The Relationship of the Value of Outcome Comparisons to the Number of Patients Per Provider

ARTHUR J. HARTZ,* EVELYN M. KUHN* and HENRY KRAKAUER†

*Medical College of Wisconsin, Department of Family & Community Medicine, 1000 N. 92nd Street, Milwaukee, WI 53226, USA; †Department of Preventive Medicine, Uniformed Services University, Bethesda, MD 20814, USA

Purpose: Monte Carlo methods were used to assess how the value of outcome comparisons depends on the number of patients per provider.

Methods: We simulated two patient data sets that have been used for well-known studies of outcome comparisons: mortality rates for coronary artery bypass surgeons from New York and Pennsylvania, and 30-day hospital mortality rates of Medicare patients from a national data set. In the simulated data sets, each surgeon or hospital provider was assigned a true or underlying probability of mortality.

Results: For the simulated CABG surgery data set, the underlying probability of mortality explained 30% of the variation in the observed mortality rate when there were 100 patients per physician, and 63% when there were 400 patients. The positive predictive value of using an observed mortality rate in the bottom 10% to identify a surgeon whose underlying probability of mortality was in the bottom 10% was 31% for 100 patients and 59% for 400 patients. The relationship between underlying and observed rates was weaker in the simulated Medicare data set with the same number of patients per provider. For a given data set, the amount of random variation in the observed rates of adverse outcomes among providers can be estimated with a simple equation.

Conclusions: The results show that the assessment of provider outcomes may be greatly affected by random variation. An indication of the amount of random variation in a given data set can be obtained from the examples in this study and an equation for estimating random variation. © 1997 Elsevier Science Ltd.

Key words: Outcomes studies, risk-adjusted outcomes, sample size, random variation, Monte Carlo.

Outcome comparisons are often criticized, however. The most frequent criticism is that differences in outcomes among providers may be due to differences in unexplained patient risk rather than differences in quality of care [13–17]. A second common criticism is that the outcomes chosen may not reflect important aspects of care [18–22]. Another concern about outcomes comparisons that has been mentioned [18,19,23], but not thoroughly explored, is that they may lack precision because of random variation.

The precision of outcome assessment has been measured using confidence intervals. These intervals provide likely bounds for either the rate of an adverse outcome or the difference between the observed and expected rate of an adverse outcome. Usually, however, the assessment of the adverse outcome rate for a given provider depends not on the rate itself but how the rate compares to others. For example, patients may select providers with the lowest rates of adverse outcomes, or health care organizations may want to eliminate from the organization providers with the highest rates. Confidence intervals do not quantify the precision for making these types of decisions.

The factor that most influences the precision of outcome studies is the number of patients per provider. In the present study, we generated simulated data to examine how the number of patients per provider affects the relationship between a provider's true relative quality and the provider's apparent relative quality. The study does not develop precise criteria for a minimal number of patients per provider necessary to assess the relationship in any data set. It does, however, show how the number of patients per provider affects the relationship in some representative data sets, and it presents a method for estimating the random component of the variation in any real data set.

METHODS

A probabilistic model was used for the computer-generated data. In this model, the patients for each physician or hospital provider were assumed to have an underlying probability of an adverse outcome that depended only on the provider; the observed rate of

Received 1 April 1996; accepted 13 December 1996.
Correspondence to: Arthur J. Hartz, Medical College of Wisconsin, Department of Family & Community Medicine, 1000 N. 92nd Street, Milwaukee, WI 53226, USA.
adverse outcomes was a random realization of this underlying probability. The provider's true relative quality was measured by the provider's relative probability of an adverse outcome, and the provider's apparent or observed relative quality was measured by the observed relative rates of an adverse outcome. To simplify the model, patient risk was kept constant.

