

---

---

# Comparing Hospital Performance Across Time Periods

James C. Vertrees, Ph.D., Elizabeth C. McCullough, M.S., Mona Z. Zhang, B.S., Richard F. Averill, M.S.

---

---

## Introduction

It is increasingly common in the United States for hospital performance comparative data to be disseminated to the public. Since comparative reporting is both more meaningful and more accepted by hospitals if the reports are adjusted for severity of illness, more than 20 states in the United States use All Patient Refined DRGs (APR-DRGs) for comparative reporting purposes.

Since mortality data is readily available and represent one indicator of quality, mortality rates are commonly included in the comparative reporting process. In addition, states may also provide comparative cost and length of stay information. These hospital performance reports generally compare hospitals in terms of each hospital's difference from an expected value computed based on a statewide average. For example, hospitals may be ranked in terms of the magnitude of the difference between the hospital's actual experience and the expected value on a severity of illness adjusted basis using APR-DRGs.

## The Problem

Mortality is a rare event. This means that mortality estimates at a local area or hospital level can be misleading due to the stochastic (random) nature of mortality. For example, if deaths for a specific cause go from one to two cases in a year, the mortality rate for this cause will have doubled. Clearly, the doubling of a mortality rate is, in this instance, unlikely to be a

cause for serious alarm.

In some instances the comparison of hospitals to a statewide severity of illness or risk of mortality adjusted "norm" has been done without taking into account the statistical significance of the observed difference between the actual and expected values. The problem with not considering the statistical significance is that the differences which are not statistically significant are not stable over time.

For example, when one state compared the rank order of the difference between actual and expected mortality for all hospitals using the data from one year against the rank order of the same hospitals computed using data from the following year, the correlation between the two sets of hospitals was low. This was counter intuitive, as one would expect the best hospitals from the first year to remain highly ranked when the second year's data are used. However, the comparison included all hospitals irrespective of the statistical significance of the difference between actual and expected value.

## Purpose

The purpose of this article is to illustrate both the problem of unstable estimates of differences between actual and expected values over time and to illustrate the use of standard tests of statistical significance as a solution to this problem. Further, the article will determine if APR-DRG adjusted hospital rank orders of the difference between actual and expected charges, length of stay and mortality are stable over

time.

### Measuring Severity of Illness and Risk of Mortality

APR-DRGs refine the basic concept underlying the Medicare Diagnosis Related Groups (DRGs), which are used for payment in the U.S. Medicare systems, by adding two sets of four subgroups to each of the base DRGs. One set of subgroups addresses patient differences relating to severity of illness and the second set addresses risk of mortality. In APR-DRGs, severity of illness is defined as the extent of organ-system loss of function or physiologic decompensation, while risk of mortality is defined as the likelihood of dying. Since severity of illness and risk of mortality are distinct patient attributes, separate subgroups are assigned to a patient.

The four severity-of-illness subgroups and the four risk-of-mortality subgroups represent minor, moderate, major or extreme severity of illness or risk of mortality. The assignment of a patient to one of these four subgroups takes into consideration not only the specific secondary diagnoses, but also the interaction between secondary diagnoses, age, principal diagnosis and certain procedures.

### Data

Claims data from 544 hospitals in California covering 1996 were used for this study. These data consisted of 3,318,972 observations (discharges) divided into two six month time periods. The first six months of data contained 1,821,123 cases; the second six months contained 1,497,849 cases. Cases in error DRGs were excluded from all analyses. In addition, cases with a length of stay of zero days or more than 365 days or charges less than \$100 or more than \$750,000

was also excluded from the analyses. Charge and length of stay information were available in both periods for 500 hospitals. Mortality information was available for both periods for 528 hospitals.

### Computing Expected Values

The expected value of charges, length of stay or mortality for a hospital is computed by using the APR-DRGs to adjust for the hospital's mix of patients. The expected value for a hospital is the average charge, average length of stay or mortality rate that would result if the mix of patients in the hospital had the same average charge, average length of stay or mortality rate by APR-DRG and subclass as the statewide database. The computation of the expected value for a hospital is as follows.

$h$  = Hospital

$g$  = Base APR-DRG

$s$  = Subclass

$R(g,s)$  = Average charge, average length of stay or percent died in APR-DRG  $g$  in subclass  $s$  (either severity of illness or risk of mortality) in the statewide database

$N(h,g,s)$  = Number of patients in hospital  $h$  in APR-DRG  $g$  in subclass  $s$

Relative to the statewide database, the expected value for hospital  $h$  is:

$$E(h) = \frac{\sum_{g,s} R(g,s)N(h,g,s)}{\sum_{g,s} N(h,g,s)}$$

Thus, the expected value is computed using a statistical approach generally referred to as indirect rate standardization.

## Test of Significance

Observed differences between a hospital's performance and the statewide norm can represent a true difference in performance or can be caused by random variation. Statistical methods can be used to determine which differences in resource use or outcomes are true differences and which may be the result of random variation. The statistical methods give the probability that an observed difference in performance between the hospital and the norm is due to random variation. A difference in performance between hospital and norm is considered "statistically significant" if this probability is small. A difference is considered statistically significant at the 0.05 level if the probability that the observed difference is due to random variation is five percent or less (i.e., less than one chance in twenty). Significance at the 0.01 level means that this probability is one percent or less.

Three interrelated factors determine whether a difference in performance is statistically significant: the number of observations, the magnitude of the observed difference in performance, and the variability in performance of the hospital and of the norm. A small number of patients, a small observed difference in performance, or high variability within either the hospital or the norm (i.e., a high standard deviation) increase the likelihood that the observed difference was due to chance and does not represent a true difference which will persist over time.

Conversely, a large number of patients, a large observed difference between hospital and norm, or low variability within both hospital and norm make it more likely that the difference was not due to chance and does represent a true difference, i.e., a difference that will persist over time. An observed difference of the same magni-

tude may be statistically significant in one comparison and not in another. For example, a half-day difference in average length of stay for normal delivery (with a large number of patients and low variability) is unlikely to be due to random variation and would probably be statistically significant. In contrast, for transplant surgery with few patients and high variability of length of stay, a half-day difference is more likely to be due to random variation and not be considered statistically significant. The conclusion that a difference is statistically significant indicates that the hospital and the norm have a true difference in performance, which is likely to be repeated in future data. For normal delivery, an observed difference of a half-day may be enough to reach this conclusion, but the same observed difference may yield only weak, inconclusive evidence for transplants.

The comparison of a hospital's performance to a norm requires the use of several distinct statistical methods. Resource variables such as length of stay and charges are continuous variables that, in general, are lognormally distributed, while outcome variables are binary variables that indicate the occurrence or non-occurrence of an event such as death. Different statistical methods are required for continuous and binary variables.

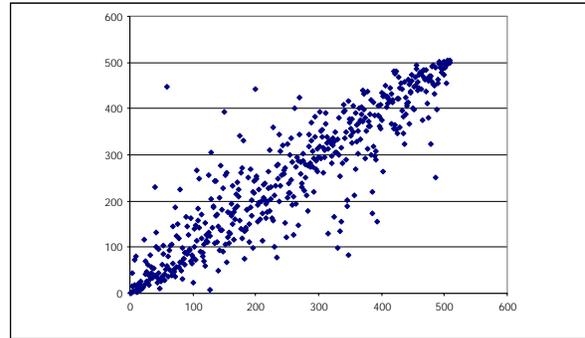
For continuous variables, the test of significance of the difference between actual and expected hospital performance can be based on the principle of stratification. The average of the hospital and the norm are compared separately within each APR-DRG and subclass. A weighted average of the difference in averages, one per APR-DRG and subclass is formed. If the averages for the hospital and norm are equal, the weighted average difference should be close to zero. The weighted average dif-

ference has a normal distribution and is considered statistically significant if the value of the difference has a value outside of the critical values of the standard normal distribution for a two tail test at a specified level of significance.

For binary variables, the test of significance of the difference between actual and expected hospital performance can also be based on the principle of stratification. The mortality rates are compared separately within each APR-DRG and subclass and then pooled across APR-DRGs and subclasses. The statistical method is the Cochran-Mantel-Haenszel test, which uses one 2 x 2 table for each APR-DRG and subclass (Mantel, Haenszel, 1959).

