International Journal for Quality in Health Care 2001; **Volume 13, Number 6**: pp. 481–488

Statistical issues in reporting quality data: small samples and casemix variation

ALAN M. ZASLAVSKY

Department of Health Care Policy, Harvard Medical School, Boston, USA

Abstract

Purpose. To present two key statistical issues that arise in analysis and reporting of quality data.

Summary. Casemix variation is relevant to quality reporting when the units being measured have differing distributions of patient characteristics that also affect the quality outcome. When this is the case, adjustment using stratification or regression may be appropriate. Such adjustments may be controversial when the patient characteristic does not have an obvious relationship to the outcome. Stratified reporting poses problems for sample size and reporting format, but may be useful when casemix effects vary across units. Although there are no absolute standards of reliability, high reliabilities (interunit $F \geq 10$ or reliability ≥ 0.9) are desirable for distinguishing above- and below-average units. When small or unequal sample sizes complicate reporting, precision may be improved using indirect estimation techniques that incorporate auxiliary information, and 'shrinkage' estimation can help to summarize the strength of evidence about units with small samples.

Conclusions. With broader understanding of casemix adjustment and methods for analyzing small samples, quality data can be analysed and reported more accurately.

Keywords: casemix adjustment, hierarchical models, quality measurement, quality reporting, regression, significance tests, stratification

The primary focus of quality measurement efforts is quite properly placed on defining measures that are valid and reliable, yet analysis and reporting of quality measures also present important challenges.

Although we usually speak of 'quality measurement', in fact our concern is often not simply to report what has already happened but rather to make an inference predicting future outcomes at the unit of interest. In this paper we consider two issues that complicate such inferences: variations in casemix across the measured units, and the effects of small sample sizes on analysis and reporting of quality data.

In the following discussion, we use a generic terminology to describe a wide variety of quality measurement applications. Quality is measured for a 'unit', which may be a hospital, clinic, health plan, medical group, individual physician, or any other institution or individual affecting quality of care. 'Patients' may be related to a unit as patients or as health plan members. The 'outcome' of a quality measure may be a clinical outcome (e.g. mortality) or a process (e.g. providing an appropriate diagnostic test or treatment, waiting time for treatment, quality of interaction with physician), and may be derived from clinical records, administrative records (e.g. claims data) or patient reports (surveys).

We assume that measurement is implemented in order to compare quality across units, possibly for quality improvement, incentive reimbursement, or consumer choice.

Casemix variation and adjustment

In comparison of quality across units, there is an implied question: how might outcomes have differed if the same patient had been treated at a different unit? This is the relevant question for decisions about where to seek treatment. Hence, differences among units attributable to characteristics of the patients rather than of the unit itself are not of interest and should be 'removed' from the comparisons.

The characteristics of patients at a unit have a distribution, called the casemix. Casemix is relevant to quality measurement if (1) outcomes are related to characteristics of the patient, within unit, and (2) the same characteristics have different distributions at different units [1]. For example, consider a common measure of quality of cardiac care, receipt of β-blockers after an acute myocardial infarction (AMI), and one characteristic, presence of chronic obstructive pulmonary disease (COPD), a contraindication for this treatment. Imagine first that at each hospital, COPD patients with an AMI

Address reprint requests to A. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115–5899, USA. E-mail: zaslavsk@hcp.med.harvard.edu

are more likely (by a fixed amount) to receive β-blockers than those without COPD, and some hospitals treat many patients with COPD while others treat few such patients. Then hospital A, with more COPD patients, would score lower than hospital B, with fewer, if their processes are identical (i.e. a COPD patient is equally likely to receive β-blockers at either hospital, and so is a non-COPD patient). This observed difference in rates creates a misleading impression that hospital A is worse than hospital B, although it makes no difference at which hospital any given patient is treated.

If at each hospital, patients with and without COPD are equally likely to receive β-blockers [condition (1) does not hold], then the varying proportions of COPD patients at the different hospitals do not affect measurement. Conversely, if rates differ by presence of COPD but the prevalence of COPD among AMI patients is the same at every hospital [condition (2) does not hold], then the relative ratings of hospitals would again be unaffected, as every hospital's rate is a fixed combination of the rates with and without COPD.

Casemix adjustment of a quality measure corrects for differences in casemix across units to estimate differences in expected outcomes for the same patients at different units [2,3]. If casemix effects are substantial relative to differences among units, then to ignore them would be unfair to the units with more patients with characteristics associated with worse performance.