The study examined how the number of patients per provider affected the relationship between the provider's true and observed relative quality. The design of the study can be summarized in three steps:

1. As described in the next two sections, underlying provider probabilities of an adverse outcome were generated so that the mean and variance of these rates were likely to correspond to the mean and variance of underlying provider mortality rates in two well-known data sets used for outcome comparisons.

2. Observed rates of adverse outcomes were randomly generated from the underlying probabilities. In the generated data sets, the number of patients per provider varied from 100 to 1000, and 1000 samples were generated for a given number of patients per provider.

3. For a given number of patients per provider, we found the relationship between the relative values of the observed adverse outcome rates and the relative values of the underlying probabilities of adverse outcomes. As described below, several methods were used to characterize this relationship.

PATIENT DATA SETS

We designed the computer generated data to correspond to patient data sets that often have been analyzed in the literature and are representative of other data sets likely to be used for quality assessment. One patient data set combined New York and Pennsylvania data on mortality rates for coronary artery bypass surgeons [3,4]. A second data set simulated hospital mortality rates for Medicare patients [6].

Each of these data sets included the ratio of the observed to predicted rate of an adverse outcome for each provider. These ratios take into account patient risk. The risk-adjusted mortality rate for the ith provider \( MR_i \) was defined as \( MR_i = R_i \times MR_0 \) where \( MR_0 \) is the mortality rate for all patients combined, and \( R_i \) is the ratio of the observed to predicted rate of an adverse outcome for the ith provider. For each data set, we computed the mean and standard deviation of these adjusted rates.

GENERATING SIMULATED DATA SETS

A simulated data set consisted of a given number of providers, each with a given number of patients. All providers in a given data set had the same number of patients to make the results easier to interpret. Each provider had a fixed underlying probability of an adverse outcome, \( p_i \), that was used to generate an observed adverse outcome rate for that provider, \( \hat{p}_i \). One thousand \( \hat{p}_i \) values were generated for the ith provider with a given number of patients. The number of patients per provider was varied from 100 to 1000.

There were 100 providers in the simulated data set corresponding to the CABG surgery patient data and 300 providers in the simulated data set corresponding to the Medicare patient data. Using more than 300 providers did not affect the results. Only one set of \( p_i \) was used for the simulated CABG surgeon data. This set was created with the SAS RANBINO function [24] so that the mean and variance of the observed mortality rates in the actual CABG surgeon data set would equal the mean and variance of the simulated rates with the same average number of patients per provider. This same process was used to create the \( p_i \) for the Medicare data set.

The number of adverse outcomes for a provider with a given \( p_i \) and a given number of patients was generated with the SAS RANBINO function [24]. The adverse outcome rate for that provider, \( \hat{p}_i \), was computed as the number of adverse outcomes for that provider divided by the number of patients.

We only reported results for simulated data that were generated from one set of \( p_i \). However, three different

| TABLE 1. Example illustrating how the observed mortality rate varies for a given provider when the underlying mortality rate is fixed |
|-----------------|--------|--------|--------|--------|
| Prob of mortality, \( p_i \) | 1.69% | 3.25% | 4.12% | 5.01% |
| Subjects per provider | 200 | 200 | 200 | 200 |
| **Sample 1** | | | | |
| Number of deaths | 4 | 10 | 9 | 8 |
| Observed mortality rate, \( \hat{p}_i \) | 2.0% | 5.0% | 4.5% | 4.0% |
| **Sample 2** | | | | |
| Number of deaths | 2 | 6 | 9 | 6 |
| Observed mortality rate, \( \hat{p}_i \) | 1.0% | 3.0% | 4.5% | 3.0% |
| **Sample 3** | | | | |
| Number of deaths | 2 | 9 | 12 | 15 |
| Observed mortality rate, \( \hat{p}_i \) | 2.0% | 4.5% | 6.0% | 7.5% |
sets of \( p \)s from the same normal distribution gave very similar results.

