENDIX A – SEMILOG MODELING</b> .....	<b>12</b>

## I. Significance

Length of stay is among the most popular outcome measures for hospitals engaging in performance improvement. Easily observed and measured, length of stay is less controversial and less emotionally charged than many outcome measures, making it an ideal choice for efforts requiring staff buy-in. Additionally, length of stay is more easily addressable and actionable than many other outcomes, further enhancing its attractiveness for performance improvement.

As an outcome measure, length of stay serves as a proxy for resource usage, reflecting how efficiently a hospital allocates staff time, space, equipment, and additional considerations per patient. Accordingly, it correlates highly with cost. In addition to these economic factors, length of stay holds quality implications for hospitals.<sup>1</sup> The longer a patient remains in hospital the longer is his exposure to various inpatient risks such as medication errors and hospital infections. Longer lengths of stay can also contribute to hospital congestion that can affect efficiency and in turn quality.

## II. Measurement Challenge

Despite its desirable attributes, length of stay is prone to varying hospital discharge policies that can bias it as an outcome measure. Hospitals that regularly transfer patients to affiliated long-term care facilities often have reduced lengths of stay. As a result, the relationship between efficiency and length of stay can be soured by the possibility of efficiency being achieved at the cost of sufficient treatment.

Also, although longer lengths of stay expose patients to increased opportunities for inpatient errors, the relationship between complications and length of stay is confounded by a “chicken and egg” type problem. Often it is impossible to discern whether a patient’s length of stay is longer due to the occurrence of a complication or whether the patient’s longer length of stay allowed him to fall victim to the complication.

## III. CareScience Length of Stay Risk Model

### ***3.1 Defining Length of Stay***

Length of Stay (LOS) is defined as the number of full days a patient stays in the hospital. It is calculated as the difference between discharge date and admission date. The shortest valid LOS is one day. If a patient is admitted and discharged on the same day, LOS is counted as one day. If a patient stays in the hospital for more than 100 days, the case is dismissed from the LOS analysis as an outlier.

---

<sup>1</sup> Assessing the relationship between length of stay and quality can be problematic, since shorter lengths of stay do not always equate with higher quality. Patients may be discharged inappropriately to achieve shorter hospital stays.

### 3.2 Risk Model Specification

The purpose of the CareScience Length of Stay Risk Model is to generate the expected or “standard” LOS (“risk” rate) under typical care, given the patient’s health status and relevant characteristics. Patient-level mortality risk is assessed via a stratified multiple regression model with the following functional form:

$$y_{ijk} = x_{ijk}\beta_k + \varepsilon_{ijk}, \forall ijk$$

where  $y_{ijk}$  is the natural log of length of stay at patient level  $i$ , provider  $j$ , and principal diagnosis  $k$ .  $x_{ijk}$  is a vector of patient characteristics and socioeconomic factors.  $\beta_k$  is the marginal effect of the independent variables on the mortality outcome measure, and  $\varepsilon_{ijk}$  is the random error component of the model. The strata ( $k$ ) are roughly based on 3-digit level ICD-9-CM diagnosis codes. Rare and insignificant diagnoses are rolled up into broad diagnosis groups, which are defined in the ICD-9-CM book. A total of 142 disease strata are analyzed.

### 3.3 Independent Variables

The following patient characteristics and socioeconomic factors comprise the set of regressors (i.e. classes of independent variables) used in the CareScience Length of Stay Risk Model.

1. **Age** (*quadratic form*)
2. **Birth weight** (*quadratic form, for neonatal model only*)
3. **Sex** (*female, male, unknown*)
4. **Race** (*white, black, asian-pacific islander, unknown*)
5. **Income** (*median household income within a zip code reported by US Census Bureau*)
6. **Distance traveled** (*the centroid-to-centroid distance between the zip code of the household and the zip code of the hospital or provider, represented as a relative term*)
7. **Principal diagnosis** (*terminal or three digit ICD-9-CM code, where statistically significant*)
8. **CACR<sup>2</sup> comorbidity scores** (*count of comorbidities within each of five severity categories on the CACR Likert scale*)
9. **Defining diagnosis** (*three digit ICD9-CM code for neonatal model only*)
10. **Cancer status** (*benign, malignant, carcinoma in situ, history of cancer, derived from secondary diagnoses*)
11. **Chronic disease and disease history** (*terminal digit ICD9-CM diagnosis codes, such as diabetes, renal failure, hypertension, chronic GI, chronic CP, obesity, and history of substance abuse*)
12. **Valid procedure** (*terminal ICD9-CM procedure codes, where clinically relevant and statistically significant*)
13. **Admission source** (*Physician Referral, Clinic Referral, HMO Referral, Transfer from a Hospital, Skilled Nursing Facility or Another Health Care Facility, Emergency Room, Court/Law Enforcement, Newborn - Normal Delivery, Premature Delivery, Sick Baby, or Extramural Birth, Unknown/Other*)

---

<sup>2</sup> Comorbidity Adjusted Complication Risk – Brailer DJ, Kroch E, Pauly MV, Huang J. Comorbidity-Adjusted Complication Risk: A New Outcome Quality Measure, Medical Care 1996; 34:490-505.

14. **Admission type** (*Emergency, Urgent, Elective, Newborn, Delivery, Unknown/Other*)
15. **Payor class** (*Self-pay, Medicaid, Medicare, BC/BS, Commercial, HMO, Workman's Compensation, CHAMPUS/FEHP/Other Federal Government, Unknown/Other*)
16. **Facility type** (*Acute, long-term, Psych.*)

Risk factors used in the CareScience risk assessment model are tailored to specific patient subpopulations and outcomes. Use of the following risk factors may vary depending on the specific subpopulation and outcome evaluated:

- diagnosis detail
- significant comorbidities
- defining procedures
- birth weight (used instead of age for neonates)

### 3.3.1 CACR Comorbidity Scores

CACR comorbidity scores are derived from principal and secondary diagnosis codes. Secondary diagnoses are first categorized according to a five point Likert scale of increasing severity (A-E) where E is most severe.<sup>3</sup> Comorbidities are calculated for each severity level as

$$N_{is} = \sum_{p_{ij} \in S} (1 - p_{ij}), \quad S = A, B, \dots, E$$

where  $N_{is}$  is the expected number of comorbidities of severity  $s$  for a patient with principal diagnosis  $i$ ,  $p_{ij}$  is the CACI probability of complication for the  $j$ th secondary diagnosis given principal diagnosis  $i$ , and  $S$  is one of the severity levels, A-E.

Common chronic diseases enter the model as dummy variables separate from comorbidities. Both comorbidities and chronic diseases are constrained to be non-negative coefficients in the model calibration.

### 3.3.2 Valid Procedures

Strictly speaking, a procedure is not a patient characteristic but rather a provider care choice. For example, two physicians may opt to pursue two different yet equally effective courses of treatment for the same patient. Although procedures represent the discretion of the care provider, they can signal important information about the patient's overall health status. Certain procedures can serve as effective proxies for lab reports and treatment history that are not available in the current database, as well as for other unobservable critical factors. To be included in the model, procedures must be designated as "valid" for the patient's particular disease stratum. Additionally, the timing of certain procedures relative to the patient's hospital admission must be considered. Valid procedures are grouped into one of two categories based on timing criteria.

Each disease stratum has a unique set of valid procedures. If a procedure falls into Category 1, timing of the procedure is not considered, and the analytic program simply searches for the

---

<sup>3</sup> Severity ratings are assigned by an internal panel of clinicians.

procedure's corresponding coefficient. (Procedures failing to be statistically significant are not included in the model and have no impact on the risk score.<sup>4</sup>)

If a procedure is mapped to Category 2, inclusion of the procedure in the model depends on the procedure's timing during the inpatient stay. If the procedure occurs within a critical time period from the patient's hospital admission, the procedure is included in the model. If not, the procedure is excluded. The critical time windows for Category 2 procedures are assigned by internal panels of clinicians.

For several disease strata, the risk model does not incorporate valid procedures. These groups include DRGs 103, 480, 481, 495, 512, and 513.

### **3.3.3 Missing Independent Variables**

As with most large databases, some records may lack one or more independent variables. Dismissing these records completely from the analysis may eliminate important patient information and in turn shrink the base sample size. This is particularly true for public data sets where missing data elements are more common. Recognizing that independent variables have varying impacts on risk scores, the risk model is designed to tolerate missing values to some extent.

#### ***Zero Tolerance***

Principal Diagnosis, Age, and Birthweight (for neonates) are mandatory elements in the risk assessment model. Patient records missing any of these required elements are excluded from the model.

#### ***Conditioned Tolerance***

For most categorical variables, such as Admission Source, there is an 'Unknown' category designated for unrecognizable or missing values. Among the categories, 'Unknown' statistically has the greatest probability of having the highest counts, since missing data are due to random errors. In risk modeling, the largest and most common category is often used as the reference group. Assigning the 'Unknown' category as the reference group is thus justifiable, however, a high proportion of 'Unknown' values risks diluting the real characteristics of the reference group.

Due to tight quality control, 'Unknown' values are very rare in private client data. In public data, however, the missing portion ranges from a couple of percent to around ten percent. It is therefore necessary to check the distribution of the data before calibration. In general, the 'Unknown' values should not represent more than one third of the entire sample in order to be used as the reference group.

---

<sup>4</sup> See Sections 4.4 and 4.5 on Model Selection.

## Value Proxy

Income and Relative Distance are derived from zip code information. In the case of Income, the patient's residence zip code is used. For Relative Distance, both the patient's residence zip code and the hospital zip code are employed. If the patient's zip code is missing, the average Distance and Income of all patients in that hospital will be applied. In cases where both patient and hospital zip codes are unavailable, the Relative Distance is set to 1, and the national average income is applied.

### 3.4 Semi-log Model

LOS is distributed with a rightward (positive) skew. Applying linear regression to data with skewed distributions of dependent variables gives rise to a number of pathologies, including inefficient and often biased, parameter estimates and predictions outside logical bounds (e.g., negative values for LOS and costs). When outcome measures are not symmetrically distributed, analysis of performance can be disproportionately influenced by outliers and extreme cases. A robust solution is to take the natural log of the dependent variable, which results in an approximately symmetric distribution and contracts the outliers inward toward the center of the data (i.e., area of greatest density within the distribution). It also ensures that all predicted values will be positive. (No matter how negative the log value is, taking the anti-log to restore the values will guarantee that they are positive.) Detail on the Semi-log model can be found in Appendix A.

#### 3.4.1 Geometric vs. Arithmetic Means

When working with a semi-log model, geometric means provide a better measure of central tendency than arithmetic means.

The arithmetic mean is the simple average, computed by adding up all values ( $x_i$ ) in the sample and dividing by the number of such values ( $n$ ):

$$\text{arithmetic mean } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

The geometric mean follows the same principle, but instead of adding the values and dividing by  $n$ , they are multiplied together and the  $n^{\text{th}}$  root of the product is taken:

$$\text{geometric mean } \tilde{x} = \sqrt[n]{\prod_{i=1}^n x_i}$$

An equivalent way to compute the geometric mean is to take advantage of natural logarithms. Defining  $y$  as the natural log of  $x$  [ $y = \ln(x)$ ], the geometric mean is the anti-log (exp) of the arithmetic mean of  $y$ :

$$\text{geometric mean } \tilde{x} = \exp(\bar{y}), \text{ where } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Because the geometric mean is based on log values and the log transformation tends to draw extreme values toward the center of the data, the geometric mean is more "robust" than the arithmetic mean; the geometric mean is less influenced by outliers and consequently is a better

representation of the data distribution. In Premier performance reports, LOS is reported as a geometric mean.

### **3.5 Out of Range Predictions**

Predicted lengths of stay of 100 days or more are considered invalid and are excluded from length of stay analyses.

## **IV. Data Source and Model Calibration**

Premier employs three main data sources: MedPAR, All-Payor State data, and private client data. All three datasets are calibrated separately.

### **4.1 MedPAR Data**

MedPAR consists of approximately 12 million inpatient visits that are covered by Medicare each year. These fiscal year data are generally consistent and updated annually with roughly a one-year lag time. (e.g. Fiscal year 2004 data were available at the end of 2005.) MedPAR covers all U.S. states and territories and is publicly available. Unsurprisingly, many research projects and publications are based on MedPAR. MedPAR covers around one-third of all hospital inpatients, almost all of which are 65 and older. Consequently, some specialties such as Pediatrics and Obstetrics are practically absent.

### **4.2 All-Payor State Data**

All-Payor State data include all inpatients regardless of payor type or other restrictions, thus providing an advantage over MedPAR. Additionally, All-Payor State data contain a larger volume: roughly 20 million records from around 2700 hospitals. Despite these advantages, the data set has limitations. The most noticeable of these is that the data are less geographically representative. All-Payor State data come from fewer than 20 states located mostly on the coasts. In addition to this handicap, the data set lacks a continuum of data for each of the states, since changing regulatory laws often affect the availability of states' data from year to year. This lack of continuous data can severely limit the feasibility of longitudinal studies. Additionally, because State data is released by individual states with their own data specifications, the data are often inconsistent across states. As a result, All-Payor State data require significant internal resources to validate and improve its quality. The two-year lag time in release prevents All-Payor State data from being chosen as the model's calibration database, because the standards of hospital care are in constant flux (reflected in part by new codes appearing every year to reflect changes in diagnosis, procedure, DRG, etc). Despite the aforementioned limitations, All-Payor State data remains a good choice for hospital ranking because of its volume and completeness of disease segments. It also serves as a reference data set for Premier's private data.

### **4.3 Private Client Data**

In addition to the public data sets, Premier collects private data from clients. Client data are submitted in compliance with Premier's Master Data Specifications (MDS), ensuring its consistency and quality. The data are updated frequently with three to six months lag and offer much richer content that allows exploration of new model specifications. Annually, there are around two million records from 140 hospitals dispersed in 35 states. Because the client base is continually changing, the number of hospitals and records may fluctuate each year. The quality and richness of the client data make it an ideal calibration database despite its significantly smaller size than the two public data sets.

### **4.4 Model Selection for Private Client Data**

To avoid overfitting, CareScience's model calibration employs Stepwise Selection for private client data with critical significance set at 0.10. Variables are added to the model one at a time with the computational program selecting the variable whose F statistic is the largest and also meets the specified critical significance. After a variable is added, the stepwise method inspects all variables in the model and deletes any whose F statistic fails to meet the specified significance threshold. Once the check is made and the necessary deletions accomplished, another variable is added to the model. This process effectively reduces the possibility of multicollinearity caused by highly correlated independent variables. The stepwise process ends when the F statistics for every variable outside the model fail to meet the significance threshold while the F statistics for every variable within the model satisfy the significance criterion.

Due to the selection criteria, the number of selected independent variables ranges from several to dozens, depending on the disease. The R-Square of the model may be smaller than that of a full model without restriction but are far more robust than an overfitted full model. For out-of-sample predictions, robust parameter estimates generate more reliable risk scores.

Chronic conditions and comorbidities are restricted to positive-only parameter estimates due to their clinical attributes.

### **4.5 Model Selection for Public Data**

Public data sets are always calibrated on themselves. Because their parameter estimates are not used for out-of-sample predictions, a full model is preferred as it provides a higher R-Square.

## **V. Performance Assessment**

Provider performance can be assessed for virtually any patient grouping (e.g. hospital-level, physician-level, principal diagnosis, DRG, procedure, etc.) through aggregation and comparison of the model's raw and risk complication rates. Positive deviations, as calculated below, indicate worse than expected (average) performance while negative deviations indicate better than expected (average) performance.

$$\text{Ln LOS Deviation}_i = \frac{1}{n} \left( \sum_{i=1}^n \text{Raw Rate}_i - \sum_{i=1}^n \text{Risk Rate}_i \right), \quad i = 1, 2, \dots, n$$

where  $n$  is the number of patients in the  $i$ th patient group.

Statistical significance tests can be used to determine whether complication deviations indicate reliable areas for opportunity. Premier performance reports flag deviations significant at 75% and 95% confidence levels.

**Figure 1: Computing Ln LOS Risk Rates and Deviations Example**

**Principal Diagnosis: Simple Pneumonia (486)**  
**Sample Patient Characteristics**

Patient	Dependent Variable	Independent Variables							
	Raw Ln LOS	Age	Age <sup>2</sup>	Gender Male=0 Female=1	Income	Comorbidities Severity A	Comorbidities Severity B	Chronic Condition 250.x2	...
1	0.69	42	1764	1	\$40,000	2	1	0	...
2	1.61	55	3025	1	\$55,000	1	2	0	...
3	2.20	63	3969	0	\$39,000	4	3	1	...
4	0.69	66	4356	0	\$25,000	3	3	1	...

**Principal Diagnosis: Simple Pneumonia (486)**

Independent Variable	Coefficient (Parameter Estimate)
Age	0.00860
Age <sup>2</sup>	-0.00005842
Gender	0.05558
Income	-0.000001080
Comorbidities Severity A	0.09464
Comorbidities Severity B	0.04309
Chronic Condition 250.x2	0.01897
...	...

**Patient-Level Risk:**

$$\begin{aligned} \text{Ln LOS Risk} &= b_0 + b_1(\text{age}) + b_2(\text{age}^2) + b_3(\text{gender}) + b_4(\text{income}) + \dots \\ &= 0.626 + 0.00860 (\text{age}) - 0.00005842 (\text{age}^2) + 0.05558 (\text{gender}) - 0.000001080 (\text{income}) + \dots \\ &= 0.626 + 0.00860 (42) - 0.00005842 (1764) + 0.05558 (1) - 0.000001080 (40,000) + \dots = 0.45 \end{aligned}$$

➤ Patient 1 has a Ln LOS risk of 0.45 (1.57 days)

(Continued next page...)

**Provider-Level Risk:**

Patient	Raw Ln LOS	Risk Ln LOS
1	0.69	0.45
2	1.61	1.34
3	2.20	1.60
4	0.69	0.98
5	2.40	2.35
6	2.77	2.01
<b>SUM</b>	<b>10.36</b>	<b>8.73</b>

