CareScience Risk Assessment Model: Hospital Performance Measurement

E. A. Kroch
M. Duan

March, 2008

Table of Contents

I. Introduction

1.1 Quality of Care

The ongoing debate over how to measure inpatient quality of care has, from time to time, focused on different aspects of Avedis Donabeidan's 1965 "structure-process-outcome" triad[1]. The focus on observable outcomes has shifted conceptualizations of hospital quality towards the ideas of Joe Juran and others[2] who define manufacturing quality as the absence of defects. By this definition, deaths, complications, unusually long hospital stays, unscheduled ICU admissions and other sentinel events that are deemed universally negative (or nearly so) signal the extent to which a care provider deviates from "good quality." Despite disagreements over the best approach to measure inpatient quality, most investigators accept that treatment quality is the absence of adverse events.

1.2 CareScience Quality Measures

CareScience measures three adverse outcomes to capture quality of care: mortality, complications, and morbidity. Each of these measures possesses its own strengths and weaknesses and is better suited to certain patient populations and applications. Together they provide a complementary and effective means for screening quality improvement opportunities.

1.2.1 Mortality

Mortality is perhaps the most widely used quality measure, since the occurrence of death seems unambiguously a defect of care. Inpatient mortality rates are also easily observed by simply counting deaths from discharges. While mortality has advantages as a quality of care indicator, it possesses drawbacks. The approach of simply counting deaths from discharges can inadvertently mask "true" mortality rates, which may be disguised by discharge policies. For instance, inpatient mortality rates can be reduced by transferring the most severely afflicted patients to another acute care facility, skilled nursing home, or hospice. Mortality rate is also prone to wide variation across diseases, rendering them irrelevant for certain populations for quality analysis. In populations where death is very rare (e.g. kidney and ureter calculus) or largely expected (e.g. admitted with DNR), mortality becomes a less meaningful quality measure.

1.2.2 Complications

Complications are a relevant quality measure for most patient populations. They range from trivial to significant and can result in increased lengths of stay or unscheduled treatments. The challenge with measuring complications is the difficulty observing them

[1] Donabedian A. Evaluating the Quality of Medical Care, Milbank Quarterly,1966; 44:166-203.
[2] Juran JM et al. Quality Planning and Analysis: From Product Development Through Use. McGraw-Hill Series in Industrial Engineering and Management Science, 1993.

and their dependence on good documentation and coding consistency. Traditionally, complications have been tracked using chart reviews during which clinicians pull and review individual patient charts. These time-consuming reviews are expensive and laborious and consequently unsuitable for large scale data analysis. CareScience has developed a unique decision-theoretic complication tracking model that uses comorbidity adjusted complication indices (CACI) to distinguish complications from comorbidities. This model assumes a nonstandard definition of complications, defining them as conditions that arise during a patient's hospital stay. By this construction, complications do not necessarily imply iatrogenic events or physician negligence.

Validation studies have shown that the comorbidity-adjusted risk (CACR) model yields similar results to chart reviews at the aggregate level, particularly for surgically treated patients[3].

1.2.3 Morbidity

Morbidity is defined as the severity of a patient's complications. Within the CareScience model, morbidity is divided into 5 severity levels, A-E, which follow a Likert scale. Complications in category D and E are considered 'morbid' complications, and they are separately measured under the label of 'morbidity'. They often result in temporary impairment, unscheduled ICU admission, and significant increase in length of stay.

1.3 CareScience Efficiency Measures

In addition to quality of care, a hospital's success depends on its financial performance and its ability to manage patients' lengths of stay, costs, and charges efficiently. These economic concerns are particularly relevant in today's climate of rising healthcare costs that routinely exceed the Consumer Price Index (CPI) by several folds. CareScience tracks and examines three efficiency outcomes: length of stay, costs, and charges.

1.3.1 Length of Stay

Length of stay is a commonly used proxy for resource usage, reflecting how efficiently a hospital allocates resources. It is easy to observe and compare across hospitals and offers the advantage of being reliably recorded. Despite these 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. Nevertheless, length of stay is widely accepted as a proxy for efficiency.

1.3.2 Charges

[3] Azimuddin K, Rosen L, Reed JF. Computerized Assessment of Complications after Colorectal Surgery. Diseases of Colon & Rectum 2001; 44:500-505.

Patient charges can serve as a supplementary efficiency measure to length of stay. Charge information can be found in electronic record databases built for billing. Although charge data are readily available, charge practices are generally not comparable across facilities or even departments and therefore require adjustments.

1.3.3 Costs

In measuring hospital performance, charges have become a useful proxy for the "costliness" of care. Unadorned charges, however, are a poor indicator of hospital expenditures, especially at the patient level. Nonetheless, by applying a well defined and hospital-specific Cost-to-Charge-Ratio (CCR), hospitals' reported charges can be adjusted to a dollar amount that more closely approximates the "true" cost of care. These adjusted costs can be used to compare hospitals that do not share the same accounting standards. Appendix A provides a detailed discussion of calculating costs.

1.4 Outcome Evaluation and Risk Adjustment

Inpatient care can be viewed as a process in which the patient's characteristics upon admission (e.g. comorbidities, etc) are the inputs and his health status and financial outcome upon discharge are the outputs. Patient health and financial outcomes are influenced not only by the care process but also by the severity of the patient upon entering the hospital; sicker patients are at higher risk for worse outcomes than patients who are less severely afflicted upon hospital admission. Risk adjustment strives to account for these differences in evaluations of care.

Evaluation of patient outcomes[4] requires benchmarking, wherein a hospital's (or physician's) outcome rates are compared to their expected rates (outcome risks) as suggested by their case-mix. Expected outcome rates for any facility or grouping of patients are based on the characteristics of those patients and a model of the relationship between patient characteristics and outcomes.

"True" patient severity upon admission cannot be directly measured. Instead, it must be inferred from the available data's recorded patient characteristics. Administrative billing data, compiled after treatments are completed, are the data most readily available for these kinds of large scale analyses. These data, optimized for reimbursement purposes and often only secondarily for quality analysis, are prone to inconsistencies and variations in coding that can distort "true" severity. Incomplete coding such as omission of secondary diagnosis codes can erroneously make patients appear healthier. Conversely, overly aggressive coding can give an inflated impression of severity and complications. Often the "real" picture upon admission can not be easily crafted from the records.

[4] "Outcome" is a term of art that includes a range of observable performance measures beyond mortality and morbidity. Length of stay and treatment costs can be considered "outcomes," since they are indicators of efficiency as well as efficacy.

Patient health status at the output end of the care process is not unambiguous. As previously described, discharge policy can affect mortality rate and length of stay. Moreover, discharge codes do not fully reflect patients' health condition. Being discharged home does not necessarily equate to fully recovery of the patient. Some patients may be re-admitted shortly. CareScience clients' data show that about 10% of patients are re-admitted to hospital within 30 days after they have been discharged. Readmission information is unavailable in public data sets.

Due to imperfect information, risk adjustment is subject to limitations. Nevertheless, it remains a widely accepted approach for measuring hospital performance by controlling for patient characteristics and allowing benchmarking to compare apples to apples.

1.5 CareScience Risk Assessment Method

The CareScience risk assessment model is estimated statistically by regression analysis of a defined population of hospital discharges. The population can be restricted to a single hospital over a quarter or can encompass a broad range of hospitals across the country over several years. The more encompassing the population, the broader is the basis for comparison. Basing the benchmark as broadly as possible permits comparisons of hospitals and their physicians across all possible markets and locations. On the other hand, benchmarking on the experiences of a small region restricts comparisons to the hospitals and physicians of that region alone.

CareScience calibrates its model of hospitalization outcomes and related performance measures on the maximum amount of discharges from both private and public sources. The aim is to construct a set of parameter estimates (coefficients or *beta* values) for each of six measures that can be used to predict outcome rates for any set of patients in CareScience products. The six outcomes include in-hospital mortality, major morbidity, complications, length-of-stay, charges, and costs. Predicted rates for these outcomes can be compared to the actual rates to evaluate performance for any set of patients based on their case-mix within a particular facility, service line, diagnosis, age grouping, or treating physician grouping.

This approach has a number of advantages over the alternative method of calibrating the risk assessment on the basis of the individual hospital or hospital system. First, a universally calibrated model makes it possible to generate predicted outcomes (risks) and other relevant statistics without rerunning regressions for every new set of patients. That reduces processing time of new data and even potentially allows real-time processing of discharges Second, outcome risks that are generated by a set of universal *beta* values can be used to compare patients within different facilities and physician groupings regardless of where they practice. Third, any set of discharges can be analyzed, no matter how small, because the model parameters themselves do not need to be estimated from the analysis data set.

II. Risk-Assessment Model Specification

2.1 Functional Form

The purpose of the model is to generate expected or "standard" outcomes (risks) under typical care, based on a patient's characteristics and socioeconomic factors. Patient-level risks for a variety of target outcomes are assessed via a stratified multiple regression model. The model has the following functional form:

$$y_{ijkl} = x_{ijkl}\beta_{kl} + \varepsilon_{ijkl}, \forall ijkl$$

where y_{ijkl} is the value for each outcome l at patient level i and provider j and principal diagnosis k. x_{ijkl} is a vector of patient characteristics and socioeconomic factors. β_{kl} is the marginal effect of the independent variables on the outcome measure, and ε_{ijkl} 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 (BDGs), which are defined in the ICD-9-CM book. There are a total of 142 disease strata and over 800 equations in the model. Details about the stratification can be found in Appendix D (*Technical Details about Model Specification).*

2.2 Dependent Variables

The following outcome measures are modeled separately with their own set of specifications:

Mortality
Complications
Morbidity
Length of Stay
Total Charges
Comparative Costs

2.2.1 Mortality

At the patient level, mortality is captured by discharge disposition. Category '20' is designated for patients who expired. The ccms_exp_flag field in the CareScience database indicates the mortality status (expired/alive) of patients based on their discharge disposition. Patients who were transferred to another acute care facility (with discharge disposition code '02') have an indeterminate mortality value and are consequently excluded from mortality analyses. The mortality risk for these patients is therefore set to 'null.' Exemptions about Mortality outcomes are included in Appendix D (*Technical Details about Model Specification).*

2.2.2 Complications

Complication is defined as the probability of having at least one complication. It is calculated as

$$1 - \prod_{j} (1 - p_{ij}),\ j = 1,2,...,m$$

where m is the number of secondary diagnosis; and p_{ij} is the probability of complication for the jth secondary diagnosis given principle diagnosis i. The probability that any given secondary diagnosis is a complication of a given principle diagnosis is determined ex ante by clinical experts. The method is called Comorbidity-Adjusted Complication Indices (CACI), which is elaborated in the section of Clinical Knowledge Base. The algorithm of calculating complications is described in Appendix D (*Technical Details about Model Specification).*

2.2.3 Morbidity

Morbidity is defined as the probability of having at least one morbid complication. It is calculated in a similar manner as complications, however, only secondary diagnoses rated 'D' or 'E' are included in the calculation. Consequently, it has a smaller value than that of complications. The same rules that govern calculating complications are applied in calculating morbidity.

2.2.4 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 and coded as inpatient, LOS is counted as one day. If a patient stays in the hospital for more than 100 days, the case, as an outlier, is dismissed from LOS analysis.

2.2.5 Total Charges

Total Charges represent the dollar amount charged to a patient during the hospital stay. The field is directly available in both private and public data. If the dollar amount is greater than 500, 000 USD, the case is excluded from both charge and cost analysis.

2.2.6 Comparative Costs

The conversion of charges to costs is a simple matter of multiplying the patient-level total charges from the discharge abstract (typically the UB-92 record) by the facility-specific cost-to-charge ratio. The computation is performed for each individual patient stay in the hospital. To calculate costs, total charges must be recorded (and fall within trimming guidelines). The calculation is performed during early data processing prior to CareScience risk assessment.

2.3 Independent Variables

The following patient characteristics and socioeconomic factors comprise the set of regressors.

Age *(quadratic form)*
Birth weight (*quadratic form, for neonatal model only*)
Sex *(female, male, unknown)*
Race *(white, black, asian-pacific islander, unknown)*
Income *(median household income within a zip code reported by US Census Bureau)*
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)*
Principal diagnosis *(terminal or three digit ICD-9-CM code, where statistically significant)*
CACR[5] **comorbidity scores** *(count of comorbidities within each of five severity categories on the CACR Likert scale)*
Defining diagnosis *(three digit ICD9-CM code for neonatal model only)*
Cancer status *(benign, malignant, carcinoma in situ, history of cancer, derived from secondary diagnoses)*
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)*
Valid procedure *(terminal ICD9-CM procedure codes, where clinically relevant and statistically significant)*
Time trend factor *(to control for inflation specific to each disease in the inpatient hospital setting, derived from discharge date, for Cost and Charge model only)*
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)*
Admission type *(Emergency, Urgent, Elective, Newborn, Delivery, Unknown/Other)*
Payor class *(Self-pay, Medicaid, Medicare, BC/BS, Commercial, HMO, Workman's Compensation, CHAMPUS/FEHP/Other Federal Government, Unknown/Other)*
Discharge disposition *(Home or Self Care, Short-term General Hospital, Skilled Nursing Facility, Intermediate Care Facility, Other Type of Institution, Home under Care of Organized Home Health Service, Left against Medical Advice, Discharged Home on IV Medications, Expired, Unknown/Other)*
Facility type *(Acute, long-term, Psych.)*

Risk factors used in the CareScience risk assessment model are tailored to specific patient subpopulations and outcomes. The 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)
- time trend (controls inflation for costs and charges)
- discharge disposition (excluded in mortality analyses)

[5] 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.

The following table summarizes the independent variables for specific outcomes and subpopulations. Some independent variables are defined and selected according to clinical relevance, and some are transformed through mathematic method. The methods are elaborated in Appendix D (*Technical Details about Model Specification)*.