**EXAMPLE OF SIMULATION**

An example of generating observed rates is shown in Table 1 for four providers. Each of the providers in the example was assumed to have 200 observations. The \( p_i \) for provider 1 in this example was 1.69%, i.e. each of the 200 observations for provider 1, the probability of an adverse outcome was 1.69%. In the first sample, provider 1 had four adverse outcomes, in the second, two adverse outcomes, and in the third, two adverse outcomes. The values of \( p \) and the number of adverse outcomes for three samples are also shown for providers 2, 3, and 4.

**MEASURING THE UTILITY OF OBSERVED OUTCOMES**

We used four measures of how well the provider's apparent relative quality reflected the provider's true relative quality. One was \( r^2 \), the square of the correlation between the observed rate of an adverse outcome and the underlying probability of an adverse outcome. The value of \( r^2 \) is the percentage of variation in the observed adverse outcome rate that can be accounted for by the underlying probability of an adverse outcome. If the providers have the same relationship with respect to their observed rates of adverse outcomes as they do with respect to their underlying probabilities of an adverse outcome, then the value of \( r^2 \) will be 1.00. If there is no relationship, the value of \( r^2 \) will be 0.00. This measure is the best indication of how well the relative observed rates reflected the relative quality of the providers.

A similar utility measure was the variation in one set of observed rates that was explained by another set of observed rates. This measure was used to determine how well performance at one time period may be used to predict performance at a later time period.

Other utility measures evaluated a criterion for selecting the best or worst group of providers. One criterion evaluated was that the provider had a rate of adverse outcomes in the lowest 10% of all physicians. We computed the positive predictive value of this criterion as a means for identifying best providers, those with an underlying probability of an adverse outcome in the lowest 10%, or good providers, those with an underlying probability of an adverse outcome rate in the lowest 25%. Since there were ties for observed adverse outcome rates, more than 10% of providers had an observed adverse outcome rate that was no worse than the lowest 10%. A similar criterion, an observed adverse outcome rate in the highest 10% of all providers, was evaluated as a method to identify the worst providers (underlying probabilities in the highest 10%) or bad providers (underlying probabilities in the highest 25%).

All measures of utility of a given outcome measure for a given sample size were evaluated in 1000 samples. We reported the average value of the utility measure in the 1000 samples.

**RESULTS**

In Table 2, the patient data sets and the simulated observed data sets are compared. Two patient data sets for mortality following CABG surgery are shown in the table, one for New York and one for Pennsylvania. The mean and standard deviation of the simulated observed rates were generated so that they would be between the values for the two states. Because the observed rates had a random variation component, the standard deviation was greater for the observed rates than for the underlying probabilities.

Adjustment made a greater difference in the standard deviation of the mortality rate for the Medicare mortality data than for the CABG surgery data. After adjustment, the standard deviation of adverse outcome rates in the Medicare data set was almost the same as in the CABG surgery data set. The average standard deviation of the simulated observed rates of the CABG surgeon data set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of providers</th>
<th>Average patients per provider</th>
<th>Mortality rate</th>
<th>Standard deviation of mortality rate among providers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Observed</td>
<td>Adjusted*</td>
</tr>
<tr>
<td>NY CABG surgeons</td>
<td>85</td>
<td>444</td>
<td>2.84</td>
<td>2.75</td>
</tr>
<tr>
<td>PA CABG surgeons</td>
<td>156</td>
<td>293</td>
<td>3.47</td>
<td>3.37</td>
</tr>
<tr>
<td>Underlying probability of mortality for simulated CABG surgeons†</td>
<td>100</td>
<td>346</td>
<td>3.05</td>
<td>3.05</td>
</tr>
<tr>
<td>Average of observed rates in 1000 simulated samples</td>
<td>100</td>
<td>346</td>
<td>3.05</td>
<td>3.05</td>
</tr>
<tr>
<td>HCFCA acute care hospitals</td>
<td>3782</td>
<td>1060</td>
<td>9.15</td>
<td>9.17</td>
</tr>
<tr>
<td>Underlying probability of mortality in simulated HCFCA data†</td>
<td>300</td>
<td>1060</td>
<td>9.17</td>
<td>9.17</td>
</tr>
<tr>
<td>Average observed rates in 1000 samples</td>
<td>300</td>
<td>1060</td>
<td>9.17</td>
<td>9.17</td>
</tr>
</tbody>
</table>

*The method for deriving risk-adjusted mortality rates is described in the section on patient data sets.