Ln LOS Raw Rate =  $10.36/6 = 1.73$

LOS Raw Rate =  $\exp(1.73) = 5.6$  days

Ln LOS Risk Rate =  $8.73/6 = 1.46$

LOS Risk Rate =  $\exp(1.46) = 4.3$  days

➤ **LOS Deviation =  $5.6 - 4.3 = 1.3$  days (excess LOS)**

## Appendix A – Semilog Modeling

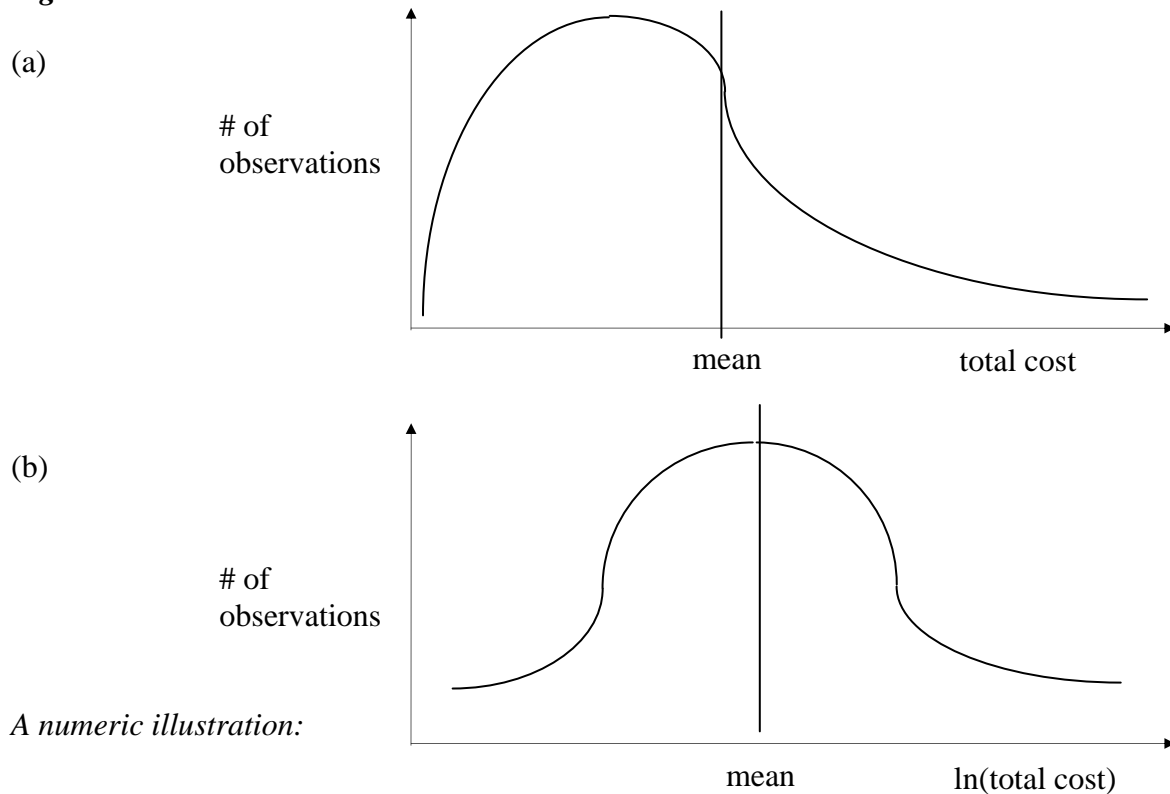
Certain outcome measures, notably costs and length-of-stay (LOS), are distributed with a rightward (positive) skew, as depicted below in Figure 1(a). Applying linear regression to models with skewed dependent variables gives rise to a number of pathologies, including inefficient, often biased, parameter estimates and predictions outside logical bounds, such as negative values for LOS and costs. When outcome measures are not symmetrically distributed, analysis of performance can be disproportionately influenced by outliers and special or extreme cases. This phenomenon can require a manual procedure for identifying and removing outliers, a subjective technique at best.

A more robust solution is to take the natural log of the dependent variable, which results in an approximately symmetric distribution and contracts the outliers inward toward the center of the data, as shown in Figure 1(b). It also ensures that all predicted values will be positive. (No matter how negative the log value is, taking the anti-log to restore the values will guarantee that they are positive.)

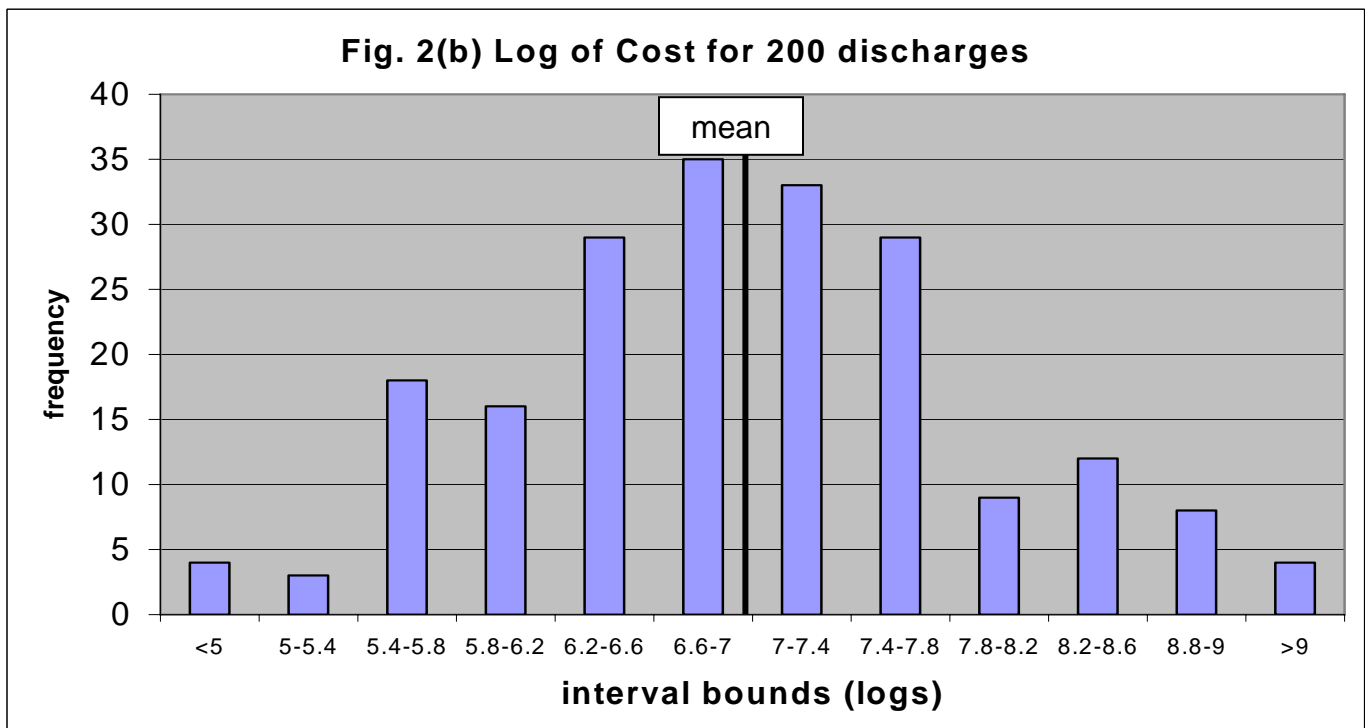
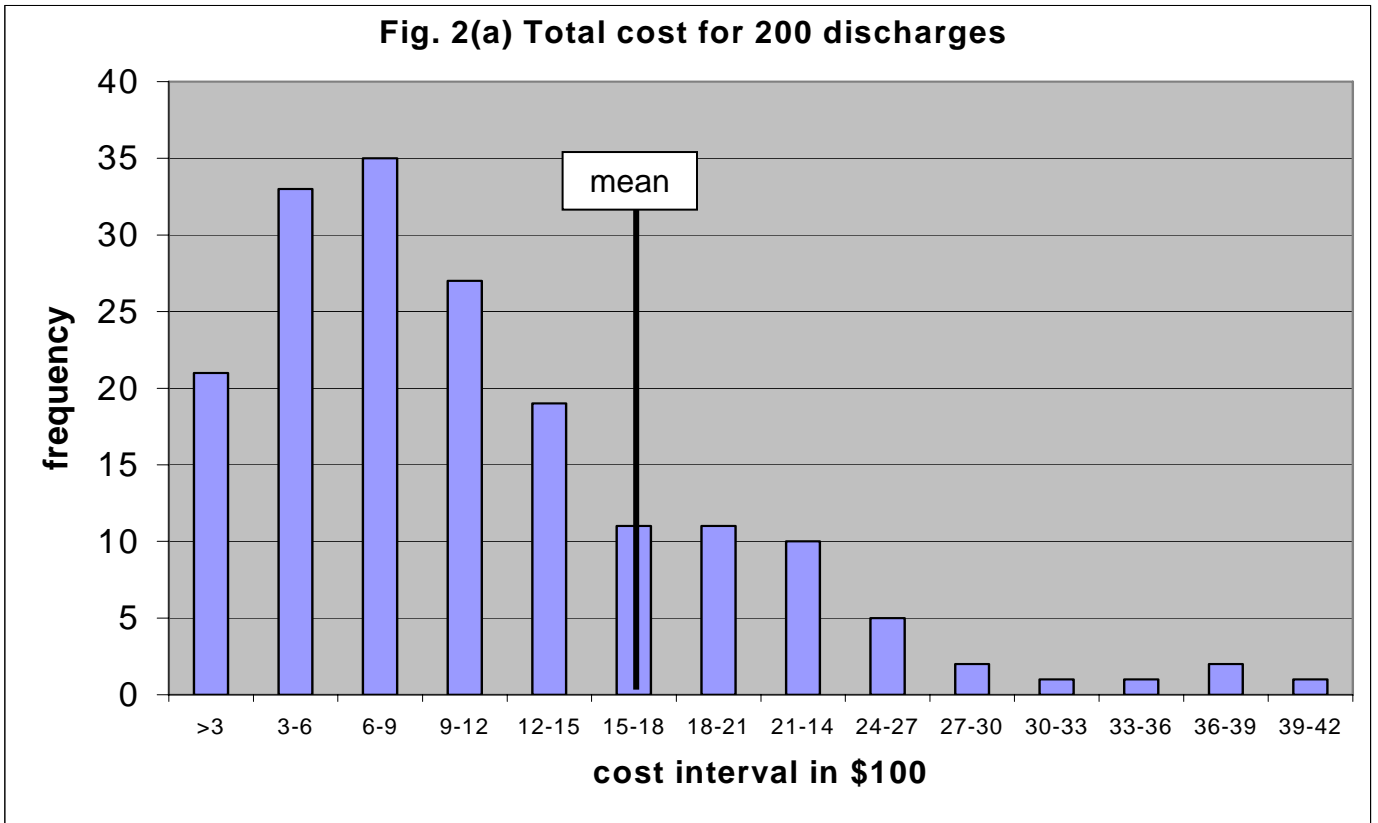
We conducted a systematic review of non-adverse outcome measures – LOS, charges, and costs – by three-digit ICD-9 code to monitor the positive skew and measure its magnitude. In symmetric distributions two measures of central tendency, geometric mean and arithmetic mean (see below), are equal. As the skew increases in unimodal distributions the ratio of the arithmetic mean to the geometric mean grows from unity.

*To illustrate skew:* Total cost is skewed right but the natural log of total cost –  $\ln(\text{cost})$  – is approximately symmetrically distributed, therefore using linear regression to forecast  $\ln(\text{cost})$  will result in much better estimates with smaller error.

**Figure 1**



Depicted below is the total cost frequency distribution for a sample of 200 hospital discharges. It displays the characteristic positive skew (skew coefficient = 2.6).



Transforming cost from Fig. 2(a) by taking the natural log gives the frequency distribution in Fig. 2(b), which exhibits the typical symmetric bell shape of the normal distribution. The **arithmetic** mean cost is marked on the first (skewed) frequency histogram, which in this illustration is \$1670. The mean of the log(cost) is marked on the second histogram at 6.95. Taking the anti-log of this value yields the **geometric** mean equal to \$1043, which is much closer to the mode of the original (untransformed) histogram. The pronounced positive skew in the original cost distribution guarantees that the arithmetic mean is much larger than the geometric mean, which tends to pull back the extreme values in the upper tail. In this illustration the ratio of the arithmetic mean to the geometric mean is \$1670/\$1043 = 1.60.

$$\text{raw arithmetic mean } \bar{x}_{jl} = \frac{1}{n_{jl}} \sum_{ik}^{n_{jl}} x_{ijkl}$$

$$\text{raw geometric mean } \exp(\bar{y}_{jl}) \text{ where } \bar{y}_{jl} = \frac{1}{n_{jl}} \sum_{ik}^{n_{jl}} y_{ijkl}$$

$$\text{risk value } \hat{x}_{ijkl} = \begin{cases} \exp(\hat{y}_{ijkl}) & \text{for all complete cases (including zeros)} \\ \exp(\bar{y}_{jl}) & \text{for all incomplete cases} \end{cases}$$

$$\text{arithmetic mean risk } \bar{\bar{x}}_{jl} = \frac{1}{n_{jl}} \sum_{ik}^{n_{jl}} \bar{x}_{ijkl}$$

$$\text{geometric mean risk } \exp(\bar{\hat{y}}_{jl}) \text{ where } \bar{\hat{y}}_{jl} = \frac{1}{n_{jl}} \sum_{ik}^{n_{jl}} \hat{y}_{ijkl}$$

where  $x_{ijkl}$  = patient.total\_charges, patient.comparative\_costs, and patient.length\_of\_stay

$$y_{ijkl} = \ln(x_{ijkl})$$

and  $\bar{\bar{x}}_{ijkl}$  = ln(total\_charges) risk, ln(comparative\_cost) risk, and ln(length of stay) risk

$i$  = patient (each row in the patient table)

$j$  = provider or grouping

$k$  = icd9 diagnosis (3 digit)

$l$  = outcome (length of stay, charges, cost)

$n$  = all observations including zeros