Model Specification Summary

Note: BW=Birth Weight, AS=Admission Source, AT=Admission Type, PC=Payor_Class, DD=Discharge_disposition, Comor.=Comorbidities
VP=Valid Procedure, TT=Time_Trend, Dx=Diagnosis, T=Terminal Digit Code, 3=Three Digit Code

		Dependent Variable																
Outcome	**Disease Groupings**	Age	BW	AS	AT	PC	DD	PDx	Comor.	Chronic	Cancer	Defining Dx	VP	TT	Sex	Race	Income	Distance
Mortality	Major Dx	Y		Y	Y	Y		T	Y	Y	Y		Y		Y	Y	Y	Y
Mortality	Broad Dx Group	Y		Y	Y	Y		3	Y	Y	Y		Y		Y	Y	Y	Y
Mortality	Normal Neonates			Y	Y	Y						Y			Y	Y	Y	Y
Mortality	Immature Neonates		Y	Y	Y	Y						Y			Y	Y	Y	Y
Mortality	Organ Transplants	Y		Y	Y	Y			Y		Y				Y	Y	Y	Y
Complication	Major Dx	Y		Y	Y	Y	Y	T	Y	Y	Y		Y		Y	Y	Y	Y
Complication	Broad Dx Group	Y		Y	Y	Y	Y	3	Y	Y	Y		Y		Y	Y	Y	Y
Complication	Normal Neonates			Y	Y	Y	Y					Y			Y	Y	Y	Y
Complication	Immature Neonates		Y	Y	Y	Y	Y					Y			Y	Y	Y	Y
Complication	Organ Transplants	Y		Y	Y	Y	Y		Y		Y		Y		Y	Y	Y	Y
Morbidity	Major Dx	Y		Y	Y	Y	Y	T	Y	Y	Y		Y		Y	Y	Y	Y
Morbidity	Broad Dx Group	Y		Y	Y	Y	Y	3	Y	Y	Y		Y		Y	Y	Y	Y
Morbidity	Normal Neonates			Y	Y	Y	Y					Y			Y	Y	Y	Y
Morbidity	Immature Neonates		Y	Y	Y	Y	Y					Y			Y	Y	Y	Y
Morbidity	Organ Transplants	Y		Y	Y	Y	Y		Y		Y		Y		Y	Y	Y	Y
Length of Stay	Major Dx	Y		Y	Y	Y	Y	T	Y	Y	Y		Y		Y	Y	Y	Y
Length of Stay	Broad Dx Group	Y		Y	Y	Y	Y	3	Y	Y	Y		Y		Y	Y	Y	Y
Length of Stay	Normal Neonates			Y	Y	Y	Y					Y			Y	Y	Y	Y
Length of Stay	Immature Neonates		Y	Y	Y	Y	Y					Y			Y	Y	Y	Y
Length of Stay	Organ Transplants	Y		Y	Y	Y	Y		Y		Y		Y		Y	Y	Y	Y
Cost	Major Dx	Y		Y	Y	Y	Y	T	Y	Y	Y		Y	Y	Y	Y	Y	Y
Cost	Broad Dx Group	Y		Y	Y	Y	Y	3	Y	Y	Y		Y	Y	Y	Y	Y	Y
Cost	Normal Neonates			Y	Y	Y	Y					Y		Y	Y	Y	Y	Y
Cost	Immature Neonates		Y	Y	Y	Y	Y					Y		Y	Y	Y	Y	Y
Cost	Organ Transplants	Y		Y	Y	Y	Y		Y		Y		Y	Y	Y	Y	Y	Y
Charge	Major Dx	Y		Y	Y	Y	Y	T	Y	Y	Y		Y	Y	Y	Y	Y	Y
Charge	Broad Dx Group	Y		Y	Y	Y	Y	3	Y	Y	Y		Y	Y	Y	Y	Y	Y
Charge	Normal Neonates			Y	Y	Y	Y					Y		Y	Y	Y	Y	Y
Charge	Immature Neonates		Y	Y	Y	Y	Y					Y		Y	Y	Y	Y	Y
Charge	Organ Transplants	Y		Y	Y	Y	Y		Y		Y		Y	Y	Y	Y	Y	Y

2.4 Semi-log Model

Length-of-Stay (LOS), Costs, and Charges are 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 B (*Semilog Modeling)*

Geometric vs. 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):

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^{th} root of the product is taken:

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:

geometric mean $\tilde{x} = exp(\bar{y})$, 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 the Care Management System tool, Length-of-Stay (LOS), Costs, and Charges are reported as geometric means.

III. Model Calibration

3.1 Data Source

Three data sources are employed in CareScience risk model calibration. They are MedPAR, All-Payor State data, and private data.

MedPAR consists of approximately 12 million inpatient visits that are covered by Medicare each year. MedPAR covers all U.S. states and territories and is publicly available. Many research projects and publications are based on MedPAR. CareScience acquires the pre-processed data annually from Solucient. The time range of the data is based on CMS's fiscal year, which contains the fourth quarter of the previous year and the first three quarters of the current year. MedPAR data is generally available with one year lag time. (e.g. Year 2004 data were available by the end of 2005.) MedPAR covers around one-third of all hospital inpatients; and almost all of its patients are 65 plus. Consequently, some specialties such as Pediatrics and Obstetrics are practically absent.

All-Payor State data includes all inpatients regardless of payor type or other restrictions, thus providing an advantage over MedPAR data. Additionally, All-Payor State data contains 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 comes 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. CareScience acquires the pre-processed data annually from Solucient. There is usually a two-year lag. Because State data is released by individual states with their own data specification, the data is often inconsistent across states. As a result, All-Payor State data requires significant internal resources from Data Management Group and Research to validate and improve the quality. The lag time in release also 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 in order 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 to CareScience's private data.

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

3.2 Missing Outcomes and Independent Variables

As with most large databases, some records may lack one or more data elements. When outcome is missing, the record will be automatically removed from the analysis. When one or more

independent variables are missing, the algorithm will set risk score to null, or explore feasible alternatives.

Principal Diagnosis, Age, and Birth Weight (for immature newborns) are mandatory elements in the risk assessment model. They are considered essential, non-replaceable risk factors. Missing any of them will result in exclusion from the risk assessment. Time_Trend is a mandatory field to predict Charges and Costs. Omission of this value results in 'null' risk scores for Cost and Charge outcomes for these specific cases.

For most categorical variables, such as Admission Source, there is an 'Unknown' category designated for unrecognizable or missing values. Given 'Unknown' is due to random error, the missing value is most likely the category with the highest frequency. For example, ER is the most common admission source. If a patient's admission source is missing, ER would be the most likely admission source. In risk modeling, the most common category is often used as the reference group. Grouping the 'Unknown' category with the most common category (the reference group) is thus justifiable, however, the high portion of 'Unknown' values risk 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, 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 reference group.

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 shall be set to one, and the national median income will be applied.

3.3 SAS Programming

Since 2003, model calibration has been executed by SAS-based programs that are created and maintained by the Research Dept. The SAS-based programs replaced the previously used CareScience Regression Language (CRL), which was maintained by the Software Engineering Dept.

3.3.1 Data Transforming

CareScience's database is on Oracle platform. The database schema and table structure are designed for specific product and its related data processing tool. It is essential for Research to pull out key data elements and transform them into SAS-recognizable data sets. The technical details are elaborated in Appendix E *(Technical Details about SAS Programming)*

3.3.2 Model Selection

Not all variables carry the same weight. Some variables may have little impact on risk scores. Some variables may have impact on only specific outcome. CareScience's model calibration employs Stepwise selection to identify significant variables at 0.10 level, which is close to the upper limit that SAS recommends. Alternative selection options are discussed in Appendix E *(Technical Details about SAS Programming).*

With Stepwise option, variables are added to the model one at a time with the 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. Only after the check is made and the necessary deletions accomplished can another variable be added to the model. This process effectively reduces the possibility of multicollinearity issue, which is 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 in the model satisfy the significance criterion. Alternatively, the process ends when the next variable to be added to the model is the one just deleted from it.

Due to the selection criteria, the number of selected independent variables ranges from several to dozens, depending on outcome and disease. The R-Square of the model may be lower than a full model without any restriction. But the parameter estimates from the selected model are far more robust than an over-fitted full model. In out-of-sample prediction, robust parameter estimates generate reliable risk scores.

Public data sets are always calibrated on themselves. No parameter estimates from their calibration are used to assess other data sets. Therefore, a full model is preferred because it provides higher R-Square.

Regardless of its significance Time Trend is exempt of stepwise selection. It is forced into the Cost and Charge models by an 'include' option in the program. Internal studies have demonstrated that Time Trend is a strong predictor for measuring inflation rate for most diseases. The average inflation rate is around 8.8%. For the 142 disease strata, Time fails to meet the 0.10 significant level in only one instance.

Chronic conditions and Comorbidities are restricted to positive-only parameter estimates according to their clinical attribute.

3.3.3 Macro Function

It is a monumental task to manipulate millions of records with hundreds of fields and subsequently run more than 800 regression equations through SAS programming. In the current model, regression equations are specified distinctly depending on disease and outcome. Regression coefficients and covariance matrices are then reshaped to fit the structure of the

analytic program's *beta* tables. Given the enormity of the process, it is essential to implement SAS Macro language, which streamlines the programming and makes it at least partially an automatically executed program. The technical details about the Macro processing are elaborated in Appendix E *(Technical Details about SAS Programming)*.

*3.*4 *Beta Tables*

Beta tables include a coefficients table (*Beta* table), a covariance table, and a regression information table. All three tables are initially generated by the model calibration process and then are manipulated to fit specific formats. These tables comprise the key components of CareScience's risk assessment tool.

3.4.1 Regression Information Table

As its name suggests, the regression information table summarizes the results of model calibration. For each model equation, regression information reports R-square, root mean square error (RMSE), number of selected independent variables, and the number of valid records in the calibration data.

R-square varies substantially across outcome and disease. Of the outcomes measured, mortality often suffers from consistently low R-square values, a susceptibility that can be attributed to two reasons. First, expiration is rare among patient discharges. At the hospital level, the mortality rate hovers around 2-3 percent. Most expiration cases are concentrated in a few high-risk diseases, such as Septicemia, AMI, and Lung Cancer. Furthermore, when mortality is an infrequent occurrence within a low-risk disease, it is more difficult to predict. The second reason is that the model calibration relies on claim data, which does not cover all clinical factors. Because the data are designed for billing purposes, it is unsurprising that financial outcomes such as LOS, Costs, and Charges have higher R-squares than for mortality. (R-square for the efficiency outcomes, LOS, Costs, and Charges can be as high as .70).

RMSE is used, along with the Covariance Table, to calculate standard error associated with predicted risk at the patient level.

3.4.2 *Beta* Table

Beta table includes all coefficients that are significant at 0.10 level. Coefficient (*Beta*) can be interpreted as the risk factor's marginal effect upon risk score. By applying a set of corresponding coefficients, risk score can be calculated for each outcome at patient level. This processing is handled within Data Manager (formerly VIMR and CRL, or CIA).

To categorical variables, each coefficient is corresponding to one category. The coefficient shows the difference between that category and the reference category. For example, Hospital Transfer (04) is one category of Admission Source. The coefficient for Hospital Transfer can be

interpreted as the change of risk relative to the reference group ER (07), the most common category of Admission Source.

Regarding valid procedures and chronic conditions, each code is treated as a separate variable in the model. Their coefficients can be interpreted as the risk difference between patients having and not-having that code.

For each model equation, the number of coefficients corresponds in both the Covariance and the Regression Information Table. This feature can be used as QA control during the manipulation of *beta* tables.

3.4.3 Covariance Table

The covariance table is derived from the covariance matrix of coefficients. Due to Data Manager's requirements, the matrix is reshaped into a two-way table, consisting of RowName and ColumnName. The covariance table contains millions of values and is much larger than the *Beta* table. Due to the legacy of an earlier tool (CRL), the correlation matrix of coefficients $(X'X)^{-1}$ is actually used to compute standard error. Consequently, the covariance matrix has been transformed into correlation matrix in the 'covariance' table.

3.5 Model Implementation Test

Two steps are involved in a model implementation test. The first is a technical test, which confirms that Data Manager (formerly VIMR and CRL, or CIA) correctly implements the beta tables to compute risks and the associated standard errors at the patient level. The second is a clinical review, which validates outcome reports from the clinical perspective.

3.5.1 Technical Test

During model calibration, SAS regression procedures generate predicted values of dependent variables and standard errors for all observations if options are correctly specified. The SAS output term 'P' (predicted) is equivalent to the risk in the Data Manager output while the SAS output term 'STDI' is equivalent to the standard error in Data Manager. The SAS-generated data set can therefore serve as a convenient reference for the technical model implementation test.

The technical model implementation test is conducted by Software Engineering and Quality Assurance Team, with Research and Data Management Group assisting them in assembling the data set. Research also consults on technical requirements, ensuring that the test data set has full coverage of all 142 disease groups (ccms_crl_group_by).

The test data may be selected from one hospital or one hospital system that has a sufficient number of cases for all disease groups. A more conservative approach, however, is to randomly extract data from the entire calibration database by disease group. The latter method guarantees full coverage of the data at a global level. It is recommended that at least 1000 cases are pulled

for each disease stratum, however, for a few strata (e.g. DRG103), this may not be possible. In these situations, all cases in the disease stratum should be included in the testing data set.

It may take a couple of months, or longer, to complete the test. Research plays an essential role in the process. When there is a discrepancy between SAS score and analytics score, finding the reason or reasons requires extensive knowledge about the methodology as well as thorough understanding of database schema.

3.5.2 Clinical Validation

The main purpose of risk-assessment is to differentiate patients based on their individual characteristics. Statistically, including more independent variables often increases R-square and creates more distinctive risk scores at the patient level. However, a model with the highest R-square may not have the best performance when coefficients are implemented in an out-of-sample prediction and risk scores are extrapolated. For instance, allowing more procedure codes in the model will certainly increase the model fit, but it may also result in inflated risk scores among patients treated by multiple significant procedures. From a statistical perspective, a few outliers are acceptable and don't exert much influence at the aggregate level, however, they may become a serious issue when grouped in CareScience front-end reports by certain criteria (e.g. DRG), causing their risk scores to deviate far beyond conventional clinical wisdom. A clinically based review can help identify problems such as this. Since the front-end tool allows users to select virtually any combination of patients, extensive reviews by the Consulting team are required to avoid issues of clinically unfounded risk scores.

Clinical reviews are also necessary to improve the quality of CareScience's clinical knowledge base used in the model. This knowledge base includes the comorbidity-adjusted complication indices (CACI), chronic condition designations, diagnosis morbidity designations, valid procedure designations, etc. This information has been gathered over the course of many years and is continually reviewed and updated by internal and external clinical experts. Despite these efforts, gaps in the knowledge base exist along with areas where the data contradict itself due to changes in treatment or coding practices. Clinical validation of the model performance provides an opportunity to identify some of these problems and accordingly upgrade our clinical knowledge base.

IV. Clinical Knowledge Base

It would be a mistake to characterize the CareScience risk model as purely a statistical model. Clinical knowledge plays a key role from the beginning of data processing to the end of risk assessment. The following section systematically examines the key components comprising the CareScience clinical knowledge base.

4.1 Comorbidity-Adjusted Complication Indices (CACI)

CareScience uses a decision-theoretic model called Comorbidity-Adjusted Complication Risk (CACR) to track complications. This model assumes a nonstandard definition of complications, defining them as conditions that arise during a patient's hospital stay. By this construction, complications do not necessarily imply iatrogenic events or physician negligence.

The CACR model is based on the assumption that most secondary diagnoses do not occur purely as comorbidities or as complications. Instead, some proportion of each of these recorded secondary diagnoses represent conditions that emerge during a hospital stay while the remaining proportion represent conditions that were present when the patient was admitted (i.e., comorbid conditions).

A comorbidity-adjusted complication index (CACI) is the probability that a given secondary diagnosis is a complication (condition developed during a patient's hospital stay) for a patient with a specific 3-digit ICD-9 principal diagnosis. For example, the CACI for a secondary diagnosis of urinary tract infection with a principal diagnosis of simple pneumonia is 90%, indicating that for 90% of patients with this principal-secondary diagnosis pair, the urinary tract infection emerged during their inpatient stay. For the remaining 10% of patients, the urinary tract infection was present at the time of admission. CACI exists for most common principal-secondary diagnosis pair combination and are assigned by Delphi[6] panels of physicians.

Because CACIs are probabilities, they can only operate effectively on aggregated data where they provide estimated complication rates. CACIs cannot be used pinpoint which patients have specific complications within a given population.

CACIs are periodically reviewed and reevaluated as a result of changes in medical practice, contestations, or additional information such as empirically derived "present on admit (POA)" data.

4.2 Diagnosis Morbidity

Diagnosis is one of the most important factors used to measure patient risk. CareScience risk model is stratified primarily according to principal diagnosis at the three-digit level. Principal diagnosis alone is able to explain a great portion of risk variations across all patient population. When principal diagnosis is the same, secondary diagnoses provide critical information to differentiate patient characteristics. Two clinical outcomes, Complications and Morbidity, are derived from secondary diagnoses. Among the independent variables, comorbidities and chronic

[6] The Delphi method, named for the famed Greek oracle, consists of several "rounds" of input and feedback. In the first round, participants are asked for their independent judgment. Once all responses are gathered, the mean response is posted. In the second round, participants are given the chance to change their response in light of the group's mean. The revised mean is posted to begin the third round, and so on.

conditions are also derived from secondary diagnoses. Because diagnoses have varying clinical significance, a five-level Likert scale was created to denote the severity or morbidity of each diagnosis.

The morbidity Likert scale levels range from 'A' to 'E' with 'A' reserved for conditions that are least severe. Secondary diagnoses in category 'A' possess minimal or no impact on patient risk. Typical diagnosis codes in this category include Headache (ICD9 Dx 7840), Backache NOS (ICD9 Dx 7245), and Diarrhea NOS (ICD9 Dx 78791). Category 'B' and 'C' denote mild conditions that may impact patient risk and the course of treatment. These conditions may increase length of stay and cost of treatment. Most common secondary diagnoses fall into categories 'B' or 'C.' Diagnoses that are classified into category 'D' and 'E' are truly severe conditions (e.g. Oclsn, cer artery NOS w/infarction (ICD9_Diag_43491)) and sometimes life-threatening. These conditions may substantially increase probability of expiration and length of stay; patients with these conditions may require additional rescue treatment, and their costs of treatment may spike.

Morbidity designations are always assigned at the terminal digit level of a diagnosis code. For instance, uncomplicated type I DM (ICD9 Dx 25001) is classified into category 'B' while type I DM w/neuro (ICD9 Dx 25061) is considered more severe, thus designated into category 'C.'

The Diagnosis Morbidity Table was originally created to measure morbid complications (Morbidity: complications with morbidity designations of 'D' or 'E'). In 2003, the table was expanded to include all common secondary diagnoses at the terminal digit level. With this expansion, the comorbidity score was broken into five categories, corresponding to the morbidity Likert scale. This modification has substantially improved the model performance. There are currently 514 secondary diagnosis codes in the table that account for about 80% of all secondary diagnoses. The less common diagnosis codes are not dropped from analysis but instead grouped into category 'U,' which stands for 'Unspecified.' As the morbidity assignments' normal distribution suggests, category 'U' diagnoses tends to share similar characteristics as the most common categories 'B' and 'C.'

4.3 Chronic Diseases and Disease History

A secondary diagnosis can either be a complication that developed after admission or a comorbid condition that existed before the patient was admitted. A few secondary diagnoses are considered "pure" comorbidities. These preexisting conditions can be identified by their complication probabilities of zero in the CACI table. These diagnoses form the basis of the risk model's chronic diseases list, which takes the form of an expanded disease-specific Chronic Condition Table. The expansion takes into account the volume of common chronic disease codes in each of the disease strata (ccms_crl_group_by). To be included, a chronic disease code must occur among at least 1% of patients in a given disease stratum. Less common codes still undergo CACI processing but are counted into one of the six comorbidity categories.

Because some chronic conditions are similar, they tend to have similar influence on patient risk assessment, e.g. 491.2x (obstructive chronic bronchitis), 493.2x (chronic obstructive asthma) and 496 (chronic airway obstruction, NEC). These kind of chronic conditions are mapped to a common chronic condition code, and share the same coefficients.

Diagnoses in the Chronic Condition Table are incorporated into the risk model as independent variables. The following example shows how the algorithm works:

> If a secondary diagnosis is on the list of chronic conditions for a given disease stratum and thus included in the Chronic Condition Table, the secondary diagnosis will not undergo CACI processing. The diagnosis will not be counted in any of the six comorbidity categories. Instead, it is treated as a separate independent variable with its own corresponding risk coefficient, which can be found in the *beta* tables assuming that it's statistically significant. If the corresponding coefficient can not found in the *beta* tables, the coefficient should be assigned a default value of zero. The chronic condition has no impact on risk score, although the chronic condition is considered clinically relevant.

For the purposes of reporting, all comorbidities and chronic conditions should be included in the comorbidity count of CareScience Quality Manager front-end reports. The total number of comorbidities is calculated as the sum of CACI_Severity_Score_Cate_A to E and U plus the number of chronic conditions. For comorbidity counts by severity category, chronic conditions are mapped to the appropriate morbidity level in the Diagnosis_Morbidity table and are then added to the corresponding CACI_Severity_Score categories.

Since the CACI and Chronic Condition Tables are reviewed separately by different clinical panels, discrepancies between the two can arise. Whenever one is updated, the other must be reviewed for consistency. If discrepancies exist, they are resolved before the update becomes effective.

4.4 Valid Procedure

The 'validity' of a procedure is defined by its clinical proxy of patient characteristics and its statistical significance upon risk score. The validity is all about patient risk assessment. An 'invalid' procedure has no influence on risk score because it lacks either clinical proxy or statistical significance. The method does not intend to judge whether a procedure is appropriate treatment to a patient. An 'invalid' procedure does NOT mean it is an un-appropriate treatment.

To qualify for valid procedure candidacy for a given disease stratum, a procedure must satisfy a frequency criterion relative to that stratum. The minimum frequency is defined as:

$0.05*N / EXP (LOG10(N) – 1),$
where N is the number of cases in the disease stratum.

Procedures meeting this frequency requirement are then reviewed by the CareScience Clinical Expert Panel and classified into four categories. Category 1 consists of procedures that are unequivocally included in the model estimation. Most procedures in this category are considered defining procedures (e.g. ICD9 Px 361.x - Coronary Arterial Bypass Graft[7]) that are often administered upon admission. Category 2 includes procedures that are considered for inclusion dependent on timing and combination with other procedures. ICD9 procedure 4443 (Gastric bleeding endoscopic control) is one such example. Use of this procedure within the first 48 hours strongly suggests it is related to patient condition upon admission. On the other hand, use of this procedure in the later stages of a hospital stay may indicate it is caused by an earlier treatment. In the latter case, the procedure should not be included as a patient characteristic. Category 3 includes procedures whose model inclusion or exclusion remains unresolved due to clinical review disputation. Most procedures in this category are diagnostic and therapeutic procedures. Category 4 consists of procedures excluded as risk factors because of coding variation or features masking attributes of the care process.

Procedures are strong risk predictors, especially for length of stay, costs, and charges. Allowing more procedures to enter the model certainly increases model fit, however, as previously discussed, a higher R-Square does not necessarily result in better model performance for out-of-sample predictions. The number of procedures patients may receive can vary from zero to several dozen despite having the same principal diagnosis. Whether a patient receives a procedure is determined by the patient's clinical status, as well as the care provider's judgment. Adding procedures without considering their clinical implications may introduce an omitted variable bias by inadvertently including provider attributes that should be excluded from risk assessment. Moreover, coding practices vary across hospitals. Some hospitals do not regularly record diagnostic and therapeutic procedures while others do. Additionally, some hospitals record only the main procedures and leave out the linking minor codes. All of these considerations contribute to the polemic surrounding the use of procedures as risk factors.

The current Valid Procedure selection process is conservatively designed and aimed at avoiding data noise to the maximum extent. Consequently, the current Valid Procedure list is shorter than earlier versions with an emphasis on 'defining procedures' (Category 1) that are clinically relevant and seldom omitted from hospital data. These procedures are minimally controversial. Category 2 procedures are only included in the model if their timing criteria are met, reducing the ambiguity between patient attributes and care provider effects. The coding accuracy of these procedures, however, is less reliable than that of Category 1 procedures.

4.5 Other Clinical Elements and Considerations

4.5.1 Relative Value Unit

[7] *All CABG procedures are rolled up into a three-digit code 361 in the model due to their similarities.*

'Relative Value Unit' or RVU describes a procedure's intensity of resource usage. Thus, significant procedures such as coronary arterial bypass grafts (CABGs), which demand greater resources than minor procedures, possess greater RVUs. Although RVUs may seem to offer useful risk assessment information, they are not patient characteristics but instead reflections of the care provider's resource usage. For example, consider two AMI patients, one receiving a CABG (RVU 56.5) and the other receiving Mechanical Ventilation (RVU 1.89). Does the high RVU of the CABG procedure indicate that the condition of the first patient is worse than that of the second who is treated with a low RVU procedure? Although it could be argued that greater resources are only used when needed (i.e., the condition of the first patient must be more severe), the answer is probably "no." The high RVU of CABG only indicates that the hospital allocated more resources to the first patient. In terms of severity, the condition of the second patient is conceivably graver as suggested by the rescue procedure. For this reason, RVUs are not directly incorporated in the risk model. They are poor predictors of clinical outcomes, and their predicting power for financial outcomes is already largely captured by valid procedures.

4.5.2 "Do Not Resusitate" Orders

'Do Not Resusitate' (DNR) orders instruct hospital staff not to attempt life saving procedures or treatments should a patient's condition turn critical. In most circumstances, once a DNR order is issued, the patient expires within a few days. In certain cases, however, the patient actually displays signs of returning to normal status after supports are withdrawn in which case the DNR order is dismissed and supports restored. A crucial study at Cooper Health System revealed that patients receiving DNR orders represent a large percentage of all expired patients. These findings suggest that consideration of DNR orders may significantly alter the overall picture of hospital expiration. More sophisticated analyses with data gathered from a range of hospitals may help improve the accuracy of the mortality model. This possibly raises the following questions: 1) Does a common guideline for DNR orders exist across hospitals? 2) What are the implications of DNR orders issued at different stages of the hospital stay in terms of patient characteristics? 3) What is the interpretation of the DNR on/off switch from a clinical perspective? 4) What kind of role does a DNR order play in the conventional measure of mortality rates? Further study in this field is needed before DNR becomes a part of CareScience risk assessment model.

V. Statistical Significance

To assist users in interpreting outcome comparison reports for a targeted "analysis set" of cases, CareScience tool provides an estimate of the **statistical significance** of each outcome deviation (actual – expected). A "significance flag" indicates the probability that the results could have occurred randomly if there were not a true underlying effect. In the front-end reports, a double asterisk (**) indicates 90% significance while a single asterisk (*) indicates 75% significance. Large deviations tend to be significant, except when great uncertainty surrounds expectations for cases in the analysis set. The choice of relatively "low" significance levels (75% and 90% as

opposed to the frequently used 95% and 99% levels) reflects the purpose of the reporting tool – it is highly sensitive but not very specific in tolerating false positives to avoid false negatives.

Statistical significance depends on the prediction error of the CareScience risk model, which derives from the properties of ordinary least squares regression analysis. Prediction error reflects how well the model fits the population on which it is calibrated. Hence, it is a **characteristic of the model calibration** and not purely a feature of the group of patients in the analysis set. As a practical matter, prediction error (hence, statistical significance) can be computed for any number of cases in the analysis set, even just one case.

The basis for computing statistical significance is to **aggregate** (as described below) the case-by-case prediction error. Just as the predicted outcome risk for each case is based on that patient's characteristics as processed by the CMS model, the model generates a prediction error for each case. The prediction error for the group of patients in the analysis set is derived by aggregating the cases in the analysis set.

This aggregation poses a challenge, because it involves combining the uncertainty around the predicted value (risk) and the imprecision of the observed outcome value, especially when the number of cases in the analysis set is small. For this reason, one must be cautious in interpreting "significant" deviations when the number of cases in the analysis set is small. In such cases, the conclusion that the deviation is "significant" is based on an **assumption that the observed rate is measured with little or no error**. Under this assumption all uncertainty in the deviation greater than a "real" **opportunity** is attributed to the prediction error. The assumption may no longer be true when the number of cases is too small. A conventional suggestion of minimum case count is around 15 – 20. If the deviation between the raw and risk (expected) values is large relative to the prediction error, then the deviation computed from the analysis group is not likely to be due to pure chance.

Remember that the prediction error is based on a model calibration on thousands, if not tens of thousands of patient observations. On the other hand, the number of cases used in certain analysis sets (such as the number of patients treated by a chosen physician for a particular condition) can be quite small. For this type of analysis, the ***raw*** mean outcome is based on a relatively small sample of cases, which makes the assumption of low measurement error less plausible. In such circumstances, the user must interpret the appearance of significance flags with caution, since the cases in the analysis may be idiosyncratic.

5.1 Claim-level Computation

Recall that standardized measures (risks) are calculated using the ordinary least squares (OLS) regression model:

$$\hat{y}_{ijkl} = x_{ijkl}\hat{\beta}_{kl}, \quad \forall ijkl$$

The following subscripts provide clarity to the notation:

i = claim_id (each row in the patient table)
j = provider or grouping
k = ICD-9-CM principal diagnosis (3 digit)
l = outcome (mortality, length of stay, charges, cost, complications, complication morbidity)

\

Hence, $\hat{y}_{ijkl}$ [8] is the predicted value for each outcome, *l*, at the patient level, *i*, for each provider, *j*, and diagnosis, *k*. x_{ijkl} is a vector of patient characteristics and other severity measures that are outside the provider's control, including clinical factors, patient demographic characteristics, and patient selection factors. $\hat{\beta}_{kl}$ is the marginal effect of the independent variables on the outcome measure. x_{ijkl} and $\hat{\beta}_{kl}$ are of dimension *q* equal to the number of linearly independent regressors (including the number of distinct responses for each categorical variable).

The goal of this effort is to calculate a confidence interval around each $\hat{y}_{ijkl}$ to determine whether the observed raw value is "close" to the predicted value and assess the effect of the risk factors on the patient outcome. The confidence interval, 100(1-α)% , around each individual patient level predicted value $\hat{y}_{ijkl}$ is calculated as

$$\hat{y}_{ijkl} \pm sep(\hat{y}_{ijkl}) \cdot t_{\alpha/2}$$ [9]

where the standard error of the predictor is

$$sep(\hat{y}_{ijkl}) = s\sqrt{x_{ijkl}(X_{kl}'X_{kl})^{-1}x_{ijkl}' + 1}$$

and

$$s = \sqrt{\frac{SSR}{n_{kl} - q_{kl}}}.$$

SSR is the sum of squared residuals from the fitted regression line, n_{kl} is the number of observations for principal diagnosis *k* for outcome *l*, and q_{kl} is the number of estimated regression coefficients for diagnosis *k* and outcome *l*.

---

[8] For the sake of consistency, this description carries the full range of subscripts, since outcome measures are calculated by principal diagnosis for each patient and each outcome and then aggregated to the desired group level.

[9] Note that we aggregate individual standard errors that include the random error in the population. Without the 1, $sef(\hat{y}_{ijkl}) = s\sqrt{x_{ijkl}(X_{kl}'X_{kl})^{-1}x_{ijkl}'}$, represents only the sampling error around the fit of the regression line, which is generally much smaller.

5.2 Aggregation

The challenge arises when aggregating the standard errors to characterize any targeted grouping of patients, such as all patients admitted by physician ***j***.

The expected value for a given outcome aggregated to the provider level, $\hat{y}_{jl}$, is the average of the patient level expected values for that provider and outcome. Mathematically it is expressed as, $\hat{y}_{jl} = \frac{1}{n_{jl}} \sum_{ik} \hat{y}_{ijkl}$, where n_{jl} = the number of patients treated by provider *j* across all diagnoses, and *l* is the relevant outcome.[10] This is compared with the average of the raw values for the same grouping, $\bar{y}_{jl} = \frac{1}{n_{jl}} \sum_{ik} y_{ijkl}$, where y_{ijkl} are the actual outcome values.
The estimated variance of the provider level estimator is

$$V(\hat{y}_{jl}) = \frac{1}{n_{jl}^2} \sum_{ik} [sep(\hat{y}_{ijkl})]^2$$ assuming iid→cov=0

and the variance of the raw outcome measure is

$$V(\bar{y}_{jl}) = \frac{1}{n_{jl}} se(y_{jl})^2 = \frac{1}{n_{jl}} \left(\frac{\sum_{ik} (y_{ijkl} - \bar{y}_{jl})^2}{n_{jl} - 1} \right).$$

In CareScience suite of products, we report a deviation score that represents the difference between the average observed and expected values for each outcome, $d_{jl} = \bar{y}_{jl} - \hat{y}_{jl}$, where d_{jl} is the provider level deviation score. To determine the confidence interval around this deviation score, we estimate the variance, $V(d_{jl})$, around it. The confidence interval allows us to gauge whether the study group's deviation could possibly be zero, indicating no significant difference between the observed and expected outcomes. Given that $d_{jl} = \bar{y}_{jl} - \hat{y}_{jl}$, the variance of the deviation score is $V(d_{jl})$. $V(d_{jl}) = V(\bar{y}_{jl} - \hat{y}_{jl})$ which can be rewritten as $V(d_{jl}) = V(\bar{y}_{jl}) + V(\hat{y}_{jl}) - 2Cov(\bar{y}_{jl}, \hat{y}_{jl})$.

[10] In general, $n_{jl} = n_j$ except where certain cases are excluded for a given outcome. For example, if a physician treats a certain number of cases in a period in which some of those cases were not in a diagnosis that observed at least one death, the total number of cases, n_j will be different for that provider's LOS risk than for his mortality risk. That said, using the new method developed to fill null values with the mean risk ensures that n_{jl} will always equal n_j whenever observations are complete.

However, we don't know the covariance of $\bar{y}_{jl}$ and $\hat{y}_{jl}$. We cannot calculate this in Data Manager, since $\bar{y}_{jl}$ and $\hat{y}_{jl}$ are aggregated to the provider level and are therefore provider dependent. We also cannot assume that the mean raw outcome rate is independent of the mean risk-adjusted outcome rate, which would yield a covariance of 0. Therefore, for the purposes of this analysis, we will treat $\bar{y}_{jl}$, the mean observed outcome rate for the provider, as non-stochastic or a non-random variable. Given this simplification, $\bar{y}_{jl}$ is simply a point rather than an estimator with a distribution and therefore has no variance and no covariance with $\hat{y}_{jl}$, the mean risk-adjusted outcome rate for the provider.

Given this simplification, the variance of the deviation score reduces to $V(d_{jl}) = V(\hat{y}_{jl})$, and we can test the null hypothesis that the deviation is equal to zero:

H_o: $d_{jl} = 0$
versus the alternative hypothesis H_a: $d_{jl} \neq 0$.

Rejection of the null hypothesis indicates a significant difference between the observed and expected outcome measures. Since we generally work with small sample sizes, we perform a t-test on the null hypothesis. More specifically, we calculate a ***t*** statistic, $t_{jl}^{\alpha} = \left(\frac{d_{jl} - 0}{\sqrt{V(\hat{y}_{jl})}} \right)$, with $n - \sum_k q_k$ degrees of freedom, where *n* refers to the total number of observations in the dataset and $\sum_k q_k$ is the total number of estimated coefficients across all diagnosis groups.

This computed statistic is compared to the critical value for the ***t*** distribution with $n - \sum_k q_k$ degrees of freedom and the desired alpha level.

As a practical matter, the degrees of freedom calculated in this manner will almost always be in the hundreds or thousands, which brings the test statistic to its limiting distribution, the normal. On the other hand, the number of cases used in a narrow aggregation, such as the number of patients treated by a chosen physician for a particular condition, can be quite small. For this type of analysis, the ***raw*** mean outcome is based on a relatively small sample of these N cases and has a t-distribution with N-1 degrees of freedom. This becomes the relevant number of degrees of freedom in conducting a test for statistical significance.

All deviation scores will have an indicator of whether the deviation is statistically significantly different from zero. For example, a **(90% significance level) indicates that there is less than a

10% probability that the deviation (standardized - raw) is due entirely to chance. Hence, we can reject the null hypothesis of zero deviation with a 10% chance of a (type I) error.

5.3 Environmental Description

1. The application reports two distinct levels of significance, 75% and 90%.
2. Significance levels are not user-controlled and are therefore standard for all clients.
3. Significance flags are indicated by the graphic "*."
4. A single asterisk (*) indicates that a deviation is significantly different from zero at the 75% confidence level while a double asterisk (**) signals that a deviation is significantly different from zero at the 90% confidence level. Descriptions of these significance notations appear on every deviation report. Deviations that round to 0.0 do not receive a flag.
5. The Data Manager program has been modified to calculate the standard error for each outcome at the patient level predicted value, $sep(\hat{y}_{ijkl}) = s\sqrt{x_{ijkl}(X_{kl}'X_{kl})^{-1}x'_{ijkl} + 1}$.
6. The program's front-end scripts then perform the mathematics to aggregate the standard errors for the specified grouping (e.g. provider, MDC, DRG, or CTC[11]). The calculations used are: (a) calculate an average variance $V(\hat{y}_{jl})$, and b) calculate each deviation score, d_{jl}.
7. In calculating deviation scores on reports, the program's front-end scripts use the non-rounded raw and risk values to calculate the deviation (raw-risk). The front-end report then rounds the raw, risk, and deviation scores to the first decimal place. The reports have a footnote stating "Raw minus risk may differ from deviation due to rounding."
8. The deviation (computed from the non-rounded raw and risk values), raw standard error, and number of observations (n) generate a *t* statistic that is compared to a critical value to determine significance.
9. The *t* statistic is calculated for each deviation score, at any aggregate level, using the following equation: $t = \frac{n \cdot deviation}{\sqrt{\sum (sep)^2}}$.
10. In determining n for the *t* statistic calculation, only valid raw values are used. (i.e. for ln values ≠ 99[12]).
11. The ccms_common.t_distributon table is populated with critical values for a two-tail ***t*** test.
12. To determine significance, locate the critical *t* value (field name: t_value) from the ccms_common.t_distribution table having the appropriate degrees of freedom (df = n-1) and significance level (sig_level = 0.75 or 0.90). If there is not an exact match for the number of degrees of freedom, choose the closest number that is smaller than the observed number.
13. If the calculated *t* statistic exceeds the t_value for a sig_level of 0.90 ($\alpha = 0.10$), the deviation score receives two asterisks. If the calculated *t* statistic exceeds the t_value for a sig_level of

[11] MDC = Major Diagnosis Category
DRG = Diagnosis Related Group
CTC = Common Treatment Category
[12] Refer to N:\Analytix\develop\methodologies\Lntrans\logreq2.doc for information on column requirements.

0.75 ($\alpha = 0.25$) but is less than that for a sig_level of 0.90, the deviation receives one asterisk.

14. The program's front end imposes a constraint preventing any score with a rounded deviation of 0.0 from receiving a significance flag. (i.e. even if before rounding the deviation $\neq 0$ and the score has received a significance flag, the flag will be removed.)

VI. Select Practice

Select Practice is a collective name for a series of CareScience methodology, product and reports. It was first developed in 2002 and 2003 as an additional feature of Quality Manager. Under Select Practice toggle, a hospital can benchmark its performance, relative to a group of selected hospitals efficiently delivering high quality of care. The methodology of identifying the selected hospitals soon gets widely accepted, and evolves into multiple applications. All of them are under the collective name of Select Practice. This documentation will mainly focus on the methodology. The mathematical details are illustrated in the Appendix C *(Select Practice Formulas).*

6.1 Setting

Outcome comparisons have long been viewed as a powerful way to motivate improvement in inpatient quality of care. These comparisons, often called practice profiles, outcomes reports, report cards, or scorecards, have captured the attention of not only health care providers but also payers and consumers. Although no single universally accepted quality of care measure exists, certain key elements are common.

Mortality is the most widely accepted measure of quality of care, but mortality alone can not fully cover all dimensions of quality. The CareScience model of quality is measured by the incidence of three adverse outcomes: mortality, morbidity, and complications, which are combined into a single quality measure using the preference weightings from the Corporate Hospital Rating Project.[13]

CareScience defines a highly rated hospital as one that delivers excellent health care in an efficient way. In the CareScience rating model, efficiency is captured by length-of-stay (LOS). Length-of-stay serves as a proxy for resource usage, reflecting how efficiently hospitals allocate resources.

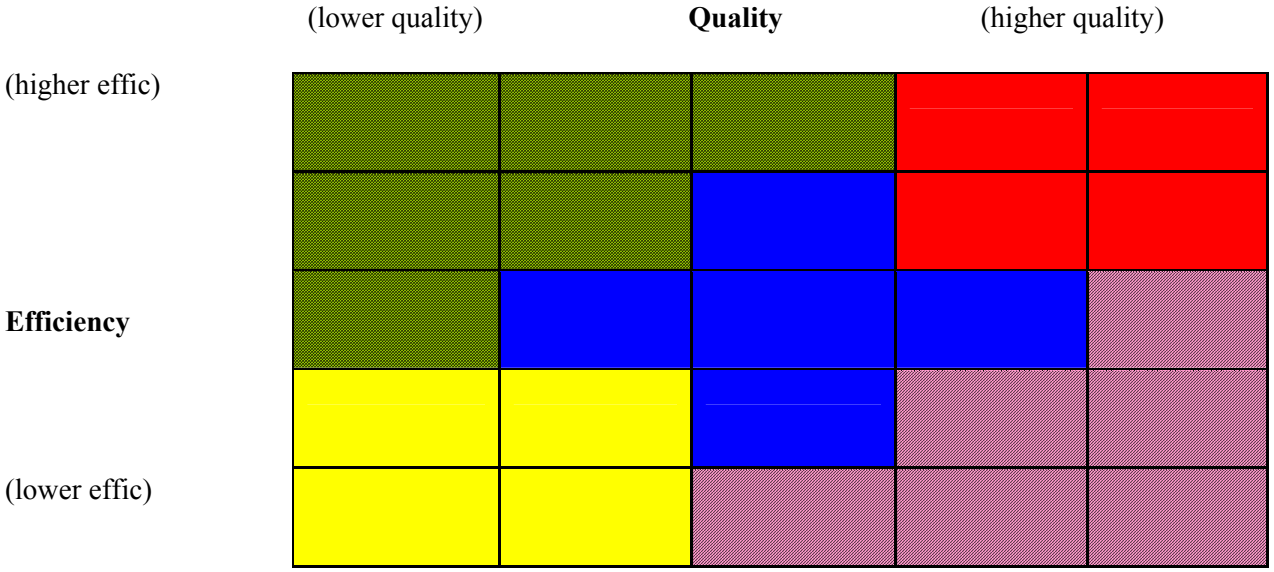
For each disease grouping, hospitals are ranked for quality and efficiency separately, with the highest rankings going to hospitals with the lowest risk-adjusted LOS and adverse outcome rates. To qualify as "Select Practice" for a given disease, a facility must be in the top two quintiles (top 40%) for both efficiency and quality measures. *Because the rating system is two-dimensional*

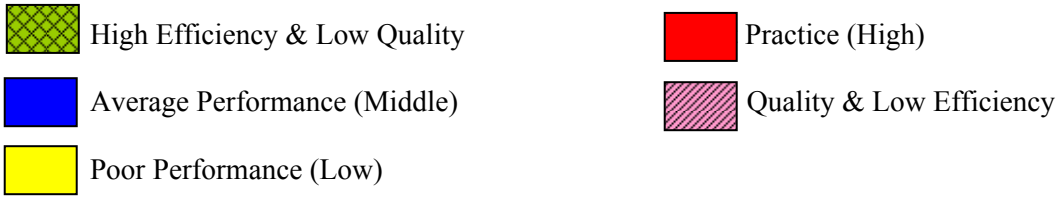
[13] Pauly MV, Brailer DJ, and Kroch EA, "The Corporate Hospital Rating Project: Measuring Hospital Outcomes from a Buyers Perspective," *American Journal of Medical Quality*, **11**(3):112-122.

and takes into account both quality and efficiency, the system makes no trade-off between these two considerations. The five by five efficiency-quality matrix is illustrated in Figure 1. For the majority of diseases, quality and efficiency rankings are weakly correlated, and the Select Practice facilities ("High") constitute 16% (40% of 40%) of all facilities that qualify for ranking. Other ranking combinations include the following: 1) placing in the bottom two quintiles of both efficiency and quality (four poor performance "Low" cells), 2) placing in the middle three groupings (five average performance "Middle" cells), 3) placing in the six low quality-high efficiency cells ("Cheap"), and 4) placing in the six high quality-low efficiency cells ("Dear").

This hospital rating system is *disease specific* for about 60 conditions (depending on the data type) that cover virtually all cases; hence, it is not explicitly a hospital level ranking. The working hypothesis is that a high performance hospital in a given disease has better chance to be high performance in other related diseases in the same service line. Nevertheless, extension of this disease-specific profiling system to rate hospitals as a whole can only be accomplished by assessing the distribution of the disease counts in Figure 1 for each hospital. A high performance hospital would therefore be one with a large number of "Select Practice" diseases (upper two quintiles for both quality and efficiency). Conversely, a poor performance hospital would have a large number of "Poor Performance" diseases. The other three categories are harder to apply to the hospital as a whole, since it is possible to be "average" in a number of ways, not just by having a preponderance of disease areas in the center of the grid, but by having a great dispersion in performance across diseases. To make a practical judgment, it may be necessary to invoke an explicit trade off between efficiency and quality and between consistency and dispersion.

Figure 1. Identification of Five Performance Categories Based on CareScience Select Practice™ – Preliminary Recommendation for CMS Study





6.2 Methodological Details

6.2.1 Data Source

The Select Practice methodology was applied to two hospital data bases: (1) state hospital association all-payor patient records and (2) MedPAR patient records from the Center for Medicare and Medicaid Services. This first database does not cover all 50 states for two reasons. First, some states do not provide the data, and second, some states charge prohibitively high prices for their data. Usually, the State data obtained contains 15 to 20 million inpatient records from over 2,000 to 2,600 facilities in 14 to 20 states. Fortunately, many states with large population are represented, including AZ, CA, FL, MA, MD, NJ, NY, PA and TX. MedPAR data contains over 12.5 million Medicare inpatient records from almost 6,200 facilities nationwide. The number of patient records in the MedPAR data has increased yearly as the aging population continues to grow. On the other hand, the number of facilities drops as hospital consolidation continues.

6.2.2 Risk Adjustment

The databases were first processed under the CareScience risk assessment methods described in the previous sections. Risk scores were generated for each of the four outcomes: mortality, complications, major morbidity, and length of stay. A risk score represents the expected or 'standard' outcome under typical care based on a patient's health status and other characteristics. Risk scores serve as benchmarks, whereby the quality and efficiency of hospital services can be evaluated across facilities, regardless of case mix. If the raw scores deviate negatively from their risk scores, the facility is considered a better provider than the benchmark.

6.2.3 Quality Index Computation

Based on an earlier developed method from the Corporate Hospital Rating Project (CHRP), risk-adjusted adverse outcome rates for mortality, morbidity, and complications are combined into a single quality measure represented by the function:

$$Q_{kh} = 0.46(T_{kh})^{0.96} + 0.29(B_{kh})^{0.91} + 0.25(C_{kh})^{0.94}.$$

Q, T, B, C, h, and k represent the quality index, risk-adjusted mortality, risk-adjusted major morbidity, risk-adjusted complications, facility, and disease, respectively. Hospitals then are ranked according to their quality index with smaller values of Q indicating better quality.

Quality index can be normalized with the following formula:

$$Normalized\ Index_{kh} = (Q_{kp} / Q_{kh})*100,$$

where Q_{kp} represents the quality index of population (P) in disease K. After normalization, higher score reflects higher quality.

The volume of discharges per hospital varies greatly across the database. Fewer discharges may not provide a statistically sound analysis and thus necessitates a minimum volume cutoff. The applied criterion is that a facility must have at least 100 discharges in a given disease (defined by principal diagnosis) to qualify for ranking in State data. In MedPAR data, the threshold is cut by half to reflect its smaller size, which only covers the 65 years old and above population. The qualified facilities are divided into five categories based on their ranks. Because all cases from a given facility are, by requirement, sorted into the same quintile, the categorization can not be precisely processed. Each category represents approximately one fifth of the total volume and all facilities.

6.2.3 Efficiency Index Computation

Length-of-stay is used as a proxy for resource usage, based on the assumption that a hospital spends more resources on patients who stay longer in the hospital for a given disease. Since length-of-stay is usually recorded very accurately for each patient, it is an ideal measurement for a patient-level model. For each disease, facilities with cases exceeding the cutoff (100 for State data and 50 for MedPAR data) are ranked according to the function,

$$RL_{kh} = exp\{(log(LOS)_{kh} - (logLOSrisk)_{kh}),$$

where RL, k, and h represent ratio, disease, and facility, respectively. Lower ratios denote greater efficiency. The efficiency index can also be normalized according to the same formula that is used for the quality index. Facilities are then divided into five categories based on the same criteria used for the quality index.

6.2.4 Cross Tabulation of Quality and Efficiency Index

Neither the quality index nor length-of-stay can alone determine the Select Practice hospitals. Our study shows that hospital rankings for quality and length-of-stay are largely independent for the majority of diseases. In other words, a hospital has roughly the same probability of falling into one of the five quality index categories regardless of its risk-adjusted length-of-stay, and vice versa. In a 5X5 cross tabulation of the quality and efficiency indices, hospitals are relatively evenly distributed in each of the 25 cells.

A Select Practice hospital is expected to deliver high quality healthcare in an efficient manner. For each disease, Select Practice hospitals are identified by choosing facilities that fall into the top 2X2 cells in the 5X5 cross-tabulation matrix. Select Practice hospitals represent roughly

16% (4/25) of cases and facilities for each disease. To prevent small samples from diluting statistical power, 200 is set as the minimum number of facilities needed for ranking. 16% out of 200 is 32. We believe that this cutoff is the minimum number of Select Practice hospitals required to keep Select Practice statistically meaningful.

No matrix is constructed if there are fewer than 200 qualifying facilities for a given disease. Diseases with low volumes are rolled up into one of the 18 major diagnosis groups (Broad Diagnosis Group), defined by the ICD9-CM classification system. For example, ICD9 codes 001 to 139 are rolled up into BDG 1 (infectious and parasitic diseases) with the exception of 038, which has sufficient volume to stand alone. All BDGs are then processed in the same manner, and a list of Select Practice hospitals is generated for each of the Broad Diagnosis Groups.

6.3 Scaling Factors

6.3.1 Scaling Factor Calculation

After the Select Practice hospitals are identified, their performance is measured as the extent to which their performance differs from the overall level. For mortality, morbidity and complications, the overall performance level for both the Select Practice hospitals and the entire population of hospitals is captured by their case-weighted arithmetic means. By comparing the risk-adjusted outcome of the Select Practice hospitals (*Select Practice case-weighted mean deviation + Population case-weighted mean raw rate*) to the population's overall case-weighted mean raw rate, a ratio is obtained for each of the three clinical outcomes for each disease. For LOS, the ratio is simply the case-weighted RL_{kh} of the Select Practice hospitals. These ratios are called 'Scaling Factors.' They are applied to 'scale' down standard risk to Select Practice risk. Scaling factors are numbers between zero and one, and they usually fall into the range between 0.75 and 0.95. The smaller the ratio is, the greater the difference is between the performance of the Select Practice hospitals and the overall hospitals.

Scaling factors are also calculated for Cost although they are not used in Select Practice ranking. The ratio of actual cost and cost risk is first calculated for all hospitals by disease, using the following function:

$$R\$_{kh} = exp\{(log(Cost)_{kh} - (logCostrisk)_{kh}),$$

where R$, k and h represent ratio, disease and facility, respectively.

The cost scaling factor is then calculated as the case-weighted $R\$_{kh}$ of the Select Practice hospitals.

6.3.2 Scaling Factor Range

Mathematically, the scaling factors for mortality, morbidity and complication may be greater than one, because the quality index that combines them allows trade-offs among them.

Extremely excellent performance in one or two outcomes may compensate for bad performance in other outcomes. Therefore, a Select Practice hospital may theoretically have worse performance than the overall population's performance in one or two outcomes. In reality, Select Practice hospitals often have balanced performance in all three clinical outcomes. Even if a few Select Practice hospitals perform badly in one or two outcomes, the group volume (at least 32 Select Practice hospitals) can reduce the outlier effect, thus, largely guaranteeing that the scaling factors of the clinical outcomes are within the reasonable range of 0.75 to 0.95.

In a few disease groups, the mortality rate is very low (e.g., ICD9-Diag 303 - alcohol dependence). For these disease groups, complications become the dominant factor in the quality index. Because mortality is hardly relevant for these rankings, the scaling factor may actually exceed 1.0, and consequently, it is capped at 1.0. At this level, Select Practice hospitals are on par with all other hospitals.

For efficiency, LOS alone determines hospital ranking. No hospital with worse performance than the population's overall performance level can rank in the top two quintiles. Therefore the LOS scaling factors are always below 'one.' Since cost is highly correlated with LOS, the cost scaling factors often trail LOS. This, however, does not mathematically guarantee that the cost scaling factors are within the reasonable range, and consequently they are capped at 'one.'

6.3.3 Scaling Factor Implementation

The detailed calculation of scaling factors and how to apply them are described in Appendix C *(Select Practice Formulas)*. The following is a simplified example of how to apply the scaling factors and interpret Select Practice risk. Mortality rate for AMI patients in a given hospital is 7.5% while the standard mortality risk is 8.0%. The deviation is -0.5%. In other words, the hospital has saved slightly more patient lives than a typical hospital with the average performance level and given the case mix. For AMI, the Mortality scaling factor is 0.89. The product of the standard mortality risk and the scaling factor is 7.1%, which is called the Select Practice risk. The Select Practice risk is the predicted outcome for a Select Practice hospital given the case mix. Compared to the Select Practice risk, the hospital's deviation now becomes 0.4%, which indicates that the hospital has saved fewer patient lives than a typical hospital in the Select Practice group with the given case mix.

Scaling factors are created and updated by Research, using SAS language. Because State data is different from MedPAR data, the SAS program differs accordingly. The Select Practice programs can not be automatically executed. The name and location of the database has to be changed every time a new database is processed. Some table and column names may also change as the Data Manager program evolves. The first half of the Select Practice program runs by 3-digit principal diagnosis. After the major 3-digit diagnoses are identified and processed, the second half of the program begins processing the broad diagnosis groups (BDGs). Because the volume of each diagnosis may vary yearly, the list of major diagnoses is subject to slight changes. It is therefore necessary to manually code the rolling up of minor diagnoses.

Scaling factors from the latest State data are applied to the production data. Before they are handed to DAU, they must be saved in a table in which 3-digit principal diagnosis is the primary key. Minor diagnoses that are rolled up in the same BDG share the same scaling factors. The actual application of scaling factors occurs in the front end report developed by Software Engineering and monitored by Product Management.

6.4 Other Implementation of Select Practice Method

Since its debut in 2002, the Select Practice method has been widely accepted for hospital ranking. It has been used as a powerful marketing tool by the Sales team. Based on MedPAR data, CareScience Select Practice Hospital List has been formally announced in 2005. The list is updated annually. Select Practice methodology has also been used for multiple consulting and research projects, bound to care providers, regulatory agencies and academia. Depending on the purpose of these activities, Select Practice method is continuously updated, and the SAS programs are continuously reshaped to accommodate new requirements.

Appendix A - Calculating Costs

In the hospital setting medical "cost" does not have a common, single definition, in part because of the diverse reasons for tracking costs. Record keeping at the patient level should ideally distinguish at least three different aspects of medical cost and its components. First, the most straightforward patient-oriented view of cost is the value of all resources directly included in the provision of individual patient care. Examples of these **variable** costs include nursing, technical, and physician labor, as well as any tests/procedures performed or supplies consumed. These costs are often the basis for clinical profiling, because they are presumed to be under the control of treating physicians. A second broader measure of **total** hospital cost includes, in addition to the aforementioned variable costs, all overhead and administrative costs (**fixed** costs). These costs are not limited to patient care directly, but rather encompass the costs of running the hospital and financing its physical plant and equipment. Most of these components are included in Medicare's Hospital Cost Report Information System (HCRIS) and are grouped by hospital departments. Third are cost reimbursements, the dollar amount paid to the provider for each case. Reimbursements are not necessarily the true treatment costs, but rather the amount that the payor has determined to be typical for the specific case mix and care requirements, in this case Medicare's RBRVS (Resource Based Relative Value Scale) computations. These, too, are available from the Center for Medicare and Medicaid Services (CMS) and are grouped by DRGs.