†Average of 1000 simulated data sets.
The relationship between the number of patients per CABG surgeon and the percent of variation in the observed rates that is explained by the underlying probabilities is shown in Fig. 1. The underlying probabilities explained 30% of the variation when there were 100 patients, 46% of the variation when there were 200 patients, and 63% of the variation when there were 400 patients. The variation that was not explained by the underlying probabilities is due to chance.

To interpret these results, it is necessary to know the number of surgeons who have a given number of patients. We obtained this information only from the Pennsylvania data because the New York data did not include surgeons who performed less than 200 cases in 3 years. In Pennsylvania 58% of the surgeons perform less than 100 surgeries a year, and 99% perform less than 400 surgeries a year. Over 3 years, 65% of the physicians performed 200 or more surgeries and 3.5% more than 600 surgeries.

Figure 1 also shows the percent of variation in adverse outcome rates for one simulation that can be explained by the adverse outcome rates in another simulation. It can be shown mathematically that, on average, this percentage is the square of the percentage of the variation in the observed rates that are predicted by the underlying rates. Even for providers who have as many as 200 patients, only 22% of the variation in another simulation can be predicted. These results suggest how well a physician's adverse outcome rate in one period may predict his or her outcome rate for the next time period. Clearly, the prediction is accurate only if there are a very large number of patients.

**Accuracy of Identifying Best Surgeons**

The criterion for identifying a best surgeon was an adverse outcome rate in the lowest 10th percentile. The chance that a surgeon who met this criterion would be a good or best surgeon, i.e., the positive predictive value of the criterion, is shown in Fig. 2. When there were 200 patients per surgeon, the criterion had a 45% chance of selecting a best surgeon, but an almost 70% chance of selecting at least a good surgeon. The sensitivity for identifying best surgeons is 62.5%, and the specificity is 91.6%. We included only the positive predictive values in a graph because patients are most concerned that they have correctly identified a best surgeon. At 600 patients per surgeon (a number achieved by only 3.5% of Pennsylvania surgeons in 3 years), the positive predictive value for identifying best surgeons is 66.4%, and the positive predictive value for identifying good surgeons is 88.9%. The sensitivity is 74.4%, and the specificity is 95.8%.

As shown in the figure, the positive predictive values were somewhat lower for identifying worst surgeons than best surgeons. The positive predictive values for identifying worst surgeons were 31.3% for 100 patients, 39.1% for 200 patients, and 63.9% for 1000 patients. The corresponding sensitivities and specificities were 42.4% and 89.7%, 48.7% and 91.5%, and 68.1% and 95.7%. The reason that the positive predictive values were lower...
The relationship of the value of outcome comparisons to the number of patients per provider for identifying worst surgeons than best surgeons is that worst surgeons have higher probabilities of an adverse outcome and therefore a greater binomial sampling variation than do best surgeons.

**COMPARISON OF $r^2$ FROM SIMULATED DATA AND FROM AN EQUATION**

The equation shown below to estimate $r^2$ was derived in the Appendix.

$$E(r^2) \approx \frac{\sum_{i=1}^{k} (p_i - \bar{p})^2}{\sum_{i=1}^{k} (p_i - \bar{p})^2 + \sum_{j=1}^{n} \sigma_i^2}.$$  (1)

The value of $r^2$ computed by substituting into the equation values of $p_i$ from the CABG surgeon data was 29.9% for 100 patients, 63.1% for 400 patients and 81.0% for 1000 patients. These numbers are very similar to the means of the corresponding $r^2$ in the simulated CABG surgeon data: 30.2% 63.3%, and 81.3%. Equation (1) provided equally good estimates of the simulated $r^2$ when the underlying $p_i$ values were from a gamma distribution instead of a normal distribution.

Because the equation can be used to accurately compute $r^2$, it can be used to identify the factors that affect $r^2$. Equation (1) shows that the percentage of variation explained by the underlying probabilities of mortality increases with the following: (1) increasing variation in the underlying probabilities, (2) more patients per provider, and (3) values of the probabilities closer to zero.

**COMPARISON OF RESULTS BY DATA SET**

To illustrate how the results may vary across data sets, we compared the results for the CABG data and Medicare Hospital Mortality data. As shown in Table 3, the variation in the underlying probability of adverse outcomes was very similar for each of the two simulated data sets (1.12 for the CABG surgery data set and 1.19 for the HCFA hospital data set). The average probability, however, was much higher in the HCFA hospital data set than in the CABG surgery data set, 9.17% compared to 3.05%. Because of the higher mortality rates in the HCFA data set, the percentage of variation explained and the positive predictive values were larger for the CABG data set. For example, when there were 400 patients per provider, the percentage of variation in the observed rates explained by the underlying probability was 63% for the simulated CABG data and 40% for the simulated Medicare data. To achieve an $r^2$ of 75% required 1800 patients per provider in the medicine data set but only 800 patients per provider in the CABG data set.

For a given number of patients per provider, the positive predictive values were also higher in the CABG surgery data set than in the Medicare data set. For example, when there were 400 patients per provider, the positive predictive value of using an adverse outcome rate in the lowest 10% for identifying a best surgeon was 59.3% in the CABG surgery data set and 42.8% in the Medicare data set.

**ESTIMATING $r^2$ IN REAL DATA SETS**

In the simulation study, we could find the value of $r^2$ from equation (1) only because we knew the value of all the underlying probabilities, $p_i$. With real data, however, only the values of the adverse outcome rates, $p_{fi}$, are observed, and the values of the underlying probabilities, $p_i$, are unknown. The equation below was derived in the Appendix to estimate $r^2$ from real data:

$$E(r^2) \approx 1 - \frac{k(p_{\bar{p}} - \bar{p})^2}{(k - 1)\sigma_{\bar{p}}^2},$$  (2)

where $k$ is the number of providers and $\sigma_{\bar{p}}^2$ is the variance of the observed rates. The calculations are shown as follows. If the variance of the mortality rates for a given

<table>
<thead>
<tr>
<th>TABLE 3. Relationship of sample size to the accuracy of provider assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample size</strong></td>
</tr>
<tr>
<td>CABG Simulation</td>
</tr>
<tr>
<td><strong>Variation explained by Underlying rate</strong></td>
</tr>
<tr>
<td>Another observed rate</td>
</tr>
<tr>
<td><strong>Positive predictive value for identifying</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

*Best providers have underlying probabilities in the lowest 10th percentile, good providers in the lowest 25th percentile, worst providers in the highest 10th percentile and bad providers in the highest 25th percentile.*
data set with 300 providers is 0.0219%, the average mortality rate is 9.17%, and there are 1000 patients per provider, then
\[ r^2 = 1 - \frac{300(0.0911)(0.00219)(299)}{0.0655} = 1 - 0.0250 = 0.618. \]

As an indication of the ability of this equation to estimate \( r^2 \), we tested it on data simulating hospitals with 100, 400, and 1000 patients, as shown in Table 4. The distribution of the estimated \( r^2 \) is shown with the correct value of \( r^2 \).

When there were 400 patients per provider, 90% of the estimates for \( r^2 \) were in a narrow range, 55.4–68.8%. Although a given estimate of \( r^2 \) will have some error, it will usually be sufficiently precise to give a good qualitative indication of the value of \( r^2 \) for a given data set.

