Exploring Variability when Interpreting Performance Rates

Conformance with guidelines is measured by constructing clinician profiles (i.e. clinicians' rates of performance on specific tasks). A performance rate, however, can best be interpreted through comparison with an external standard. A fair comparison between clinicians or clinician groups requires allowing for differences due to patient case mix, as described later in this section. This adjustment of the estimate should also be supplemented with confidence limits, allowing for variability due to both sampling error and clinician-related differences, such as differences in the patients served by different clinicians. (To avoid confusion with the concept of standards of quality discussed in relation to performance measures, the terms adjusted and adjustment are used throughout this paper in place of the epidemiological terms standardized and standardization.)

For example, assume a health maintenance organization (HMO) is being characterized according to the number of acute asthmatic attacks with treatment not conforming to guideline in 1 year. \( N \) represents the number of patients seen for an asthmatic attack during the year; \( C \) represents the number of patients who had non-conforming treatment. Then a simple summary profile is:

\[
\text{Rate of non-conforming treatment} = \frac{C}{N}.
\]

It is not necessarily useful to compare such a simple summary rate to a standard of performance or to compare it between HMOs, because different HMOs may see different types of patients, for whom receiving the recommended care is more or less critical. For instance, a committee may be less likely to recommend expending resources on quality improvement if non-conformance to treatment guidelines is restricted to mild asthmatic attacks that might well resolve spontaneously. Before making any comparisons, the committee therefore adjusts for severity of the asthmatic attacks. Patients are divided into distinct categories, and HMOs are expected to perform similarly when treating patients within such a category. The case-mix adjustment is made by using either of the following:

- **Direct adjustment**, in which the HMO profile is reconstructed, adjusting the rate to reflect a standardized case mix.
- **Indirect adjustment**, using standardized rates within each category of patients and calculating the expected rate for the case mix observed for a given HMO. The observed rate is then divided by the expected rate.

While either method could be used in principle, direct adjustment may not be possible if there is only a small number of patients for each clinician. In this case, the clinician performance rate cannot be calculated if no patients for a particular category were reviewed. When small sample sizes are common, it is more practical to pursue indirect adjustment.

In an example of indirect adjustment, patients with acute asthmatic attacks are categorized by the "severity" of the attack (severity determined by peak flow rate): category 1, mild asthma; category 2, moderate asthma; and category 3, severe asthma.

Then,

\[
N_1 = \text{the number of patients in category 1, mild asthma;}
\]
\[
N_2 = \text{the number of patients in category 2, moderate asthma;}
\]
\( N_3 \) = the number of patients in category 3, severe asthma;
where \( N_1 + N_2 + N_3 = N \), the total number of patients in the HMO, and

\[ R_1 = \text{proportion of patients with mild asthma whose care does not conform to guideline, derived from some external standard;} \]
\[ R_2 = \text{proportion of patients with moderate asthma whose care does not conform to guideline, derived from some external standard;} \]
\[ R_3 = \text{proportion of patients with severe asthma whose care does not conform to guideline, derived from some external standard.} \]

The expected number of non-conforming treatments for the observed HMO is

\[ E = (N_1 \times R_1) + (N_2 \times R_2) + (N_3 \times R_3). \]

This forms the indirectly adjusted rate:

\[ \text{Observed rate/expected rate} = \frac{C}{N} / \frac{E}{N} = \frac{C}{E} \]

This indirectly adjusted rate has a simple interpretation. If the rate equals 1, then the HMO is performing as would be expected according to the standard. On the other hand, if the adjusted rate is greater than 1, then the HMO’s profile is higher than would be expected, and if the adjusted rate is less than 1, then the HMO’s profile is lower than would be expected.

While the adjusted rate is accurate and interpretable as a summary of HMO performance, adjusted for patient case mix, the indirectly adjusted rate remains only an estimate based on a sample of the HMO’s patients. The more patients in the sample, the more likely the estimated rate is close to the true rate for the HMO. However, with a small sample, the estimate can be far from the truth. In providing an HMO with a report of its performance, providing upper and lower limits between which the truth is likely to be included (as well as the estimated rate) would be important. To provide such limits, a confidence interval is created around the estimated rate.

A confidence interval would also have a natural and familiar interpretation. If the confidence limits contain 1, then the HMO could be performing at the standard even though the estimated rate may be different from 1. On the other hand, if the confidence limits do not contain 1, then the HMO is probably performing at a rate above or below the standard; whether above or below is indicated by the estimated rate. The confidence limits are constructed according to the methods in Breslow and Day (1987) [1].

It is assumed that the total number of non-conforming treatments in a year has a Poisson distribution (a commonly used distribution describing rare events). Then, a 95% confidence interval for the indirectly standardized rate is given by the lower and upper bounds, in which

\[ \text{Lower} = \left[ \frac{C}{E} \right] \times \left[ 1 - \frac{1}{9C} - \left( \frac{1.96}{3\sqrt{C}} \right) \right] \]
\[ \text{and} \]
\[ \text{Upper} = \left[ \frac{C + 1}{E} \right] \times \left[ 1 - \frac{1}{9(C + 1)} + \left( \frac{1.96}{3(C + 1)^{1/2}} \right) \right]. \]

Carrying through the previous example, it is assumed that the adjusted ratio for non-conforming treatments, \( C/E \) (\( C \), the number of non-conforming treatments, over \( E \), the expected number of non-conforming treatments), equals 1.25. Consider HMO A, for example, for which the actual number of non-conforming treatments is 60 and the expected number of non-conforming treatments is 48, for a 95% confidence interval between 0.95 and 1.60. With the low sample size of 60 cases, there is not enough evidence to reject the possibility that the HMO is performing within the standard, even though its non-conforming treatment rate is higher than normal (1.25). Now consider HMO B, whose number of non-conforming treatments is 125 and expected number of non-conforming treatments is 100, which places the 95% confidence interval between 0.95 and 1.60. With the larger sample of 125 cases, there is enough evidence to state that HMO B’s non-conforming treatment rate, although it is the same rate as HMO A’s (1.25), is significantly higher than normal.

