White Paper

Enhancing the Reliability of Physician Performance on Hospital Outcome Measures

Robert Sutter, RN, MBA, MHA
Brian Waterman, MPH

October 2013
Table of Contents

Introduction ......................................................... 1
Reliability of Provider Performance Measures ............. 1
Reliability Adjustment ............................................. 2
Results ................................................................. 2
Summary ............................................................... 5
References ............................................................ 6
About the Authors .................................................. 7
Introduction

Physician executives seeking to assess and improve physician performance on patient outcome measures face a number of specific challenges in practice. In previous work, we have discussed the importance of using appropriate analytics to determine the extent to which variation in patient outcomes can be attributed to physicians versus other aspects of hospital care.\(^1\) Research has also shown that physician-level performance profiles are frequently unreliable, due largely to small sample sizes.\(^2,3\) Failure to appropriately handle such challenges in practice can lead to incorrect conclusions, waste of limited quality improvement resources, and, in some cases, alienation of physician partners.

Reliability of Provider Performance Measures

Reliability is a key component of any effective performance measurement system. The reliability of a measure is the extent to which the result is the same on repeated trials.\(^4\) For example, measuring a physician’s inpatient average length of stay (LOS) for a clinical condition is a common performance measurement. If it is reliable, this measure would be the same when measured every month, quarter, year, etc. However, repeated measurements are never exactly equal: The physician’s average LOS for a given clinical condition is never exactly same from month to month, for example.

Reliable measurement systems have a tendency to be consistent over time with repeated measurements. More consistent results over repeated measurements increase the reliability of the measurement system, while less consistent results lower the reliability. It is important to note that all measurement systems are subject to a certain amount of inconsistency and random error, and the amount can vary from small to large.

Random error comprises all of the chance factors that confound the measurement system.\(^4\) The amount of random error in a measurement system is inversely related to the reliability of the measurements. That is, the more random error present in a measurement system, the less reliable are the measurements. This implies that a reliable measurement is not greatly impacted by random error.\(^4\)

The amount of random error present in a summary measurement is also affected by the sample size in an inverse manner. Measurements with larger sample sizes are on average more reliable than those with small sample sizes. For example, physicians with larger case counts will have more reliable inpatient average LOS measurements compared to physicians with small case counts.
Physicians with small case counts can be problematic in practice for the physician executive. It is difficult to discern whether a high or low inpatient average LOS is due to random error or the quality of care provided. For example, a very sick patient with an unusual clinical condition and comorbidities who experiences an unusually long inpatient LOS would constitute a random error — which may notably increase the physician’s average inpatient LOS.

**Reliability Adjustment**

In recent years, statistical models for handling reliability concerns with risk-adjusted outcomes have emerged. These more sophisticated methods are becoming increasingly popular due to growing public visibility in high-profile outcomes measurement efforts like the Value-Based Purchasing Program from the Centers for Medicare & Medicaid (CMS). Unlike standard risk-adjusted performance indices that express provider-specific performance based solely on the data observed and expected for the provider, these techniques derive estimates that are “shrunken” to the average of all sampled providers to varying extents based on the reliability of individual estimates. This approach has been shown to dampen the influence of noise that occurs in small samples and produce optimal standard errors and confidence intervals for the entire set of physicians of interest.

In order to examine the impact of accounting for variation in physician-level measures of risk-adjusted excess LOS (the difference between observed and risk-adjusted expected), we utilized a random effects hierarchical linear model. The model produced reliability-adjusted estimates and compared results with results based on standard applied methods with a sample of 1,195 congestive heart failure patients seen by 48 attending physicians over a 3-year period at a medium-sized hospital in the southeast United States.

**Results**

Figure 1 depicts the impact of applying reliability adjustments to heart failure risk-adjusted excess LOS. The length of the red line indicates the degree of adjustment to a physician’s measurement; the longer the line the greater the adjustment. The variability in the degree of adjustments is notable: The range of adjustment is -0.75 days to 2.7 days, and 25 percent of physicians had an adjustment of at least 1.3 days. This implies that a noteworthy amount of noise has been removed from these measurements.

**Figure 1: Risk-Adjusted Excess LOS, Physician Standard vs. Reliability-Adjusted Mean**

Source: Truven Health Analytics™ research, strategic service engagement.
The reliability adjustment is a function of both the physician case count (which reflects the noise) and the degree of variation across physicians (the signal). For example, physicians with small case counts have lower reliability and are shrunk more toward the average risk-adjusted excess LOS. Whereas for physicians with large case counts, the risk-adjusted excess LOS measurements are more reliable and are shrunk less toward the average risk-adjusted excess LOS.

The reliability-adjusted, risk-adjusted excess LOS is considered the best estimate of a physician’s true performance. Incorporating reliability adjustment can have a notable impact on evaluating physician performance. In Figure 2, the physicians depicted in red changed performance categories after reliability adjustment is applied. Note that physicians 5, 12, 14, and 46 went from “As Expected” to “Worse Than Expected” after reliability adjustment. And physicians 6, 17, and 22 went from “Better Than Expected” to “As Expected” after reliability adjustment. One physician, 31, went from “Worse Than Expected” to “As Expected.”

**Figure 2: Heart Failure Risk-Adjusted Excess LOS**

![Standard Confidence Intervals](image)

![Reliability-Adjusted Confidence Intervals](image)

*Source: Truven Health Analytics research, strategic service engagement.*

Overall 16.7 percent of the physicians’ heart failure risk-adjusted LOS performance categorizations were revised by incorporating reliability adjustment. As Figure 3 depicts, the percentage of physicians categorized as “Better Than Expected” was reduced from 10 percent to 4.2 percent after reliability adjustment, and the “Worse Than Expected” category increased from 4.2 percent to 10 percent.

The result of reliability adjustment is that the physician executive has much more confidence that the 10 percent and 4.2 percent of the physicians categorized as “Worse Than Expected” and “Better Than Expected” respectively are accurately portrayed and not the result of random error. Consequently, armed with this data, the physician executive is positioned to make better decisions regarding physician performance improvement opportunities.
Figure 3: Heart Failure Risk-Adjusted Excess LOS, Physician Performance Category

The magnitude of the implications of reliability adjustment can be appreciated by examining the effects across the entire clinical spectrum. We examined the difference in physician performance categories associated with reliability adjustment in 138 clinical reporting groups* where the case count was at least 30 and had at least four distinct physicians. This encompassed 2,057 physician performance assessments.

Figure 4 depicts the following results of applying reliability adjustment:

- The “Worse Than Expected” category increased from 4.9 percent to 13 percent
- The “Better Than Expected” category decreased from 11 percent to 4.9 percent
- The “As Expected” category decreased from 84 percent to 83 percent

Figure 4: All Clinical Reporting Groups Risk-Adjusted Excess LOS, Physician Performance Category

Additional note:

* Truven Health Analytics creates clinical reporting groups by combining clinically similar DRGs.
Table 1 depicts the specific adjustments for each performance category across all conditions. It is worth noting that all performance categories were affected by reliability adjustment. However, no physicians were reclassified from the “Worse Than Expected” to “Better Than Expected” categories or from the “Better Than Expected” to the “Worse Than Expected” categories.

The largest adjustment occurred in the “Better Than Expected” category where 61.9 percent of physicians were reclassified to the “As Expected” category after reliability adjustment. The “Worse Than Expected” category had the second largest adjustment: 21.8 percent of physicians were reclassified to the “As Expected” category after reliability adjustment. And the “As Expected” category underwent the least adjustment with 10.4 percent of physicians reclassified to the “Worse Than Expected” category and 0.9 percent to the “Better Than Expected” category.

<table>
<thead>
<tr>
<th>Traditional Performance Category</th>
<th>Reliability-Adjusted Performance Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worse Than Expected</td>
<td>Worse Than Expected</td>
<td>79 (78.2%)</td>
</tr>
<tr>
<td></td>
<td>As Expected</td>
<td>22 (21.8%)</td>
</tr>
<tr>
<td></td>
<td>Better Than Expected</td>
<td>0</td>
</tr>
<tr>
<td>As Expected</td>
<td></td>
<td>180 (10.4%)</td>
</tr>
<tr>
<td></td>
<td>As Expected</td>
<td>1541 (88.7%)</td>
</tr>
<tr>
<td></td>
<td>Better Than Expected</td>
<td>17 (0.9%)</td>
</tr>
<tr>
<td>Better Than Expected</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>As Expected</td>
<td>135 (61.9%)</td>
</tr>
<tr>
<td></td>
<td>Better Than Expected</td>
<td>83 (38.1%)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>259</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1,698</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2,057</td>
</tr>
</tbody>
</table>

Note: The percentage depicted in parentheses represents the row percent.

Source: Truven Health Analytics research, strategic service engagement.

These results have important implications for the physician executive measuring and evaluating physician performance. A large percentage (61.9) of physicians thought to be performing “Better Than Expected” with standard methods are not really the role models of performance that standard methods would lead one to believe. Also, a fair percentage (21.8) of physicians thought to be performing “Worse Than Expected” with standard methods are not the inferior performers that standard methods would highlight.

Summary
When measuring physician performance, physician executives are faced with the quandary of determining whether the physician performance measurements represent a single or random error. This uncertainty impedes the physician executive’s ability to take appropriate performance improvement actions based on physician performance measurements.

Incorporating reliability adjustment into performance measurement reduces the impact of random error on measurements and produces a better approximation of likely future performance. While not a substitute for traditional metrics, which are core and appropriate for day-to-day management and Ongoing Professional Practice Evaluation (OPPE) reporting, reliability-adjusted measures provide supplemental information that can aid a physician executive when making decisions that require accurate classification of likely performance in the longer term. For instance, when identifying best practices to serve as a model for care path creation, identification of “Better Than Expected” performers is critical; reliability adjustment can be used in this case to help ensure selection of higher-performing examples that are more likely to be stable over time.
Reliability adjustment applied to physician-level performance is a relatively new area. As others have noted, it is important to keep in mind that reliability adjustment adds complexity to the production, interpretation, and utilization of results. Further study is needed to test these methods in practice and explore relevant extensions (i.e., explicit modeling of procedure volumes, which have been shown to be a determinant of cardiac surgery outcomes).

Notwithstanding, these case study results demonstrate that reliability adjustment of physician performance measurements is feasible and provides a view of performance that can serve as a helpful supplement to typical profiling measures. A physician executive who is interested in catalyzing performance improvement through focused, effective physician performance improvement is well- advised to consider the value of incorporating reliability adjustments into his or her performance measurement system.

References
6 Dimick JB, Staiger Do, Birkmeyer JD. Ranking Hospitals on Surgical Mortality: The Importance of Reliability Adjustment. *Hlth Serv Res*. 1614-1629.
About the Authors

Robert Sutter, RN, MBA, MHA
Robert Sutter brings more than 25 years of experience to the Truven Health Strategic Consulting Services practice. He is an experienced quality improvement and data analysis professional with healthcare and other industry experience. As a member of the Strategic Services practice, Mr. Sutter focuses on clinical performance, particularly the improvement of hospitals through quality transformation and transforming data into information. His expertise includes quality infrastructure, data analysis, performance measurement, and process improvement.

Brian Waterman, MPH
Brian Waterman works a Senior Product Marketing Manager for Truven Health Analytics. He is responsible for ongoing positioning and development of a hospital-based clinical performance measurement platform designed to help hospital leaders target meaningful drivers of variation in quality and cost outcomes. Mr. Waterman’s career efforts have primarily focused on applied analytics to support clinical and management decisions around measuring and improving quality. His research interests focus on enterprise surveillance, applied statistical methods, and decision support analytics aimed at measuring and identifying superior care.