### DISCUSSION

We evaluated the relationship between the number of patients per provider and the reliability of comparing mortality rates. The evaluation was performed using two computer-generated data sets that simulated important examples of data sets used for outcome comparisons: data on CABG surgeons from New York and Pennsylvania and Medicare hospital mortality data. In the simulated data sets, variation among providers was due only to variation in quality and random variation; there was no variation due to inadequate risk-adjustment.

A significant percentage of the outcome variation in the simulated data sets was due to chance. It required at least 400 cases per surgeon in the CABG surgery data set and 1000 cases per hospital in the Medicare data set for the percentage of variation due to random factors to be less than 40%. Since comparisons of hospitals often involve a subset of patients, e.g. those who have a myocardial infarction, there may be not be enough patients in the hospitals to obtain an accurate assessment of their relative underlying probabilities of adverse outcomes.

### FACTORS INFLUENCING THE VALUE OF RISK-ADJUSTED OUTCOMES

Sample size greatly influenced the precision of risk-adjusted outcomes as measures of quality of care. With increasing sample sizes, there was decreasing random variation and increasing positive predictive values for identifying the best or worst providers. It is clear from the differences in the results for the two simulated data sets, however, that the validity of risk-adjusted outcome comparisons was affected by factors other than sample size.

Equation (1) can be used to explain the different results for the two data sets. As shown in this equation, the percentage of variation in observed rates that is not due to chance, i.e. \( r^2 \), is increased if the underlying probabilities have a greater variance and are closer to zero. The value of \( r^2 \) was higher in the simulated CABG surgery data set than the Medicare hospital data set because the latter data set had higher adverse outcome rates and no corresponding increase in the variation of the underlying adverse outcome rates. In general, there is likely to be less underlying variation among hospital or other institutional providers after adjusting for patient risk than among individual physicians since the results for institutional providers are the average of the results of many individual physicians. Variation among hospitals will be further reduced if outcomes are compared for all patients (e.g. hospital adverse outcome rate) rather than for specific conditions (e.g. adverse outcome rate for myocardial infarction patients).

The value of \( r^2 \) can be estimated easily and with reasonable accuracy using equation (2). To estimate \( r^2 \), it is necessary to know the average rate of adverse outcomes, the variance in the observed rates of adverse outcomes, and the number of patients per provider. A low estimate of \( r^2 \) can be obtained by substituting into equation (2) the number of patients for providers with relatively few patients, e.g. in the lowest 10th percentile. If this estimated value of \( r^2 \) is low, e.g. less than 50%, then comparison involving providers with the low sample size will be inadequate, and these providers should be eliminated before making comparisons.

### CONCLUSIONS

The results from this study suggest that, even if risk adjustment is adequate and the outcome measure is appropriate, limitations due to sample size may remain. Although the imprecision in statistical estimates due to small sample sizes is true in all statistical studies, the effect of small sample size has been largely ignored in studies of outcomes research. The present study quantified the degree that the sample size may limit the accuracy of quality assessment in certain types of commonly used data sets. The results suggest that, unless the number of patients per provider is large, judgments of the quality of
individual providers should be made cautiously and need to be substantiated with other evidence.

Acknowledgements: The authors appreciate Ms Teri Wermager’s work in preparing the manuscript, the helpful editorial comments of Ms Clare Guse, MS, and the statistical help of Dr Robert Kuhn.