### Illustrating the Problem

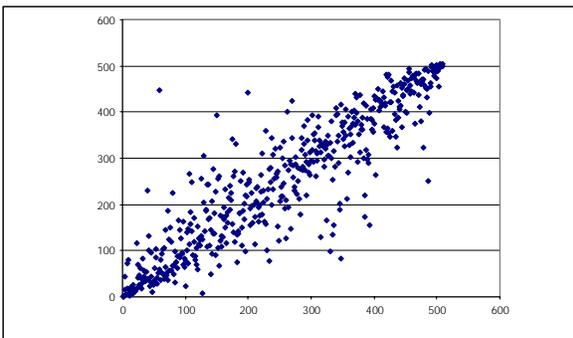
To illustrate the problem, hospitals were ranked from high to low in terms of the difference between their actual and expected value across all APR-DRGs. This was done for charges, length of stay and mortality. Figure 1 plots these rank orders for charges based on the first six months data against the rank orders for the same hospitals using the second six months data. Figure 2 presents the same information for length of stay, and Figure 3 presents the



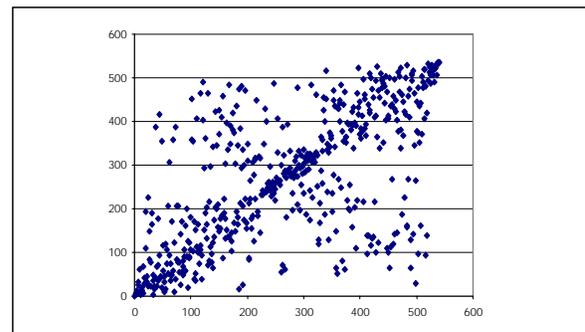
**Figure 2** Rank Order of Difference between Actual and Expected LOS  
First Time Period vs. Second Time Period  
500 Hospitals  
Rank Order Correlation: 0.9084

same information for mortality (i.e. percent died). If the rank orders were stable, the plots would be close to a 45-degree straight line. That is, if the rank order is stable, a hospital will have the same (or similar) rank in the second six months as in the first.

As the scatter diagrams indicate, the relationship for actual versus expected charges is, for most hospitals, close to a straight line. However, there are clearly hospitals where this relationship does not hold. A similar scatter plot for length of stay is more random than for charges while the relationship for mortality departs substantially from a straight line. The



**Figure 1** Rank Order of Difference between Actual and Expected Charges  
First Time Period vs. Second Time Period  
500 Hospitals  
Rank Order Correlation: 0.9527



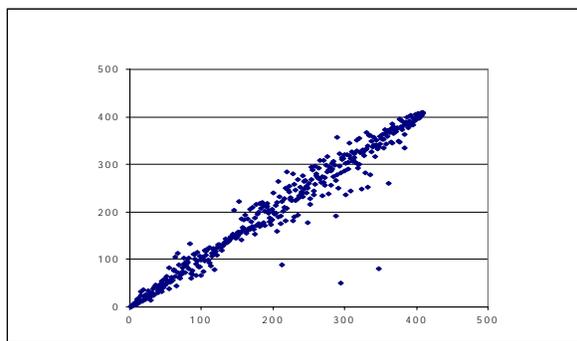
**Figure 3** Rank Order of Difference between Actual and Expected Mortality  
First Time Period vs. Second Time Period  
528 Hospitals  
Rank Order Correlation: 0.6587

results for mortality depart further from a straight line than the results for charges and length of stay because mortality is a rare event, and, therefore, more difficult to measure with statistical precision.

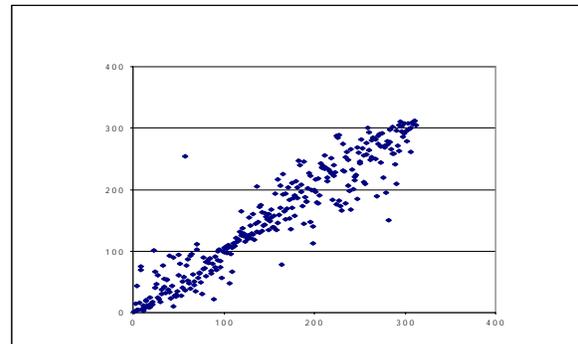
The scatter plots presented above give a sense of the relationship (or lack thereof) between actual and expected values over time. A summary measure of this relationship is given by the correlation coefficient. Correlation values close to 1.0 indicate a close relationship between the measure in time period 1 and the same measure in time period 2. The correlation coefficients are based on all observations, irrespective of whether the observed difference between actual and expected values was or was not statistically significant. The rank order correlation for charges is 0.9527, for length of stay 0.9084, and for mortality 0.6587. As these results contain the same information as the scatter diagrams, the pattern is the same.

**Illustrating Tests of Statistical Significance as a Solution**

Figures 4 through 6 presents the same information as was presented in Figures 1 through 3, but restricted to hospitals where

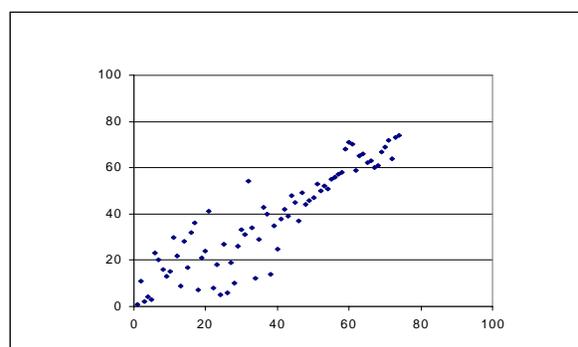


**Figure 4** Rank Order of Difference between Actual and Expected Charges  
 First Time Period vs. Second Time Period  
 Difference Statistically Significant at 0.05 Level in Both Time Periods  
 410 Hospitals  
 Rank Order Correlation: 0.9716



**Figure 5** Rank Order of Difference between Actual and Expected LOS  
 First Time Period vs. Second Time Period  
 Difference Statistically Significant at 0.05 Level in Both Time Periods  
 312 Hospitals  
 Rank Order Correlation: 0.9428

the difference between actual and expected value was statistically significant at the five percent level in both time periods. As is evident from Figures 4-6, restricting comparisons to hospitals where the difference between the actual and expected value was statistically significant greatly improves the stability of the rank order of the hospital over time. This is expected since differences, which are not statistically significant, can, by definition, arise from chance alone. Therefore, it is not expected that the rank order of differ-



**Figure 6** Rank Order of Difference between Actual and Expected Mortality  
 First Time Period vs. Second Time Period  
 Difference Statistically Significant at 0.05 Level in Both Time Periods  
 74 Hospitals  
 Rank Order Correlation: 0.9049

ences between actual and expected values which are not statistically significant, would be stable over time as the difference may only reflect the innate variability of average values and occurrence rates in small samples.

The number of hospitals where the observed difference was statistically significant is less than the total number of hospitals. For example, for mortality the number of hospitals with a statistically significant differences between actual and expected mortality at the five percent level in both time periods is 74 of the 528 hospitals. In contrast, 410 out of 500 hospitals had a statistically significant difference in both time periods for charges and 312 out of 500 hospitals had a statistically significant difference in both time periods for length of stay. Since mortality is a rare event, it is more difficult to have sufficient data to reach the conclusion that an observed difference is not due to chance.

Table 1 summarizes the results of applying a test of significance to the difference between actual and expected values. As the results in Table 1 demonstrate, the application of a test of significance improves the rank order correlation across the two time periods for charges, LOS and mortality. The improvement is particularly substantial for mortality (53.7 percent increase).

Variable	All Hospitals		Hospitals with Statistically Significant Difference	
	Number of Hospitals	Rank Order Correlation	Number of Hospitals	Rank Order Correlation
Charges	500	0.9527	410	0.9716
LOS	500	0.9084	312	0.9428
Mortality	528	0.6587	74	0.9049

**Table 1** Rank Order Correlations for all Hospitals and Hospitals with Statistically Significant Difference between Actual and Expected Values

## Conclusions

Any comparison of hospital performance

to normative data needs to include a statistical test of significance. Observed differences between actual and expected values can be due to chance and this may not represent true differences. Hospital rankings of the difference between actual and expected values adjusted for severity of illness and risk of mortality using APR-DRGs were found to be stable over time for those hospitals for which the difference between actual and expected values was statistically significant.

## References

Mantel, N, Haenszel, W, "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease", *Journal of the National Cancer Institute*, Vol. 22, 1959, p. 719-748.

© 3M 1999