One possible problem may remain. The confidence limits that were calculated allow only for sampling variability. Would it be expected or desirable for all HMOs to have performance rates that are exactly the same? There may be a number of reasons why this might be an unreasonable expectation for any particular HMO:

- The case-mix adjustment may not be perfect.
• There may be regional variations related to varying prevalence of the guideline-related disease or relevant comorbidities.
• There may be variations in access to facilities or third-party coverage that affect patient willingness to cooperate in treatment.
• There may be a variability in the application of the review criteria.

There might be a certain amount of natural, even expected, variability among the "true" performance rates of HMOs. The importance of co-factors that might account for this variability can be explored, for instance, by expanding the confidence limits for a particular HMO to allow for this natural variability as well as for the sampling variability considered previously.

To make this allowance, the following hierarchical model has been developed for use in comparing performance among clinicians in the DEMPAX project [2]. It is still assumed that the number of non-conforming events for a given clinician follows a Poisson distribution. It is also assumed that the mean of the Poisson distribution is specific to each clinician. Then, among different clinicians, the Poisson means are themselves random variables drawn from an exponential distribution.

Under these assumptions, a standard calculation for mixture distributions shows that, for the clinician for whom a profile is being created, the number of non-conforming events follows a geometric distribution. The mean of the geometric distribution depends on the amount of variability allowed between clinicians.

The amount of additional variability introduced through these assumptions is illustrated as follows:
• If only sampling variability is assumed, then the number of events of guideline-related interest is drawn from a Poisson distribution and will have a variance best estimated as C.
• If the hierarchical model is assumed, allowing for variability between clinicians as well as sampling variability, then C will have a variance best estimated as C + C^2.

This additional variability becomes incorporated into wider confidence intervals. Assume that the C non-conforming treatments occurred as follows: C1 among patients with mild asthma attacks, C2 among patients with moderate asthma attacks, and C3 among patients with severe attacks, and C1 + C2 + C3 = C. Then, under the hierarchical model, the variance of C equals C + C1^2 + C2^2 + C3^2. This is in contrast to the simpler assumption of only sampling variability, which states that the variance of C equals C.

The variance from the hierarchical model is then used to construct the 95% confidence intervals for the indirectly adjusted rate, using an approximation in which the lower and upper limits are as follows:

Lower = (C/E) - [1.96*(variance(C)^1/2)/E]
Upper = (C/E) + [1.96*(variance(C)^1/2)/E]

As an illustration of the impact of the additional variability, it is assumed, as before, that C equals 60 and E equals 48, with C1 equaling 30, C2 equaling 20, and C3 equaling 10. Then, the 95% confidence interval, with only sampling variability, is [0.93, 1.57]. In contrast, the 95% confidence interval that allows for clinician-to-clinician variability as well as sampling variability is [0.45, 2.05]. It is recommended that these hierarchical intervals be developed further and considered for use in reporting guideline conformance.

SUMMARY AND FUTURE CONSIDERATIONS

The accuracy of the confidence intervals described in the previous section is directly tied to the sampling scheme and sample size employed. As seen in the examples, large numbers of patients may be required to construct tight bounds that allow accurate discrimination between "normal" and "abnormal" rates. That is, if a threshold standard were set for further investigation or intervention, large numbers of patients may be needed before we are confident that we can distinguish acceptable from unacceptable rates. If stratified sampling is done, with an intent to analyse rates separately in each stratum, then these large sample sizes must be available in each stratum. As a particular, common, example, stratifying by clinician in order to evaluate the performance of each clinician separately may require a very large review effort to ensure meaningful comparisons. Often, attempts to evaluate individual clinicians require such large sample sizes that the studies either become unworkable or are carried out...
with insufficient power to lead to useful conclusions. It is generally more practical to evaluate clinicians as a group, such as the staff of an HMO or hospital, where large-scale sampling is practical, using stratification to ensure adequate representation in terms of patient case mix.

Although this statistical discussion has been restricted to the construction of confidence intervals, the same issues and remarks apply to decision-making and the implementation of intervention programs. The same small sample sizes that create wide confidence bounds and prevent a decision about whether a particular clinician is truly performing according to standards also create power problems when a decision is needed to implement an intervention program. A natural analog to developing confidence intervals is the designing of a record sampling study in which a fixed number of records is reviewed to determine whether a more intensive investigation or an immediate intervention program is necessary. This sort of "lot sampling design" is already a common industrial tool [3] that could be modified to account for repeated samples from the same clinicians (i.e. the same clinician-to-clinician variability issue addressed in the construction of confidence intervals). However, the chance of making an incorrect decision and either abstaining from a needed intervention or implementing an unnecessary one is tied directly to sample size. Hence, any attempt to evaluate individual clinicians requires very large samples. Again, the best recommendation would be to focus on groups of clinicians when global interventions such as education or system improvements can have a powerful impact.


REFERENCES

1. Breslow N E and Day N E, 


John Orav, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA