**Benchmarking Clinicians**

Farrokh Alemi, PhD

Version of Sunday, December 04, 2017

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In hiring, promoting and managing clinicians, managers often need to understand the efficiency and effectiveness of clinical practices. Managers are focused on the survival of their organization, which often translates to two overarching objectives: improve productivity and increase market share. Clinicians' decision making affects both the organization's productivity and its market share. Poor quality clinicians are bad for the patient as well as for the organization [[[1]](#endnote-1)]. Inefficient clinicians increase cost of care for everyone. Clinicians with poor quality of care affect the reputation of the organization and de facto its market share [[[2]](#endnote-2)]. Managers who ignore poor quality of care among their clinicians and focus on non-clinical issues, are failing to see the real causes of their organization's malaise. If a manager is serious about improving the long term financial performance of his/her organization, he/she has no choice but to address clinicians' practice patterns.

For a long time, managers have avoided addressing the quality of clinical decisions on the grounds that they do not have sufficient training to understand these decisions and because such managerial interventions would be an un-welcomed intrusion in the patient/provider relationship. But are these criticisms valid? First, managers do not need to know medicine to understand practice patterns. Managers can profile a physician by looking at the outcomes of his/her patients. They may not understand how a patient should be managed but they certainly can understand patient outcomes such as mortality, morbidity, satisfaction, health status and numerous other measures. Managers can then compare clinicians to each other and see who is performing better. Across encounters and over time, the manager can detect patterns and use this information to bring about lasting changes in practice patterns.

Second, the concern that benchmarking intervenes in physician and patient relationship might be a red herring. After all, practice profiles are constructed after the fact, when the patient is gone. Practice profiles do not tell us how an individual patient should be managed, they identify patterns across visits. In short, these profiles leave the management of individual patients in the hands of the physician. There is no interference in these clinical decisions. No one tells the clinician to prescribe certain drugs or to avoid some surgeries for a specific patient. Practice profiles document the net impact of physicians’ performance on groups of patients.

Benchmarked practice profiles can be used in several different ways. Typically, the information is provided back to clinicians, who can compare their performance to their peer and emulate best practices. Sometimes, financial incentives are tied to better clinical outcomes, as in the case of value-based reimbursements. Other times, benchmarking information is used to improve hiring and contracting practices.

**Simple Comparisons, Misleading Results**

A relatively simple method for comparing two clinicians is to compare their average outcomes [e.g. [[3]](#endnote-3)]. A statistical procedure for the analysis of means/rates of two samples (mean/rate of outcomes for the clinician and mean/rate of outcomes for the peer providers) is well established. Excel software contains a tool for such analysis. The methods of comparing means or rates were discussed in previous chapters in this book. Analysts can use these procedures to see if the difference in means is statistically significant.

The direct comparison of means/rates of outcomes makes sense only if the clinician and his/her peer are taking care of the same patient. For example, if patients are randomly assigned to their provider [[[4]](#endnote-4)], then we can be reassured that the two groups are similar. It is not easy to randomly assign patients to different providers. Patients select their own providers and have long established relationships with them. In absence of randomization, a statistical procedure is needed that balances the data and ensures that the patients in the clinicians and his/her peer do not differ in significant ways from each other. Propensity scoring, covariate balancing, and stratification are examples of available methods to statistically balance the data so that patients seen by the clinician and his/her peer do not differ [[[5]](#endnote-5)]. In this chapter, we illustrate the use of covariate balancing in benchmarking clinicians.

**Comparing Clinicians to Peers on Patients of Similar Severity of Illness**

We present the concept of risk adjusted benchmarking in two steps. In the first step, we assume that a measure of patients’ severity of illness is available. A large number of severity indices exist and are reported in the literature under prognostic indices. Later, in the second step, we remove the need for a well formed severity index and show that the analysis can be done using various relevant patient features. The choice of which approach to take will depend on what data are available.

When a severity index is available, then expected outcomes can be estimated. Clinician’s observed outcomes on a set of patients can then be compared to expected outcome of the peer group on the same set of patients. The performance of the peer providers is simulated on the patients’ of the clinician. If is the probability of observing the patient in the severity group "i", is the average outcome for the clinician for severity group "i", then the following formula is used to calculate expected outcomes for the clinician:

The same calculation can be done on the peer clinicians.

The observed peer outcome, is not a reasonable benchmark because it is calculated on a different set of patients than the clinician’s patients. The simulated peer outcome,, adjusts for the differences in the patient populations. It estimates what would have happened if the clinician’s patients were seen by his/her peer. It is calculated by switching the severity of patients managed by the clinician with the patients managed by the peer group.

These simulated values are counterfactual calculations, in the sense that these statistics report the outcomes if the peer physicians had seen the patients of the clinician. The peer physicians have not really done so. We need to simulate what the situation would have been, if they had done so.

The outcomes for the clinician and the simulated outcomes for the clinician are now measured on the same set of patients and therefore are apple-to-apple comparisons. The difference between the two expected outcomes shows the average opportunity for improvement. The total improvement in outcome for the clinician is seen as:

An example can demonstrate. Consider that a clinician and his/her peers have had the outcomes displayed in Table 1. Is this clinician better or worse than the peer providers? To answer this question, the analyst must compare the expected outcomes for the clinician to the expected outcomes for the peer providers simulated on the same patients as the clinician.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Clinician A** | | **Peer Group of A** | |
| Severity of patients | Number of patients | Average length of stay of patients | Number of patients | Average length of stay of patients |
| Low | 20 | 3.1 | 40 | 4.1 |
| Medium | 30 | 3.4 | 40 | 3 |
| High | 70 | 5.2 | 50 | 4.5 |
| **Table 1:  Severity Adjusted Comparison of Performance of Several Clinicians** | | | | |

The first step is to calculate the probability of finding a patient in a different severity groups. This is done by dividing the number of patients in a severity group by the total number of patients seen by the clinician being evaluated. The probability of having a low severity patient is 20/120, medium severity patients is 30/120 and high severity patient is 70/120. This clinician mostly sees severely ill patients. Once the probabilities are calculated, the second step is to calculate the expected length of stay for the clinician:

To understand if 4.4 days is too high or low, the analyst needs to compare this clinician's performance to his/her peer providers. But the peer providers do not see as severely ill patients as does the clinician being evaluated. To simulate the performance of the peer providers on the patients seen by the clinician, the analyst uses the frequency of severity among that clinician's patients to weigh the outcomes of the peer providers:

The clinician, whose data we were analyzing, seems to be less efficient than the average of his peer group. Note that in both analyses we used the same frequency of having low, medium and high severity patients. Therefore, the differences cannot be due to the severity of patients. Of course, the analysis can be misleading if the classification of patients into various severity groups is at fault or if observed differences are due to random variations and not to real practice differences. But if the classification of patients into severity groups is correct, the fact that the projected length of stay for peer providers was lower than the observed length of stay for the clinician being evaluated suggests that the two may differ in their performance.

**Comparing Clinicians When Patient's Severity of Illness Is Not Known**

In the previous section, we divided patients in broad categories of severity (Low, Medium and High) and compared care provided within each category. To calculate these benchmarks, one needs access to a reliable and valid measure of the severity of illness. Sometimes, such a measure is not available or not trusted to adequately measure the full spectrum of severity of the patient's illness. This section provides an alternative method of benchmarking that does not require the availability of a valid and accurate severity index.

When no severity index is available, an analyst must match patients of clinicians and his/her peer feature by feature. If we think of the clinician as cases and the benchmarking physicians as controls, then this is similar to doing a case control study, where covariates in cases are matched to the same covariates in controls, and the average outcomes are examined among matched cases and controls. A great deal has been written on matching and many methods exists to match cases to controls. Here we explore a method that we designed and relies on decision trees. Suppose that the variables j through m describe the characteristics of the patients seen by the clinician. In this situation, the expected observed outcomes for the clinician is calculated as:

Where j through m indicate a particular combination of features of the patients, indicates the probability of the combination of features occurring and indicates the clinician's outcomes when these features were present. We can simulate the performance of the peer group on patients seen by the clinician by substituting the probability distribution of features with the distribution of patients seen by the clinician:

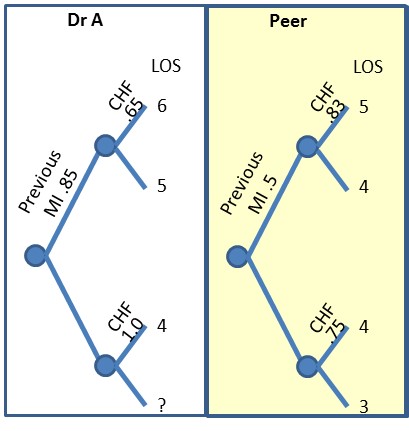
In this simulated calculation, the probabilities are based on frequency of features in the patients seen by the clinician but the outcomes are based on experience of peer providers. Thus we are simulating how the peer providers would have performed if they had the same clinician’s patients.

An example can demonstrate the use of this procedure. Table 4 shows 20 patients of one clinician and 24 patients of his peer providers. These patients were admitted to a hospital for myocardial infarction. In each case, we have recorded two features (existence of a previous myocardial infarction, MI, and presence of congestive heart failure, CHF). Obviously, a patient with a previous MI and with CHF has a worse prognosis than a patient without these features. The analyst needs to separate outcomes for patients with and without specific characteristics.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Clinician's Patients | | | | Peer Provider's Patients | | | |
| Case | Previous MI | CHF | Length of stay | Case | Previous MI | CHF | Length of stay |
| 1 | Yes | Yes | 6 | 1 | MI | CHF | 5 |
| 2 | Yes | No | 5 | 2 | MI | CHF | 5 |
| 3 | Yes | Yes | 6 | 3 | No MI | CHF | 4 |
| 4 | Yes | Yes | 6 | 4 | No MI | No CHF | 3 |
| 5 | Yes | Yes | 6 | 5 | No MI | CHF | 4 |
| 6 | Yes | No | 5 | 6 | No MI | CHF | 4 |
| 7 | Yes | Yes | 6 | 7 | MI | CHF | 5 |
| 8 | Yes | No | 5 | 8 | MI | CHF | 5 |
| 9 | Yes | Yes | 6 | 9 | MI | CHF | 5 |
| 10 | Yes | No | 5 | 10 | MI | CHF | 5 |
| 11 | Yes | Yes | 6 | 11 | MI | CHF | 5 |
| 12 | No | Yes | 4 | 12 | No MI | No CHF | 3 |
| 13 | No | Yes | 4 | 13 | No MI | CHF | 4 |
| 14 | No | Yes | 4 | 14 | No MI | CHF | 4 |
| 15 | Yes | Yes | 6 | 15 | No MI | CHF | 4 |
| 16 | Yes | Yes | 6 | 16 | No MI | CHF | 4 |
| 17 | Yes | Yes | 6 | 17 | No MI | CHF | 4 |
| 18 | Yes | No | 5 | 18 | No MI | No CHF | 3 |
| 19 | Yes | No | 5 | 19 | MI | No CHF | 4 |
| 20 | Yes | Yes | 6 | 20 | MI | CHF | 5 |
|  |  |  |  | 21 | MI | CHF | 5 |
|  |  |  |  | 22 | MI | CHF | 5 |
|  |  |  |  | 23 | MI | No CHF | 4 |
|  |  |  |  | 24 | No MI | CHF | 4 |

**Table 4: Patients of Clinician and Peer Providers May Differ in Significant Ways**

An event tree can organize and summarize the data. An event tree is a decision tree without a decision node. Each feature of the patient (e.g. previous Myocardial Infarction, MI, or Congestive Heart Failure, CHF) can be used to create a new branch in the event tree. The tree then ends with the consequences (length of stay) presented to the right of each branch. A branch on the tree shows a particular combination of patient features. When data are in table format these combinations of features or branches are called strata. The idea of apple-to-apple comparison is to make sure that both the clinician and the peer group are matched on the same branches or same strata. For example, the event tree for the patients seen by the clinician and peer group are provided in Figure 1. In this tree, previous MI is the first event, then CHF is the second event. The Length of Stay is given to the right of the tree. The probability of previous MI and the conditional probability of CHF given a previous MI is given on the arcs. To make it easier to read the tree and because within each node the probabilities add up to 1, the probabilities of negative events, such as not having CHF, are not provided.

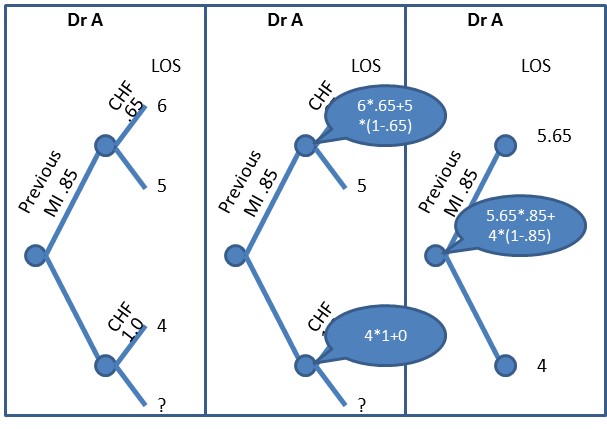


**Figure 1: Decision Trees for the Clinician's & Peer Group Practices**

CHF=Congestive Heart Failure, MI=Myocardial Infarction, LOS=Length of Stay

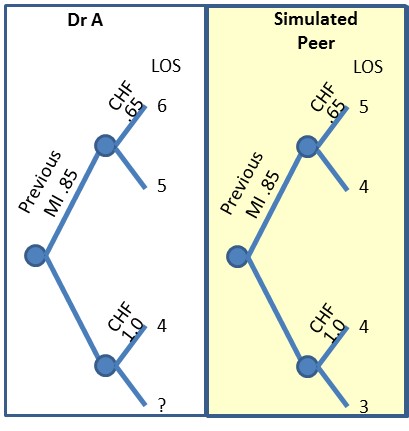
The expected length of stay for the patients of the clinician being evaluated was: 5.4 days. This is obtained by folding back the tree to the root node as shown in Figure 2. Starting from the right upper corner of the tree, we first fold the event of occurrence of CHF. This is calculated as: 6 days \* 0.65 + 5 days \* (1 - 0.65). Next we do the same for CHF in the lower right side: (4 days \* 1.0 + 0). The process of folding back continues by replacing the uncertain event with the calculated expected values. Each node is replaced with the expected length of stay. This is done until we reach to the root of the tree. The expected value for the entire tree is calculated as:

Expected length of stay = (6 \* 0.65 + 5 \* (1 - 0.65)) \* (0.85) + (4 \* 1.0 + 0) \* (1-0.85) = 5.40

****

**Figure 2: Folding Back in 3 Steps**

To simulate how the same patients would have been cared for under care of peer providers, the event tree is kept as before, but now the average length of stay of each patient grouping is replaced with the average length of stay of patients of the peer providers. Figure 3 provides the resulting trees.



**Figure 3: Simulated Performance of Peer Group on Patients’ of the Clinician**Outcomes from peer group and probabilities from Dr. A’s patients

Note that the tree in Figure 3 is a hybrid. Part of it comes from the clinician’s frequency of seeing different types of patients. The other part, the outcomes, comes from the peer group. The expected length of stay of the patients' of the clinician was calculated to be 5.40 days. If the peer group sees the patients of Dr. A, then the simulated expected length of stay is 4.55.

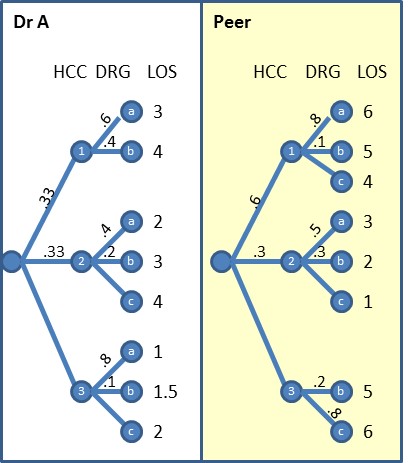
(5\*.65+4\*(1-.65))\*.85 + (4\*1+3\*(1-1))\*(1-.85) = 4.55

Thus, Dr. A’s patients stay on average 0.85 days longer than his peer group treating the same group of patients. In 20 patients, Dr. A’s patients stay 17 days longer than if they were treated by Dr. A’s peer. Note that if we had not switched the probabilities, we could not claim that peer’s and clinician’s performance was calculated on same patients. This example highlights the importance of simulating the performance of peer providers on the patients seen by the clinician.

**The problem with Overlap of Cases among Clinicians and Their Peers**

As the number of features increases, the number of data points that fall within each branch on the decision tree becomes smaller. Soon most branches will have no patients. Many peer providers' cases cannot be matched feature by feature to the clinician's patients. When the features available do not match, the analyst can used expected outcomes to replace missing information.

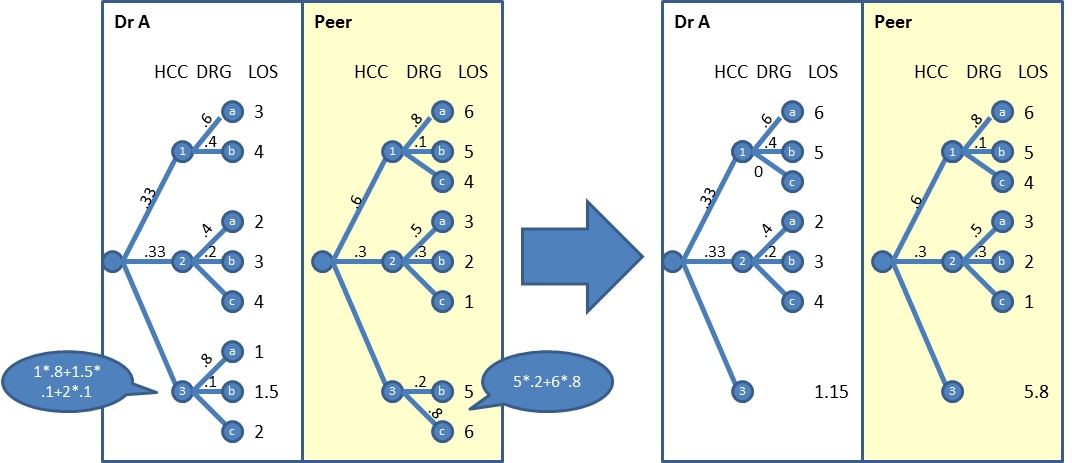
For example consider the following data on Dr. A and his peer doctors. For each patient managed by these doctors we have information on the Centers for Medicare and Medicaid Service's (CMS) Hierarchical Condition Category (HCC) risk adjustment model, and the assigned Diagnostic Related Groups (DRGs). For simplicity, we have divided the HCC scores into three groups: low, medium and high shown as 1, 2, and 3. Also for simplicity, we have assumed there were only 3 DRGs shown as a, b, and c. Since HCC is always available, we will start the event tree using this variable. DRGs are not always occurring for all physicians so these values may be null. The tree shows the probability of various HCCs and DRGs. Note that the probabilities shown for the DRGs are conditional on HCC values. So the product of probability of HCC probability and conditional probability of DRG shows the joint probability of the combination of HCC and DRG. Also further suppose that Dr. A and his peer have the length of Stay (LOS) indicated in Figure 4. A glimpse at the tree suggests that Dr. A’s peer may be seeing patients with higher severity and thus longer stays. The tree for Dr. A and the tree for his peer group are structurally different. There are some branches in Dr. A’s tree that are not in the peer group and vice versa. Given these differences, it is not possible to switch the probability events from one tree with another.