One other dimension to tracking costs and charges deserves note. Aggregate or departmental dollars can be assembled for the entire facility or restricted to the inpatient population only. Facility level cost-to-charge ratios are often much greater than those limited to inpatients. There are two reasons: (1) outpatient costs are much closer to billings; and (2) facility level aggregations more naturally allow the inclusion of overhead (fixed) costs.

A number of published studies have assessed the measurement of the costs of providing medical care in hospital settings. When comparing costs derived from departmental cost-to-charge ratios to costs derived from RVUs[14], CCR-calculated costs were shown to be a reliable methodology for estimating hospital costs and for comparing average costs for a collection of patients in a given DRG among hospitals.[15] Hospital-level CCRs, however, are not a reliable way to assess costs for individual patients when evaluated by the RVU "gold standard." Nonetheless, when large representative samples of patients are the basis for comparison, hospital-level CCRs can be useful for overall cross-hospital profiling.

The accuracy of cost-to-charge ratios is influenced by a number of hospital factors, including nursing and physician labor, teaching status, patient risk and demographics, insurance status, and even location. An alternative to using RVUs to assess the reliability of CCRs is to compare hospital charges to costs that are derived from internal cost accounting systems, for those

[14] RVU, relative value units, describes a procedure's intensity of resource usage. The details about RVU can be found in the section of Clinical Knowledge Base.
[15] Shwartz M, Young DW, Siegrist R. The Ratio of Costs to Charges: How good a basis for estimating costs? Inquiry 32:476-481. 1995.

hospitals that have them.

Some medical centers, for example, have adopted a cost accounting and management information system developed by Transition Systems, Inc. (TSI). This system relies on DRGs to allocate costs and involves the following process: First, all billable procedures in the hospital are grouped into intermediate products which are then mapped into departments. Within each department, each intermediate product is assigned a relative weight for up to nine cost categories – variable labor[16], variable supplies, variable other, fixed direct labor[17], fixed direct equipment, fixed direct facilities, fixed direct other, variable indirect, and fixed indirect. This relative weight is then multiplied by the annual volume of the product and then summed over all products per department to determine the total relative weight units expended by the department. These costs are broken down into the nine cost categories. One main problem with this system is that hospitals differ in how they define intermediate products and in how they map them into departments.[18]

The CareScience data model explored a number of alternatives and after extensive testing found that the best approach was to exploit the Medicare cost and charge information from the minimum data set hospital cost report (HCRIS) to create cost-to-charge ratios (CCR) by department. The cost and charge information is for the entire facility, organized by departments, and focuses on ancillary services without distinguishing variable and fixed costs. The application of these ratios to client data poses challenges, because the CareScience Data Manager tracks charges at the patient level rather than by department. Some studies have shown that the overall hospital CCR, in contrast to departmental CCRs, is not a reliable basis for comparing costs across hospitals. These computed CCRs contain some extreme values, which require trimming. After trimming (excluding the top and bottom 1 percentile) the average ratio is 0.62 with a standard deviation of 0.23. These total facility ratios are somewhat smaller than the average (unweighted) departmental ratios with which they correlate at about 86 percent.

Our ultimate goal is to accurately account for costs and resources attributable to an individual patient's care. Meeting it would require detailed information from our clients both in terms of charge breakdowns by department and complete cost accounts. This kind of information could help us build a model of CCR-based costs. CCR derived costs have been shown to be reliable when comparing the relative cost of patients in a DRG in one hospital to the average cost of patients in the same DRG in a group of hospitals.[19] This could be accomplished if we obtained charge break downs on a departmental level. These charges could then be multiplied by departmental CCRs obtained either from the hospitals own total annual departmental costs and charges or those calculated from the CMS minimum data sets.

---

[16] Variable labor is direct patient care or services.
[17] Fixed labor includes strictly administrative staff and caregivers when they perform administrative duties.
[18] Shwartz M, Young DW, Siegrist R. The Ratio of Costs to Charges: How good a basis for estimating costs? Inquiry 32:476-481. 1995.
[19] Shwartz M, Young DW, Siegrist R. The Ratio of Costs to Charges: How good a basis for estimating costs? Inquiry 32:476-481. 1995.

Appendix B – 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.

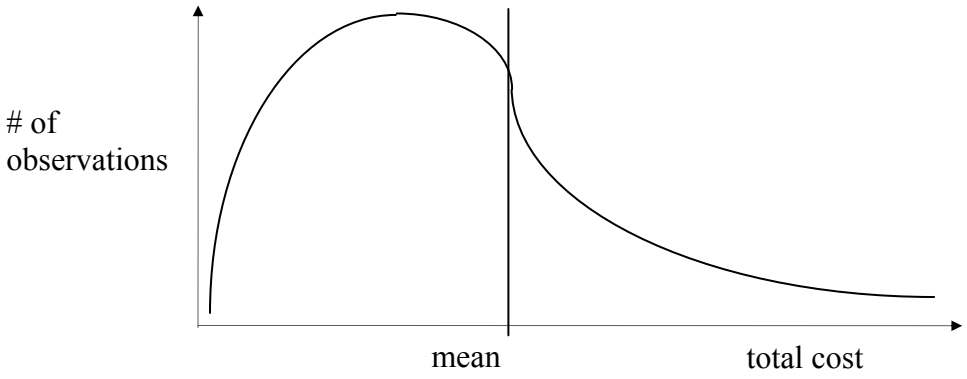
A more robust solution is to take the natural log of the dependent variable, which results in an approximately symmetric distribution and contracts the outliers inward toward the center of the data, as shown in Figure 1(b). It also ensures that all predicted values will be positive. (No matter how negative 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 (se below), are equal. As the skew increases in unimodal distributions the ratio of the arithmetic mean to the geometric mean grows from unity.

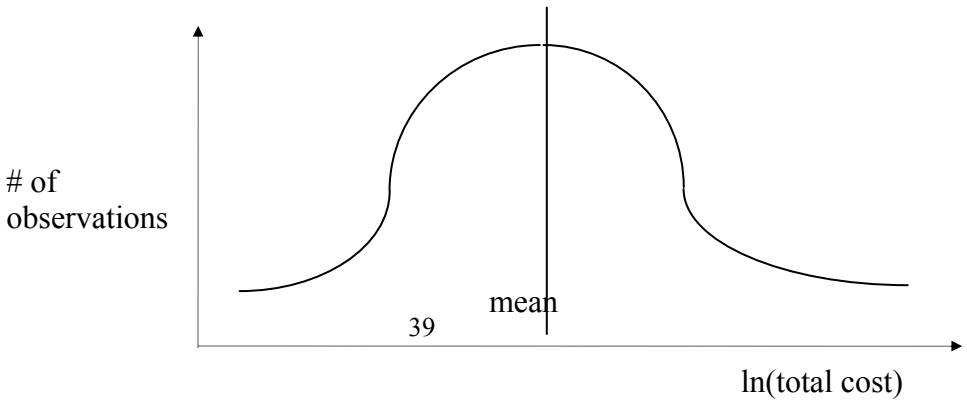
To illustrate skew: Total cost is skewed right but the natural log of total cost – ln(cost) – is approximately symmetrically distributed, therefore using linear regression to forecast ln(cost) will result in much better estimates with smaller error.

Figure 1

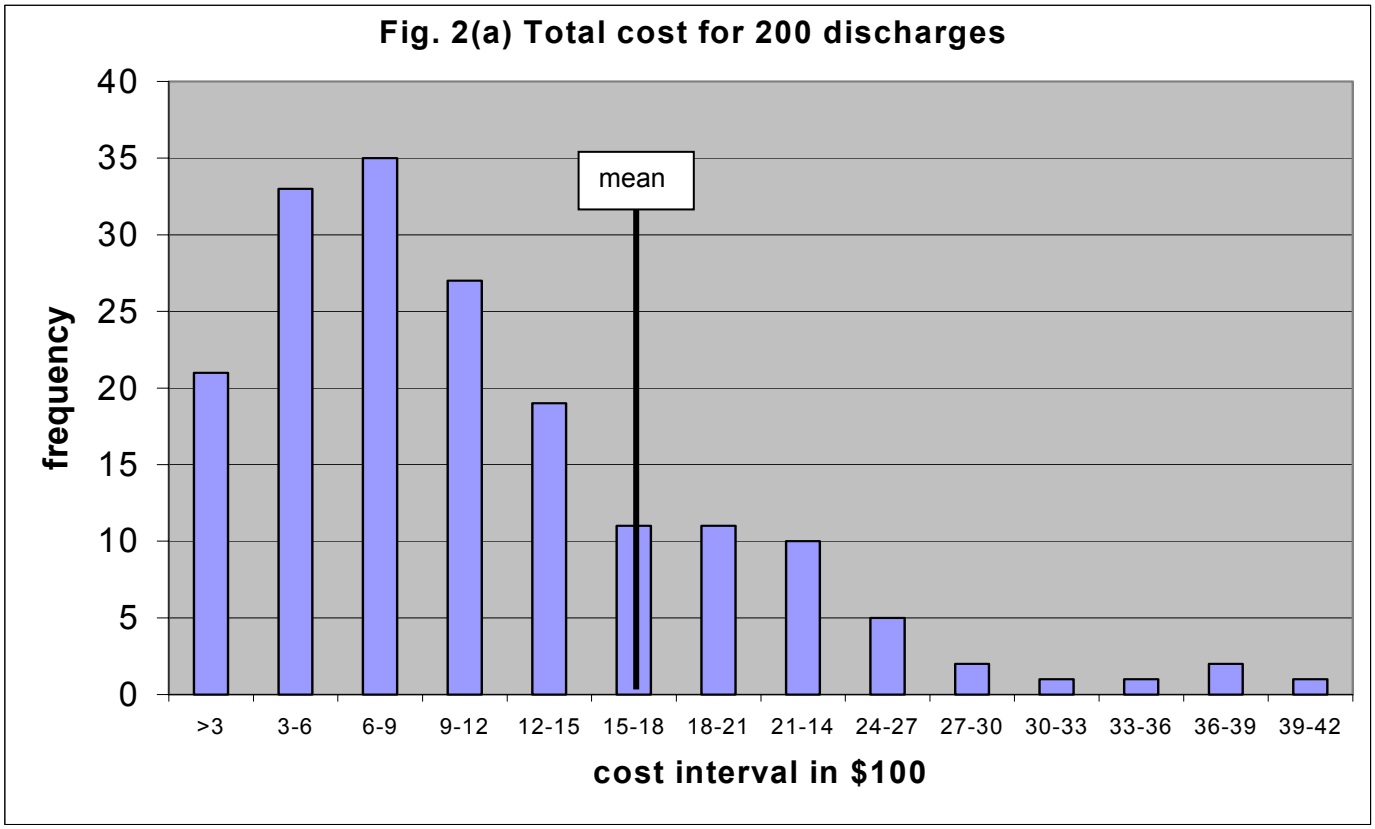
(a)

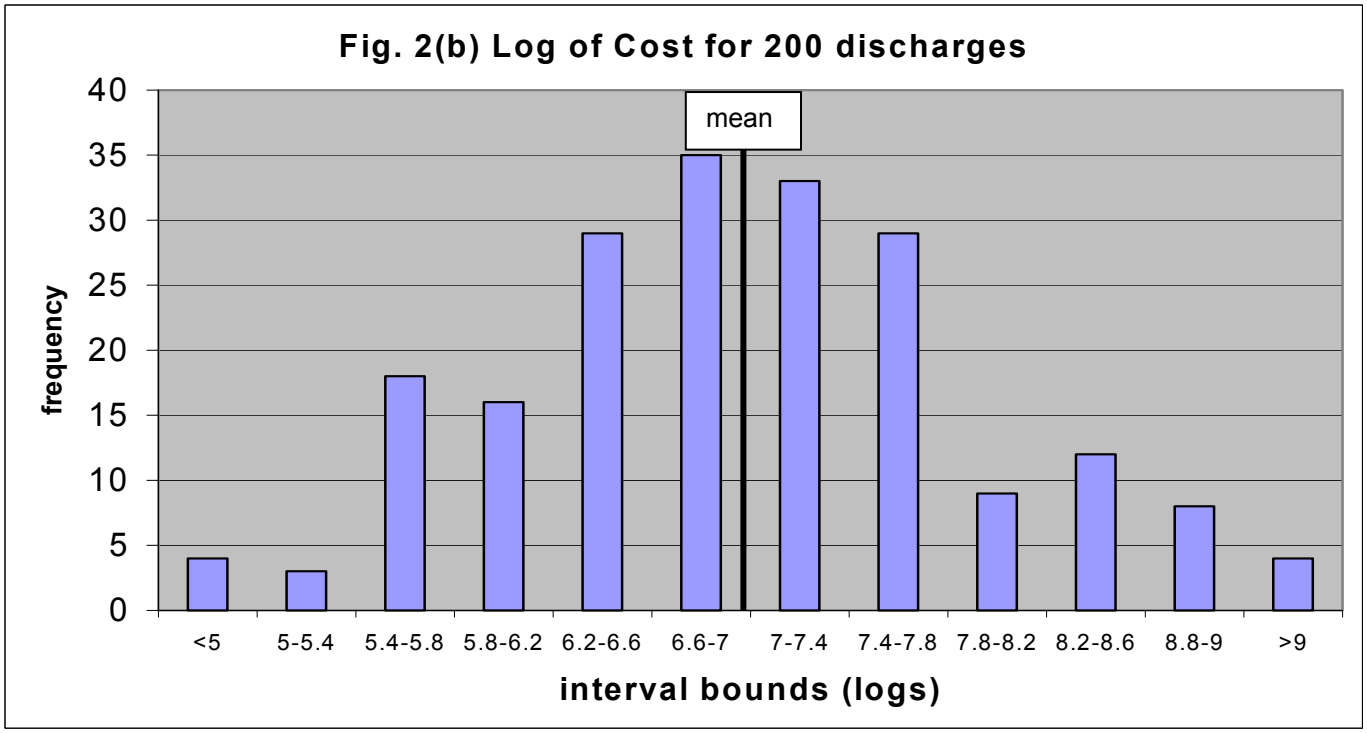


(b)



A numeric illustration:
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).





Geometric vs. 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 uses the same principle, but instead of adding the values, they are multiplied together and instead of dividing by n, the n^{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 just 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 back toward the center of the data, the geometric mean is more "robust" than the arithmetic mean. "Robust" here means less influenced by outliers.

Back to the cost example from 200 hospital discharges:

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.

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

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

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

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

geometric mean risk ***exp***$(\bar{\hat{y}}_{jl})$ 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 $\hat{y}_{ijkl}$ = 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

Modeling Requirements:

1. Create an additional column in the patient/episode table to hold ln_x, ln_x_risk, x_risk($e^{ln\_x\_risk}$) and ln_x_stderr where x represents the dependent variables.

2. Populate this column with the ln(total_charge), ln(comparative_costs), and ln(ccms_length_of_stay) respectively. The ln values will be populated with a '99' when costs and charges are zero.[20]

3. Regress the ln(total_charge), ln(comparative_costs), and ln(length of stay) on the original vector of independent variables. Cases with a null value or a 99 for the dependent variable as well as incomplete cases will not be included in the regression. Nevertheless fitted values (risks) and their standard errors will be generated for **all** complete cases[21]. We shall use **n** to designate the number of complete observations including those with null or '99' dependent values; **m** indicates the number of observations included in the regression (excluding incomplete and those with null or '99' dependent values). To illustrate, suppose a given model stratum has 100 observations of which 95 are complete; and of these 95, ten have cost equal to zero (are given a value of '99' in the log column). Then n=95 and m=85. The regression is run on m=85 cases and fitted values (risks) together with their standard errors are generated for n=95 cases.

4. The back-end values are left in log form and antilogs are applied only after aggregation on the front end.

[20] 99 is a placeholder used by Data Manager to identify observations that should be excluded from the regression because the dependent variable is undefined (ln of 0 is undefined).
[21] Complete cases are defined as having values for all independent variables required for the regression.

5. The front-end software application then performs the appropriate calculation (sum, average, etc.) on the **log** values to display the raw, standardized, and deviation results in the reports. (The calculations that are relevant to this conversion are on the previous page.)

6. Deviations are based on geometric means:

   Geometric Deviation = $\exp(\overline{y}_{jl}) - \exp(\overline{\hat{y}}_{jl})$.

7. The front-end software calculates the p-value to determine significance with all measures in logs (i.e. without converting raw, risk, or standard errors to the original units.) The calculation will not change from what is currently in use but will be based on the **m** nonzero cases.

$$t_{jl}^{p} = \left(\frac{\overline{y}_{jl} - \overline{\hat{y}}_{jl}}{\sqrt{V\left(\overline{\hat{y}}_{jl}\right)}} \right)$$

8. The deviations column on all CareScience Quality Manager reports must equal the raw minus the standardized values up to rounding error in the first decimal place, such that the deviation is no more than 0.1 different from the difference between raw and the standardized value.

Implementation Comments:
Implementation is a combination of front-end and back-end changes. The database must hold logarithmic values – ln(total_charges), ln(comparative_costs), and ln(length of stay) – and standard errors in log form. All computations of confidence intervals and significance are in logs, including necessary aggregations. All computations on risk values are done ***before*** conversion back to "levels" (in log units), hence excluding cases with zero values in the raw data. This approach to aggregation generates geometric (not arithmetic) means. Moreover, the log transformation method guarantees that expected level values (after taking the antilog) be positive, which eliminates the need for front-end data trimming.

Comparative Costs as an example:

1. Assignments:
   a = exp(avg(ln_comparative_costs))
   b = exp(avg(ln_comp_cost_risk))
   c = sum(decode(ln_comparative_costs,null,0,1))
   d = sqrt[sum(ln_comp_cost_risk_stderr ^ 2)]
   k = avg(ln_comparative_costs) - avg(ln_comp_cost_risk)

2. Computations:
   Charge deviation = a - b
   Charge sig flags: **t**-value= k * c / d (with degrees of freedom: c - 1)

Addendum on LOS

Within CareScience database, patients discharged the same day as admitted are assigned length-of-stay = 1, not 0. That conforms to most billing practices. LOS is defined as the number of days present, not including the day of discharge with a minimum LOS = 1. This algorithm eliminates the possibility of undefined value of ln(length_of_stay) when LOS = 0.

Appendix C - Select Practice Formulas

1. Determine case-weighted arithmetic means by disease (3-digit ICD-9 code, k) for
   - mortality (T_k)
   - morbidity (B_k)
   - complications (C_k)

   Note: Records without a risk score are excluded.

2. Compute risk-adjusted rate for each facility (h) for each adverse outcome:
   - $T_{kh} = T_k + \Delta T_{kh}$, where Δ indicates mean deviation (actual - risk)
   - $B_{kh} = B_k + \Delta B_{kh}$
   - $C_{kh} = C_k + \Delta C_{kh}$

   Note: T_{kh}, B_{kh}, and C_{kh} can never be negative.

3. Compute quality indicator by disease k for each facility (h) based on CHRP weights:

   $Q_{kh} = 0.46(T_{kh})^{0.96} + 0.29(B_{kh})^{0.91} + 0.25(C_{kh})^{0.94}$

   Note: Smaller Q_{kh} denotes higher quality.

4. Compute efficiency ratio for length of stay by disease k for each hospital h:
   - $RL_{kh} = \exp\{(\text{logLOS})_{kh} - (\text{logLOSrisk})_{kh}\}$ for h×k level means

     Note: Smaller RL_{kh} denotes higher efficiency.

5. Cross tabulate Q_{kh} and RL_{kh} by quintile across facilities for each disease k.

   Note: This is a 5 x 5 cross-tabulation matrix.
   Small facilities with less than 100 cases for disease K were excluded from the tabulation.
   No matrix is constructed with fewer than 200 qualifying facilities.

6. Identify the select practice facilities for each disease k (set H^*_k) by choosing facilities falling into the top 2x2 cells in the 5x5 cross-tabulation matrix.

   Note: Select practice represents roughly 16% (4/25) of cases for each disease k. The reason for the imprecision is that select practice is a facility characteristic, requiring all cases from a given facility be sorted into one quintile only.
   For the 50 highest-volume diseases, the correlations between quality (Q_{kh}) and efficiency (RL_{kh}) are weak.

7. Compute select practice risk for each disease k as
   - T^*_k = case-weighted mean T_{kh} for all h $\in$ H^*_k
   - B^*_k = case-weighted mean B_{kh} for all h $\in$ H^*_k
   - C^*_k = case-weighted mean C_{kh} for all h $\in$ H^*_k
   - RL^*_k = case-weighted geometric mean RL_{kh} for all h $\in$ H^*_k
   - $R\$^*_k$ = case-weighted geometric mean $R\$_{kh}$ for all h $\in$ H^*_k

   Note: Compute ratio for total costs by disease k for each hospital h:

   $R\$_{kh} = \exp\{(\text{logCost})_{kh} - (\text{logCostRisk})_{kh}\}$ for h×k level means

8. Compute the scaling factor for each outcome as the ratio of select practice to mean:
   - $R_{tk} = T^*_k / T_k$ for all k
   - $R_{bk} = B^*_k / B_k$ for all k
   - $R_{ck} = C^*_k / C_k$ for all k
   - $R_{lk} = RL^*_k$ for all k, where l indicates LOS
   - $R_{\$K} = R\$^*_k$ for all k, where $ indicates Total Costs

9. ICD-9 codes that fail to pass Step 5 are rolled up into one of 18 major diagnosis groupings (Dx_Groups) of the ICD-9-CM classification system. For example, ICD-9 codes 460 to 519 (excluding the six codes that could stand alone) are rolled up into Dx_Group 8, which represents the diseases of the respiratory system. The scaling factor for each Dx_Group is computed as outlined by Steps 5 to 8. ICD-9 codes rolled up into a broader Dx_Group share the same scaling factor.

C.1 Applying the Scaling Factors

The purpose of the scaling factors is to generate risk scores based on Select Practice hosptials. Deviation (Actual – Risk) and significance flag can then be computed on the new risk scores. The deviation and significance flags indicate the facility's performance level compared to the nation's Select Practice facilities for the particular disease group.

1. Compute risk score and stderr of Select Practice at patient level:

   Ln_los_risk_select = Ln_Rlk + Ln_length_of_stay_risk
   Ln_los_risk_stderr_select = Ln_Rlk + Ln_length_of_stay_risk_stderr

Ln_cost_risk_select = Ln_R$k + Ln_cost_risk
Ln_cost_risk_stderr_select = Ln_R$k + Ln_cost_risk_stderr

comp_risk_select = comp_risk * Rck
comp_risk_stderr_select = comp_risk_stderr * Rck

mort_risk_select = mort_risk * Rtk
mort_risk_stderr_ select = mort_risk_stderr * Rtk

comp_morb_risk_select = comp_morb_risk * Rbk
comp_morb_risk_stderr_select = comp_morb_risk_stderr * Rbk

(R: Scaling Ratio; k: 3-digit ICD-9 code; L: LOS; $: cost; c: complication; t: mortality; b: morbidity)

2. Report the Select Practice deviation and significance flag at the aggregate level:

Use the same method as the standard-practice report.

Appendix D - Technical Details about Model Specification

Model Stratification

The stratification is roughly based on 3-digit level ICD-9-CM diagnosis codes. A clinical and statistical review was conducted, basing on State All-Payer 1999 data. The stratification processing is described below:

1. A major diagnosis code, representing more than 0.1% of total discharges, stands as a separate stratum. For example, ICD9 410 (AMI) stands as a separate stratum because its volume (2% of total discharges) is greater than 0.1%, the minimum requirement.
2. Within the minor diagnosis codes that fail the volume criterion, a clinically significant diagnosis code stands as a separate stratum. The clinical significance is determined at threshold of 5% mortality rate. For example, ICD9 151 (Malig. Stomach Neoplasm) stands as a separate stratum because its mortality rate (11%) is above the threshold of 5%, even though its volume is lower than 0.1%.
3. Similar diagnosis codes are rolled up into a common stratum. For example, ICD9-204, 205, 206, 207 and 208 are rolled up into Leukemia.
4. The remaining diagnosis codes are rolled up into Broad Diagnosis Groups according to the designation of ICD-9-CM. For example, ICD9_740 - 759 are rolled up into BDG14: Congenital Anomalies.
5. Exemption One: Newborns are determined by three principal diagnosis codes at two-digit level: 76, 77 and V3. If any of a newborn's diagnosis codes begins with 764, 765 or V213, this case is classified into low birth-weight immature newborn. If none, this case is classified into normal newborn.
6. Exemption Two: Major organ transplant patients are stratified separately due to issues specific to this patient group (e.g. heart transplant). They are identified by DRG.

Mortality Exemptions

Hospital level mortality rate is usually around 2 or 3 percent. Expirations do not evenly occur across the 142 model strata. Among some disease strata, mortality rate is very close to zero. For instance, mortality rate is less than one-tenth of a percent among intervertebral disc disorder patients (ICD9-diag 722). It is extremely difficult to build a robust model to accurately pinpoint the very rare expirations. As a result, these disease groups are proposed to be omitted from mortality analysis rather than forced into a poor model.

The CareScience mortality model is based on linear regression, and consequently the predicted mortality risks may fall out of the range between zero and one at the patient level. Out-of-range risks are acceptable unless they exceed the "reasonable range" of -0.5 <= and <= 1.5 at which point they are considered invalid. If negative risks occur in aggregate reporting, they are rounded to zero.

Complications Algorithm

Complications are derived from principal and secondary diagnosis codes. Ideally, complications should be recorded as binary outcomes. However, there is no absolute way to classify diagnoses as complications. Clinically, a diagnosis may be considered a complication in one case but a comorbidity in another. The POA flag (present on admit) is helpful to identify existing conditions prior to admission. But the flag is often unavailable in either public or private data. Moreover, a diagnosis that was captured during an inpatient stay does not necessarily indicate its development after admission. Although chart reviews are a reliable way to supplement this information, they are unsuitable for large-scale data processing efforts. CareScience has therefore developed a unique comorbidity-adjusted complication index (CACI) to approximate the probability that a diagnosis is a complication given its accompanying principal diagnosis. These complication probabilities were determined ex ante by a panel of medical experts (please see section of Clinical Knowledge Base for details).

The following algorithm is in use when complication is calculated:

1. Find the pair of principal and secondary diagnosis in CACI table and select the corresponding probability;
2. If the principal and secondary diagnosis code combination can not be found in the CACI table, the program selects the secondary diagnosis code's default probability (independent of principal diagnosis) from the CACI2 table;
3. If the secondary diagnosis code can not be found in either the CACI table or the CACI2 table, the diagnosis is excluded from the calculation;
4. If the patient has no secondary diagnosis codes, his complication probability rate is set to zero;
5. If the first three digits of the secondary diagnosis code are equal to the first three digits of the principal diagnosis code, the secondary diagnosis is excluded from the calculation;
6. For Obstetrics patients, the program selects probabilities for ALL diagnoses from the CACI2 table;
7. Newborns, as defined by the principal diagnosis code, are not included in complication analyses.

CACR Comorbidity Scores and Chronic Diseases or Disease History

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. If a secondary diagnosis is not present in the Diagnosis_Morbidity table, it receives a designation of "unspecified" and is correspondingly grouped to category U. Secondary diagnoses are subsequently evaluated according to the CACI algorithm. Comorbidities are calculated as

$$\sum_{1\to n}^{s}(1 - \text{Pij}),\ \ j = 1,2,...,n$$

where n is the number of secondary diagnoses, s is the severity category, and p_{ij} is the probability of complication for the jth secondary diagnosis given principal diagnosis i. The probability that

a particular secondary diagnosis is a complication of a given principal diagnosis is retrieved from the CACI table. Probabilities of comorbidities in the same severity category are summed together. As a result, comorbidity score may not be an integer.

Comorbidity score is closely related with complications. Comorbidity scores are calculated by the similar algorithm that is used to calculate complications.

Chronic diseases and disease history are determined from patients' secondary diagnosis codes. *(Note: Chronic diseases were previously included in patients' comorbidity scores.)* To differentiate patient characteristics, common chronic diseases enter the model separately from comorbidities. Comorbidities and chronic diseases are restrained to positive coefficients in the model calibration.