REFERENCES


APPENDIX

Using standard methods from analysis of variance, it is possible to partition the total variation of the observed adverse outcome rates into two components: (1) the variation in the underlying probabilities of an adverse outcome among providers and (2) the random variation in observed adverse outcome rates within each provider. Assuming that the average of all of the observed rates for a given provider equals the underlying probability of an adverse outcome for that provider, the equation for this partitioning is:

\[ \sum_{i=1}^{k} \sum_{j=1}^{1000} (\hat{P}_{ij} - \bar{P})^2 = \sum_{i=1}^{k} \sum_{j=1}^{1000} (p_i - \bar{P})^2 + \sum_{i=1}^{k} \sum_{j=1}^{1000} (\hat{P}_{ij} - p_i)^2, \]

where there are k providers and 1000 simulations per provider, \( \hat{P}_{ij} \) is the observed rate of an adverse outcome for the \( j \)th simulation of the \( i \)th provider, \( p_i \) is the underlying probability of an adverse outcome for the \( i \)th provider and \( \bar{P} \) is the mean of the \( p_i \). Since \( p_i \) are constant for each simulation, the expression

\[ \sum_{i=1}^{k} \sum_{j=1}^{1000} (p_i - \bar{P})^2 \]
can be simplified to

$$
\sum_{i=1}^{k} 1000(p_i - \bar{p})^2.
$$

Since $r^2$ is the proportion of the total variation that is due to the variation in the $p_i$ among providers,

$$
r^2 = \frac{\sum_{i=1}^{k} 1000(p_i - \bar{p})^2}{\sum_{i=1}^{k} 1000(p_i - \bar{p})^2 + \sum_{i=1}^{k} \sum_{j=1}^{1000}(\hat{p}_j - \bar{p})^2}.
$$

By substituting $p_i(1-p_i)/n$ for $(\hat{p}_i - p_i)^2$, we obtain the following estimate of the expected value of $r^2$.

$$
E(r^2) \approx \frac{\sum_{i=1}^{k} (p_i - \bar{p})^2}{\sum_{i=1}^{k} (p_i - \bar{p})^2 + \sum_{i=1}^{k} \frac{(1-p_i)}{n_i}}.
$$

(A2)

To estimate the value of $r^2$ from real data, it is necessary to estimate the value of $\sum(p_i - \bar{p})^2$. This estimate was obtained and dividing all terms by 1000 from equation A1. Here, equation A1 was simplified by assuming one simulation instead of 1000 simulations. By taking the expected value of both sides of this simplified equation and rearranging the terms, we get

$$
\sum_{i=1}^{k} (p_i - \bar{p})^2 = E\left(\sum_{i=1}^{k} (\hat{p}_i - \bar{p})^2\right) - E\left(\sum_{i=1}^{k} (\hat{p}_i - p_i)^2\right)
$$

$$
= E\left(\sum_{i=1}^{k} (\hat{p}_i - \bar{p})^2\right) - \sum_{i=1}^{k} \frac{p_i(1-p_i)}{n_i}
$$

$$
\approx \sum_{i=1}^{k} (\hat{p}_i - \bar{p})^2 - \sum_{i=1}^{k} \frac{\hat{p}_i(1-\hat{p}_i)}{n_i}.
$$

Substituting estimates for actual values in equation (2) gives

$$
E(r^2) \approx \frac{\sum_{i=1}^{k} (\hat{p}_i - \bar{p})^2 - \sum_{i=1}^{k} \hat{p}_i(1-\hat{p}_i)}{\sum_{i=1}^{k} (\hat{p}_i - \bar{p})^2}.
$$

A simple approximation to the numerator can be expressed by substituting $\frac{k\bar{p}(1-\bar{p})}{n}$ for

$$
\sum_{i=1}^{k} \frac{p_i(1-p_i)}{n_i}.
$$

With this simplification,

$$
E(r^2) \approx 1 - \frac{(k\bar{p}(1-\bar{p}))/n}{(k-1)s^2_p}.
$$

(A3)

where $k$ is the number of providers and $s^2_p$ is the variance of the observed rates. We used $(k-1)s^2_p$ instead of $\sum(\hat{p}_i - \bar{p})^2$ because the variance of the observed rates is more likely to be reported or easily obtained than $\sum(\hat{p}_i - \bar{p})^2$. 

Substituting estimates for actual values in equation (2) gives