**Figure 4: Probability and Outcomes for Dr. A and His Peer Group**

Dr. A does not see any patients who have DRG c and HCC 1. Likewise, the peer group does not see patients with HCC 3 and DRG a. In general, in all comparisons of clinicians to benchmark groups, some strata are missing for one or the two groups. In the literature the extent of match in branches of the tree is referred to as “tree overlap.” It is rare to have a perfect overlap. Procedures are needed on how to manage the data when the two trees have partial overlap. Of course, portion of the data where the two trees do not overlap can be ignored but doing so will throw away a lot of data including crucial data about clinicians who see rare but severely ill patients.

The folding back method provides a way of adjusting the two trees so that the branches in both trees are exactly the same. If the peer group branch is too long then folding back can be used to trim it. If Dr. A’s branch is too long, then folding back can be used to trim it. One can fold-back either the right or the left tree till one finds a situation where the two trees match exactly. Where the structure of the two trees differ, we can use the folding back and expected values to replace the nodes with missing branches. In Figure 5, we have trimmed the two trees so that they have same branch structures. The trees for the clinician and the peer group have two areas where the branches do not match. Sometimes the peer group sees patients that the clinician does not see. Since we are comparing the clinician and the peer group on patients of the clinician, this miss-match does not matter. We can imagine that the clinician sees the miss-matched patients with probability of zero. Then the two trees have the same structure on the right side top part of Figure 5. The reverse is not true. When data are missing in the peer group, then assuming an arc with probability of zero is not reasonable. In these situations, folding back can help us arrive at a situation where the two tree’s structures are the same (see the right side bottom part of Figure 5). Now that both trees have the same branches, we are ready to switch the probabilities.



**Figure 5: Trimming the Two Tress to Common Branch Structure**

If the peer group sees the patients of Dr. A, and the two tree structures are the same, then we can simulate the effect of the peer group seeing patients of Dr. A by switching the probabilities in the two trees and calculating the expected value for the clinician and the peer group. The expected value for Dr. A is calculated with the distribution of his/her own patients. In the following we have put the probabilities in bold to assisted tracing where they come from:

**0.33**\*(**0.6**\*3 +**0.4**\*4) +**0.33**\*(**0.4**\*2 +**0.2**\*3 **+ 0.1**\*4) + **0.34**\*1.15 = 2.50 days

The expected value for the peer group seeing the patients of Dr. A is calculated by using the distribution of the patients of DR. A but the outcomes of the peer group. Again, probabilities are shown in bold to highlight that these are the same as in above equation:

**0.33**\*(**0.6**\*6 +**0.4**\*5+**0.0**\*4) +**0.33**\*(**0.4**\*3 +**0.2**\*2 **+ 0.4**\*1) + **0.34**\*5.8 = 4.48 days

Based on these calculations, we conclude that on average, and on the same set of patients, Dr. A is 4.48 - 2.50 = 1.98 days, or 21%, more efficient than his/her peers.

**Analysis Using Excel**

We can use Excel to do the necessary calculations. First, cases are classified into categories using Pivot table. In particular, rows of the table include the combination of the strata; the columns include Dr. A and Peer Group cases. The values inside the Pivot table need to calculate the probability of observing a patient in each cell and the average length of stay (LOS). This pivot table reconstructs the tree structure in a tabular format. Table 5 shows the tabular presentation of the two trees. Note that where the branches are missing, the tabular values are also missing.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Strata** | **Dr. A** | | | **Peer Group** | | |
| **HCC probability** | **DRG probability** | **LOS** | **HCC probability** | **DRG probability** | **LOS** |
| HCC1, DRG a | 0.33 | 0.6 | 3 | 0.6 | 0.8 | 6 |
| HCC1, DRG b | 0.33 | 0.4 | 4 | 0.6 | 0.1 | 5 |
| HCC1, DRG c | 0.33 |  |  | 0.6 | 0.1 | 4 |
| HCC2, DRG a | 0.33 | 0.4 | 2 | 0.3 | 0.5 | 3 |
| HCC2, DRG b | 0.33 | 0.2 | 3 | 0.3 | 0.3 | 2 |
| HCC2, DRG c | 0.33 | 0.4 | 4 | 0.3 | 0.2 | 1 |
| HCC3, DRG a | 0.34 | 0.8 | 1 | 0.1 |  |  |
| HCC3, DRG b | 0.34 | 0.1 | 1.5 | 0.1 | 0.2 | 5 |
| HCC3, DRG c | 0.34 | 0.1 | 2 | 0.1 | 0.8 | 6 |

**Table 5: Experience of Dr. A and Peer Group**

The second step is to fold back in situations where length of stay of the peer group is not available, i.e. the DRG or HCC variables for the peer group is null. In this situation, we calculate the expected value using the Sum Product function for portion of the strata that is available (in this case HCC). For example, in the third to the last row in Table 5, the LOS for the peer group is missing. A new row of data is calculated to replace all three last rows in Table 5, situations where HCC is 3. The LOS value in the new row is calculated as:

=sumproduct(three first rows of DRG p for Dr. A, three first rows of LOS for Dr. A)= .6\*3+.4\*4 = 3.40

=sumproduct(three first rows of DRG p for Peer Group, three first rows of LOS for Peer Group) = .8\*6+.1\*5+.1\*4 = 5.70

The net results of these alterations are shown in Table 6:

|  |  |  |  |
| --- | --- | --- | --- |
| **Strata** | **Probability for Dr. A Patients** | **LOS for Dr. A** | **LOS for Peer Clinicians** |
| HCC1, DRG a | 0.198 | 3 | 6 |
| HCC1, DRG b | 0.132 | 4 | 5 |
| HCC1, DRG c | 0 | NULL | 4 |
| HCC2, DRG a | 0.132 | 2 | 3 |
| HCC2, DRG b | 0.066 | 3 | 2 |
| HCC2, DRG c | 0.132 | 4 | 1 |
| HCC3 | 0.34 | 1.15 | 5.8 |
|  | **Expected** | **2.50** | **4.48** |

**Table 6: Revised Table with No Missing Values**

The third and final step is to calculate the observed expected length of stay for Dr. A and the simulated expected length of stay for the peer group. This again is done through Sum Product functions:

Expected Value for Dr. A = sumproduct(rows of HCC p for Dr. A, rows of DRG p for Dr. A, rows of LOS for Dr. A) = .33\*1\*3.40 + .33\*0.4\*2 + 0.33\*0.2\*3 + 0.33\*0.4\*4 + .34\*1\*1.15 = 2.50

Expected Value for Peer Group=sumproduct( rows of probability of HCC for Dr. A, rows of probability of DRG for Dr. A, rows of LOS for Peer Group) = . 0.33\*(0.6\*6 +0.4\*5+0\*4) +0.33\*(0.4\*3 +0.2\*2 + 0.4\*1) + 0.34\*5.8 = 4.48

A way to think about this procedure is that we have simulated the performance of peer providers on the patients’ of the clinician by replacing the probabilities with the corresponding values from the clinician's experience -- whenever such probabilities are available. When such values are not available, we replace the missing values with locally estimated expected values.

**SQL Code**

The following provides the SQL code for accomplishing the goals of the benchmarking effort. In the first two steps, we group the data by HCC and DRG. A particular combination of HCC and DRG constitute a stratum or a branch on the tree.

/\*\*\*\*\*\* Code to Match using Expected Values & Folding Back \*\*\*\*\*\*/

USE Benchmarking

-- Create tables of strata for the clinician and benchmark group.

-- In massive data, these steps make the data more manageable.

DROP TABLE #DrA

DECLARE @total as float

SET @total = (SELECT COUNT([ID])FROM #data WHERE Dr=1)

SELECT [DRG] as DRGa

,[HCC] as HCCa

,Avg(CAST([LOS] as Float)) as LOSa

,COUNT([ID]) as NumA

,COUNT([ID])/@total as ProbA

INTO #DrA

FROM #data

WHERE Dr=1 -- Select Dr you wish to examine

GROUP BY [DRG], [HCC]

DROP TABLE #DrB

DECLARE @totalb as float

SET @totalb = (SELECT COUNT([ID])FROM #data WHERE Dr=2)

SELECT [DRG] as DRGb

,[HCC] as HCCb

,Avg([LOS]) as LOSb

,COUNT([ID]) as Numb

,COUNT([ID])/@totalb as ProbB

INTO #DrB

FROM #data

WHERE Dr=2

GROUP BY [DRG], [HCC]

-- \*\*\* Match clinicians and peer group on common strata/branches \*\*\*

-- This step produces the data in Table 5

DROP TABLE #Match1

SELECT CASE When HCCa IS null Then HCCb Else HCCa END as HCCa

, CASE When DRGa IS null Then DRGb Else DRGa END DRGa

, LOSa

, CASE When NUMa IS null Then 0 Else NUMa END NUMa

, CASE When ProbA IS null Then 0 Else ProbA END AS ProbA

, CASE When HCCb IS null Then HCCa Else HCCb END as HCCb

, CASE When DRGb IS null Then DRGa Else DRGb END AS DRGb

, LOSb

, CASE When NUMb IS null Then 0 Else NUMb END NUMb

, CASE When ProbB IS null Then 0 Else ProbB END AS ProbB

INTO #Match1

FROM #DrA a Full Join #DrB b on DRGa=DRGb and HCCa = HCCb

Select \* from #match1 order by HCCa, DRGa

HCCa DRGa LOSa ProbA HCCb DRGb LOSb ProbB

1 a 3 0.198 1 a 6 0.480

1 b 4 0.132 1 b 5 0.060

1 c NULL 0.000 1 c 4 0.060

2 a 2 0.132 2 a 3 0.150

2 b 3 0.066 2 b 2 0.090

2 c 4 0.132 2 c 1 0.060

3 a 1 0.272 3 a NULL 0.000

3 b 1.5 0.034 3 b 5 0.020

3 c 2 0.034 3 c 6 0.080

\*/

-- Decide where to fold

-- Assign 0 to situations where we should fold back

DROP Table #Match2

SELECT HCCa as HCC

, Min(CASE When LOSb IS Null Then 0 Else 1 END)as DontFold

INTO #Match2 FROM #Match1

GROUP BY HCCa

SELECT \* FROM #Match2

/\*

HCC DontFold

1 0

2 1

3 1

\*/

-- At folded and unfolded categories, calculate probabilities and LOS

DROP TABLE #Match3

SELECT HCCa as HCC

, IIF(DontFold=0, ’O’, DRGa) as DRG

, SUM(ProbA) As ProbA

, SUM(ProbA\*LOSa)/SUM(ProbA) as LOSa

, SUM(ProbB\*LOSb)/Sum(ProbB) as LOSb

INTO #Match3 FROM #Match1 a Left Join #Match2 b on a.HCCa=b.HCC

Group by HCCa, IIF(DontFold=0, ’O’, DRGa)

SELECT \* FROM #Match3 order by HCC, DRG

/\*

HCC DRG ProbA LOSa LOSb

1 a 0.198 3 6

1 b 0.132 4 5

1 c 0 NULL 4

2 a 0.132 2 3

2 b 0.066 3 2

2 c 0.132 4 1

3 F 0.34 1.15 5.8

\*/

-- Expected values using distribution of Dr. A’s patients

SELECT sum(ProbA\*LOSa) AS DrA

, SUM(ProbA\*LOSb) AS Benchmarks

FROM #Match3

/\*

DrA Benchmarks

2.503 4.48

\*/

The results indicate major difference between the clinician and peer group. If we simulate how the peer clinicians might see patients of Dr. A, the expected outcome is 4.48 days. Dr. A takes 2.50 days. Dr. A is more efficient by an average of 1.98 days, or 21%.

**Is it Reasonable to Benchmark Clinicians?**

Risk assessment is not as benign as it first looks. Risk assessment and practice profiling has its own risks [[[6]](#endnote-6),[[7]](#endnote-7)]. When clinicians' performance is measured and the clinician is provided with feedback, several unintended consequences may occur:

1. Measurement may distort goals. Clinicians may improve their performance on one dimension but inadvertently deteriorate on another. For example, if the manager emphasizes length of stay, clinicians may improve on this measure but inadvertently increase the probability of re-hospitalization because they have sent the patient home too early. Everyone focuses on what is measured and may ignore other issues. To avoid this shortcoming, it is important to select the benchmarking goals broadly and to select multiple benchmarks.
2. Measurement may lead to defensive behavior. Clinicians may put their effort or time in defending their existing practices as opposed to improving them. To avoid this pitfall, it is important to engage the clinicians in the selection of the performance indicators and the severity index. Managers can ask clinicians what they want to be evaluated on and how they wish to measure the severity of their patients. Furthermore, it is important to make sure that feedback to each clinician is provided privately and without revealing the performance of any other single peer provider. It is ok to share the average of the peer providers as long as the identity of each provider remains anonymous. The focus of feedback should be everyone not just the clinicians with poor performance. An environment needs to be created where no one is blamed and all clinicians are encouraged to seek improvements as opposed to arguing about the results.
3. Inadequate measure of severity may mislead the analysis. A poor severity index, one that is not predictive of the patients’ prognoses, might give the impression of severity adjustment but in reality be no better of a random guess of outcomes. In these circumstances, an unadjusted benchmark is better as at least it does not give the appearance of what it is not. To avoid this pitfall, it is important to select a severity index that has high predictive power.
4. Too much measurement, may lead to too little improvement. Sometimes analysts who conduct benchmark studies take considerable time to collect information and analyze it. In these circumstances there may be too little time spent on discussing the results, selecting a new course of action and following up to make sure that the change is an improvement. It is important to keep in mind that the goal of benchmarking is improvement. Conducting an accurate analysis is only helpful if it leads to change and improvement, otherwise it is a waste of time.

**Presentation of Benchmarked Data**

Presenting benchmarked data should be done in a fashion that helps clinicians improve their practices as opposed to acting defensively. Poorly presented information leads to unnecessary and never ending debates about the accuracy of the information presented. The Agency for Healthcare Research and Quality, Centers for Medicare & Medicaid Services and the Office of Personnel Management sponsored a working group on how to talk about health care quality. The group made a number of suggestions on the presentation of benchmarked information. The following is a summary of the group's suggestion:

Before the Meeting

* Check the data. A simple mistake in benchmarked data will undermine the perception of the validity of the entire data set. To avoid this, check the accuracy of the data thoroughly prior to the meeting, making sure that all variables are within range and that missing values are appropriately handled. Prepare histograms of each individual variable in the analysis and review to make sure they seem reasonable.
* Prepare graphs and pictures. In order to help clinicians have an intuitive understanding of statistics, it is important to provide them with visual displays of data. Show data and summary statistics using bar-charts and x-y plots.
* Distribute paper reports. Prepare handouts for discussion during the session. Distribute handouts ahead of the meeting to participants. Make sure that handouts are stamped draft and that the date of final report is clearly reported.
* Prepare stories. Supplement numeric data with anecdotal information that have the same message. Make sure that the anecdotes do not reflect judgments about quality of care but focus on the data being reported (i.e. patient's condition or patient outcomes). Provide an example of the patients' typical complaint (usually in the form of an audio or video short tape). It is important to weave the story or the anecdotal data in with the voice of the customer.

At the Meeting

* Confidential evaluation. Make it clear that the evaluation is confidential. If talking to a group of clinicians, do not identify who is the best or worst. You may let each clinician privately know how they performed against the group but do not publicly provide this information.
* Brief introduction. Make a brief introduction of the purpose of the session. Introduce your project team and ask clinicians to introduce themselves. Even if they know each other, still ask them to introduce themselves so they feel better.
* Limitations of benchmarking. Acknowledge the limitation of the practice profiling method. Explicitly say that numbers could be misleading if the measures of severity are not adequate or the sample size is small. Point out that the focus should be on improvement and not measurement issues.
* Start with a customer's story in the customer's voice. Start the meeting by playing a brief tape of a customer talking in his/her own words about what happened to him/her. Use both positive and negative vignettes.
* Present the data and not the conclusions. Present the findings without elaboration about causes or explanations. Do not give advice about how clinicians can do things differently. For example, say "Data show that patients stay longer in our hospital." Don't say "You should shorten the time it takes for hip fracture patients to get discharged." Its up to the clinicians to change and decide how to change, benchmarking just points out the variation in outcomes and facilitates the clinicians to focus on specific issues. What clinicians do depends on them. During the presentation, the analyst guides the clinicians through the data. The data and not the analyst help clinicians arrive at a conclusion and to act.
* Ask for input. Explicitly ask for the audience's evaluation of the data after each section of the report is presented. Allow the clinicians to talk about the data by pausing and staying quiet. For example, "Data show large variation on how long it takes us to discharge a patient with hip fracture, what do you think about that?" Pause and let them talk about the variation in hip fracture data. The point is not to troubleshoot and come up with solutions on the spot but to discuss the issues and think more about causes.
* Accept criticism. Do not defend the practice profiling method, the benchmarking effort or any aspect of your work. Let your work speak for itself and accept their suggestions for future improvements. Shift the discussion from blaming the study to what can be done to improve in the future.
* Plan for next step. Thank the clinicians for their time and describe next steps (e.g. I will correct the report and get it back to you within a week.)

After the Meeting

* Revise the report. Summarize the comments made during the meeting and append it to the report.
* Summarize available resources. Describe resources available through the management to help them (e.g. travel funds to attend meetings, invited presentation of best practices).
* Distribute report. Send a written report to each clinician. Please note that reports have a way of lasting well beyond the time for which they were generated. Make sure that all providers' identities are removed.
* Ask for feedback. Ask the clinicians to comment on:

1. What worked well regarding the practice profiles and what needs improvement?
2. Do clinicians plan to change their practice and in what way?
3. Was it worthwhile to gather data and do benchmarking? Why?
4. Announce the next round of benchmarking. Set the time of next benchmarking report.

**References**

1. Andel C, Davidow SL, Hollander M, Moreno DA . The economics of health care quality and medical errors. J Health Care Finance. 2012 Fall;39(1):39-50. [↑](#endnote-ref-1)
2. Changes in Consumer Demand Following Public Reporting of Summary Quality Ratings: An Evaluation in Nursing Homes. Werner RM, Konetzka RT, Polsky D. Health Serv Res. 2016 Jun;51 Suppl 2:1291-309. [↑](#endnote-ref-2)
3. Callahan M, Fein O, Battleman D. A practice-profiling system for residents. Acad Med. 2002 Jan;77(1):34-9. [↑](#endnote-ref-3)
4. Cargill V, Cohen D, Kroenke K, Neuhauser D. Ongoing patient randomization: an innovation in medical care research. Health Serv Res. 1986 Dec;21(5):663-78. [↑](#endnote-ref-4)
5. Alemi F, ElRafey A, Avramovic I. Covariate Balancing through Naturally Occurring Strata. Health Serv Res. 2016 Dec 14. [↑](#endnote-ref-5)
6. Iezzoni LI. The risks of risk adjustment. JAMA. 1997 Nov 19;278(19):1600-7. [↑](#endnote-ref-6)
7. Krumholz MH, Rathore SS, Chen J, Wang Y, Radford MJ. Evaluation of a Consumer-Oriented Internet Health Care Report CardThe Risk of Quality Ratings Based on Mortality Data. JAMA. 2002; 287(10):1277–1287. [↑](#endnote-ref-7)