Birth Weight and Defining Diagnosis

Neonates represent approximately 10% of all admissions and are therefore an important analysis population. The overwhelming majority of neonates are healthy full-term babies. Attention is focused on high-risk neonates, who are primarily immature, low-weight newborns. Our study shows that birth weight is the most important predictor of survival, treatment pattern, and resource requirements. Weight class is encoded in the fifth digit of immature neonate diagnosis codes (764, 765, and V213). If the fifth digit denotes unspecified weight, the birth weight is set to null, and the record is excluded from outcome analyses. If the fifth digit denotes a birth weight range, the birth weight is set to the midpoint of the range. For example, a fifth-digit value of '4' indicates a weight between 1 and 1.25 kg, and so the birth weight is set to 1.125kg. It should be noted that the fifth digit of V213 indicates a different weight range from that of 764 and 765.

Since 2003 gestational age became a new field in newborn data, however, it has remained inconsistently reported. Due to lack of data, the relationship between gestational age and birth weight has yet to be quantified. In CareScience risk model, if a record only contains a gestational age code, birth weight is not be assigned, and the record is excluded from analysis. If a newborn does not possess a code that indicates immature status, the newborn is assigned to the normal neonatal group for which birth weight is not a risk factor.

The neonatal model does not include procedures and CACR comorbidity scores. These factors are considered less relevant newborn characteristics. On the other hand, certain diagnoses are deemed significant attributes defining a newborn's health status. These codes are directly incorporated into the neonatal model at the three digit level and are called defining diagnoses.

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 (ccms_crl_group_by) 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 the *beta* tables to find the procedure's corresponding coefficient. If the coefficient is not present in the *beta* tables, its value is set to zero. Category 1 procedures with coefficients of zero have no impact on the risk score. (It should be noted that although a procedure may be considered clinically relevant, it may not be statistically significant for a particular outcome. Procedures failing to be statistically significant are not included in the model and have no impact on the risk score.)

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. More specifically, the Valid Procedure table contains a field called 'Timing' that specifies the maximum period of time from admission during which a procedure must occur to be included in the model. For example, a Timing field value of "48" indicates that a procedure may enter the model if it occurs within the first 48 hours of the hospital stay. For patient records, timing of individual procedures can be calculated as the difference between the Admission_Date and the Diagnosis_or_Procedure_date. If the difference is within the timing requirement in the Valid Procedure table, the procedure will be counted, and the algorithm, as described for Category 1, will be applied. If the timing field is missing, the procedure will be excluded.

For several disease strata (ccms_crl_group_by), the risk model does not incorporate valid procedures. These groups include Normal_Neonates, Immature_Neonates, DRG 103, DRG 480, DRG 481, DRG 495, DRG 512, and DRG 513.

Appendix E – Technical Details about SAS Programming

Data Transforming

1. Transforming Date-Time Format

SAS/Access software allows SAS to pull out data directly from Oracle database. Most formats, except date-time, are kept the same in SAS data set. Oracle date-time formatting is recognized in the SAS environment, however, SAS automatically transforms Oracle formatting into a unique SAS date-time format when Oracle data are read into SAS. This feature affects numerous timing-related fields, including Treatment_or_Admit_Date, Inpatient_Disch_Date, and Diagnosis_or Procedure_Date, as well as numeric fields derived from timing-related fields. Time_Trend, a field derived from Inpatient_Disch_Date, falls into the latter category. Because SAS date-time format is used in the subsequent SAS data processing environment, care is taken to ensure the accuracy of variables whose format has been transformed.

2. Transforming Categorical Variables

For some categorical variables, the transformation is simple and straight forward (e.g. sex can easily be converted to a dummy variable with value '0' for all male patients and '1' for all female patients.) Some categorical variables, e.g. admission source, may consist of multiple categories. But each visit has only one corresponding category. They can be directly incorporated into model statement with '*class*' option. They can also be transformed into dummy variables with *'if, then'* clauses. For the purpose of manipulating parameter estimates and covariance matrix, the latter method is preferred.

Some categorical variables, on the other hand, may consist of various values for each visit. For example, one visit may have three chronic conditions while the other has none. For this kind of categorical variables, '*class*' option is invalid, and *'if, then'* clause is not technically feasible. SAS/Macro offers a solution to this issue. The details are elaborated in the section of Macro Function.

3. Adding New Variables

When the risk model is recalibrated, new variables are often introduced. The following three methods are commonly used to obtain new variables. First, some variables may not be included in the previous model calibration; but they do exist in the database For example, Race has long been available in client data; but it has been incorporated into model only after 2005. Second, new variables may not exist in the current data but can be derived from existing fields. Birth weight, which can be derived from existing diagnosis codes, is one such example. Third, new variables may not be present or able to be derived in the current data but can be requested. In these instances, Research proposes changes to the Master Data Specification requirements that

define what fields are collected from clients. After sufficient accumulation of the data, the new variable can be implemented into the model calibration.

4. Modifying Existing Variables

The introduction of new variables may require modification of existing variables. For example, the separate inclusion of chronic conditions as independent variables in the model necessitates the adjustment of the CACR Comorbidity Score.

Alternative Model Selection Options

1. Forward Selection

Forward selection begins with no variables in the model. For each of the independent variables, the forward selection process calculates F statistics that reflect each variable's contribution to the model if included. The p-values for the F statistics are then compared to a user-specified critical value for entering the model. If no F statistics have a significance level greater than the critical value, the forward selection process stops. Otherwise, the process continues by adding the variable with the largest F statistic to the model. The iterative process continues until no remaining variable produces a significant F statistic. Once a variable is added to the model in the forward selection process, it remains in the model.

2. Backward Selection

Backward selection begins by calculating statistics for a saturated model that includes all independent variables. The variables are then deleted from the model one at a time until the only variables remaining in the model produce F statistics greater than the critical value specified for staying in the model. At each step, the variable showing the smallest contribution to the model is deleted. Given the number of potential variables in our models, backward selection is less efficient than forward selection, since only a handful of variables typically meet the significance requirement.

Macro Function

SAS/Macro language employs two main devices: Macro variables and Macro processing. Macro variables enable SAS users to dynamically modify text in a SAS program through symbolic substitution. When the substitution expands to compiled SAS programs, the term 'Macro' is applied. When a Macro is called, the compiled programs are executed automatically. This feature called Macro processing. The benefits of Macro language are obvious: reducing the number of SAS statements in a program, decreasing manual mistakes, and saving time for SAS user. But Macros do NOT reduce data processing time. The following example illustrates how the Macro language works.

Patients may be treated with multiple procedures. Depending on the patient's disease stratum, some procedures may qualify as valid procedure candidates and require dummy variables assigned to them. The conventional SAS command for assigning dummy variables is *'If … then…; else…;'* Because the valid procedure list is disease-specific and contains thousands of rows, the SAS command for assigning dummy variables must be repeated tens of thousands of times, since a record may be mapped to dozens of procedure codes. The scope of this task makes it impossible to accomplish manually, however, the use of Macro programming statements renders it feasible.

More specifically, a series of Macro variables, corresponding to the 142 disease groups, are frist created using the SAS command *'select … into … from …'* Next, a Macro is created to repeat the data processing for the142 disease strata. In the Macro, the same SAS command *'select…into…from'* is used to create a series of Macro variables corresponding to the Valid Procedures of a disease stratum; for each of the valid procedure Macro variables, a SAS program is executed that scans all procedure codes, picking up the patients that were treated by the procedure. The processing is automatically repeated until all valid procedure Macro variables have undergone the process. After all records have been assigned valid procedure dummy variables in a disease stratum, the Macro proceeds to the next disease stratum, and the valid procedure Macro variables are automatically re-written to reflect the different Valid Procedure candidates of the second disease group. The processing continues until all 142 disease strata undergo the Macro.

The Macro function simplifies SAS statements in data processing immensely. But the Macro logic is far more complicated than plain SAS statements, especially when Macro variables or multiple Macros are assembled into one global Macro. To that end, it is always recommended to test Macros independently before implementing them full-scale into a model calibration.

Appendix F - New Methods on Horizon

Logit Modeling of Mortality

1. The concept of Logit Model

Mortality is a binary outcome; a patient either lives or expires upon discharge. Mortality risk is a predicted probability within a (0, 1) interval. A problem with the linear probability model is that valid probabilities are bounded by 0 and 1, but linear functions are inherently unbounded. A logit model can resolve this discrepancy.

Transforming the probability to an odds ratio removes the upper bound. The lower bound is then removed by taking the logarithm of the odds. Setting the result equal to a linear function with the explanatory variables yields a logit model. For k explanatory variables and i=1,…, n individuals, this model is expressed as

$$\text{Log}\,[Pi / (1- Pi)] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

where Pi is the probability that Yi=1. The expression on the left-hand side is usually referred to as the logit or log-odds. Similar to an ordinary linear regression, the x's may either be continuous or dummy variables. The logit equation can be solved for pi to obtain

$$Pi = \text{EXP}\,(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) / (1 + \text{EXP}\,(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}))$$

We can further simplify this equation by dividing both the numerator and denominator by the numerator itself:

$$Pi = 1/ (1 + \text{EXP}\,(-\alpha - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik}))$$

The resulting equation has the desireable property that regardless what values are substituted for the β's and x's, Pi will always be a number between 0 and 1.

2. The Algorithm of Calculating Risk

Logit model implementation requires the following equations to calculate risk at the patient-level:

[1] $\text{Logit of Mortality Risk} = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \ldots + \beta_n * X_{in}$

The Standard Error of Logit is defined as:

[2] $\text{Logit Mortality Risk Standard Error} = \text{Sqrt}(\, X_i\, [\text{COV}_b]\, X_i')$,

where COV_b is the nonlinear ML estimated variance-covariance matrix of the estimated β coefficients.

Note that the *Log of the Mortality Risk* is NOT the same as the *Logit of the Mortality Risk,* since the logit value (Logit p) is equivalent to the log of the ratio that represents the risk (log (p/1-p)). Therefore, the logit ratio can not be equal to a log ratio. The same logic applies to the Standard Error value.

3. The Algorithm at Front-End Report

When reporting Mortality Risk, CareScience analytics first determine the arithmetic average Logit Mortality Risk and then transforms it into the real Mortality Risk using the following equation:

[3] Aggregate Mortality Risk = [1/(1+exp((-1)* Avg(Logit Mortality Risk)))]*100

Mortality Risk will always be a number between zero and one. The Raw Mortality Rate and deviation are computed as:

[4] Raw Mortality Rate = (Number of Expired Cases/Number Of Eligible Cases)*100

[5] Deviation = Raw Mortality Rate – Mortality Risk

4. Computing the Significance Flag of the Mortality Deviation

"Significance flags" indicate whether the deviation could plausibly be interpreted as the probability that the results could have occurred randomly if there were no a true underlying effect. In other words, significance flags represent the probability or confidence interval at which the value generated from a variable in the raw data (sample) reflects the value generated from the same variable in the calibration data (entire population).

Rounding of deviation values is performed to the first decimal place. For example, a value of 0.04% is rounded to 0.0%. Similar to other outcomes, if the Mortality Deviation equals 0.0% after rounding, the significance level is not computed. Accordingly, the calculations below are not performed.

In order to determine the significance level of the Mortality Deviation, the following algorithm is applied:

a. First, transform the Raw Mortality value into the Logit Raw Mortality value by applying the following steps and equation[22]:

[22] Note: ABS =Absolute and AVG = Average

[6] Logit Raw Mortality = ln(Raw Mortality/(1 – Raw Mortality))

b. Compute the Logit Mortality Deviation as:

[7] Logit Mortality Deviation = Logit Raw Mortality – Logit Mortality Risk

c. Calculate the T_VALUE (Z_VALUE) as:

[8] T_VALUE = ABS{(Logit Mortality Deviation)/ [AVG(Logit Mortality Risk Standard Error)]}

d. Compare the T_VALUE to the T_DISTRIBUTION table and obtain the significance level.

Note: If the Raw Mortality Rate is either 0 or 1 OR the Average Logit Mortality Risk Standard Error is 0, the Logit Raw Mortality value is considered *undefined.* The significance level is set at the highest significance level (i.e. 90%).

5. Concerns about Implementing Logit Model

In-hospital death is rare among many patient populations. At hospital level, the survival-death split is around 98- 2. The split can be more extreme among many patient populations. For a given sample size, the standard errors of the coefficients depend heavily on the split on the dependent variable. As a general rule, we are better off with a 50-50 split than with a 95-5 split. Logit model has a unique sampling property. We can do disproportionate stratified random sampling on the dependent variable without biasing the coefficient estimates. The intercept does change under such sampling schemes. This solution is extremely useful in our situation. But the data set has to be specifically tailored to each disease stratum.

Convergence failure is a common issue with logit model. We know that most of the independent variables are categorical variables. They enter the model equation as series of dummy variables. Some of the dummy variables may have the following property: at one level of the dummy variable either every case has a 1 on the dependent variables or every case has a 0. That causes complete separation or quasi-complete separation. In either case, logit model will not converge. Removing problematic dummy variables can achieve convergence. It is equally effective to collapse uncommon categories. But again, the data set has to be specifically tailored to each disease stratum.

Logit model does not generate standard errors for the model. Only the covariance matrix of parameter estimates could be used to calculate standard error for individual patients during out-of-sample prediction. Therefore, the significance level on the front-end report is less robust, comparing to that of linear model.

At the aggregate level, logit model generates similar results to linear model. At the patient level, however, logit model offers better face validity. Implementing the logit model is a costly endeavor requiring the overhaul of current methods and programs. From a methodological perspective, the logit model offers greater advantages in the long term.

Prospective Risk

A specialized version of the model deals with "prospective" patient risk for adverse outcomes and resource demands. This model restricts explanatory patient risk factors to those known and recorded upon admission to the hospital. While most CareScience patient-level variables qualify for inclusion, the model excludes discharge disposition factors and most procedure information. The comorbidity score for this version of the model is based purely on known chronic conditions rather than on all secondary diagnoses from the discharge abstract. The model compensates for information unavailable on the day of admission by including information from prior hospital admission by linking visit records to unique patient identifiers.

Still to be developed is a version of the CareScience prospective risk model that makes use of test results and signals clinical findings within the first 24 hours of admission. An increasing number of CareScience clients are supplying such clinically rich data elements, but we are not yet able to construct a database that is large enough to build a statistically robust model.

Discharge Disposition Model *(To be added)*

Modeling Complication as a Binary Outcome *(To be added)*