Casemix adjustment is formally similar to risk adjustment of prospective payments, designed to pay each unit the typical costs for patients with characteristics like those it treated [4]. Risk adjustment is essential to pay units fairly for providing care when some patients are more expensive to treat than others [5]. Further rationales for risk adjustment have parallels in casemix adjustment of quality measures. Inadequate risk adjustment threatens the financial stability of units that treat more adverse (expensive) patient populations; failure to casemix adjust a quality measure may threaten competitiveness of units with adverse casemix if patients, referring providers, purchasers, or accreditation bodies make decisions based on quality measures. Similarly, inadequate risk adjustment gives units an incentive to avoid treating patients with high expected costs. In theory inadequate casemix adjustment for quality creates similar incentives. Quality measures, however, have a less direct impact than payments on the unit's institutional success, and the effects of patient characteristics on quality measures are less manifest than effects on costs. Hence such perverse incentives may have less force in quality measurement.

Casemix adjustment involves a judgement that the differences are 'not attributable to the unit being evaluated', which is normative and clinical, not statistical. This concept defines the question of interest: casemix characteristics are those hypothesized to remain the same if the patient were assigned to a different unit. Thus, characteristics and processes of the unit are not appropriate casemix adjustors. Although handwashing by surgeons is known to be (negatively) predictive of postpartum infection, we would not adjust for this variable because if the patient moved to a hospital where hands are more assiduously washed, outcomes would predictably be better. Intermediate outcomes also are rarely casemix adjustors: we would not usually adjust postsurgical mortality for the frequency of postsurgical complications. We might do so, however, if we were specifically evaluating rescue of complicated patients, or if we believed that the patients would have experienced the same complications at any hospital. Similarly, we would typically adjust for differences between measures for medical and surgical patients, but not if the decision to treat medically or surgically for a particular condition were determined by the policies of the unit being evaluated.

Table 1 Illustration of effect of casemix differences with two strata

	Hospital A	Hospital B	Hospital C
(1) Complication rates, medical	45%	50%	30%
(2) Complication rates, surgical	25%	30%	??
(3) Percent medical patients	60%	20%	100%
(4) Percent surgical patients	40%	80%	0%
(5) Observed overall rate =			
(1) × (3) + (2) × (4)	37%	34%	30%
(6) Adjusted rate[1]	33%	38%	??

[1] Adjusted rate assuming average casemix of 40% medical, 60% surgical, calculated as (1) × 40% + (2) × 60%.

We now consider several casemix adjustment methods, considering first approaches yielding a single-number summary of quality and then those conveying additional information by reporting more than one summary.

To illustrate direct standardization, imagine that we wish to report a single summary of complication rates unaffected by the mix of medical and surgical patients. We might calculate rates for each stratum and then combine them in fixed proportions, such as the average fractions of surgical and medical patients across all units being compared. Table 1 illustrates the calculations for two hospitals with very different medical/surgical mixes. Hospital A has lower rates than hospital B in each stratum, but a higher unadjusted rate, due to its heavier caseload of medical cases. Standardization removes the casemix effect, revealing the superior performance of hospital A. Confidence intervals can be calculated for directly standardized rates [6].

Direct standardization is simple and independent of modeling assumptions, but has limitations. It may be impossible to calculate a standardized score for a unit because it has no cases in a stratum; in Table 1, hospital C has no surgical patients which means the standardized rate cannot be calculated. If a hospital has unusually few cases in a stratum, those cases may determine a disproportionate part of the standardized score, making it statistically unstable. Furthermore, direct standardization is not adapted to adjusting simultaneously for many variables or for graded (continuous) variables.

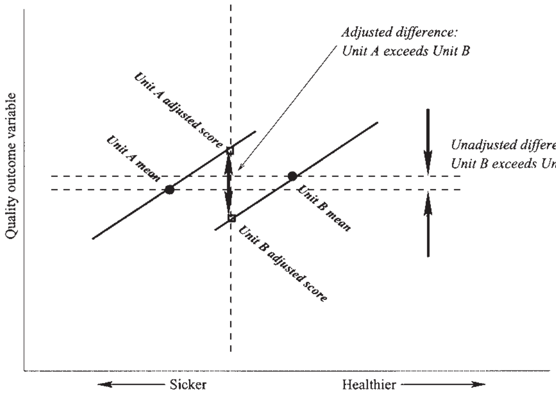

Figure 1 Casemix adjustment using linear regression.

An alternative approach, indirect standardization, deals with the first of these problems by calculating an expected incidence for each unit using rates by stratum from a standard population, and then comparing the observed to the expected rates. This approach uses models more than direct standardization but has some of the same limitations.

A more flexible approach to adjustment is through regression modeling. Typically, we model the predicted outcome for each patient as the sum of a component due to measured patient characteristics and one due to the unit at which she is treated. Mathematically, $y_{up} = \beta\ x_{up} + \mu_u + e_{up}$, where y_{up} is the outcome for patient p at unit u, x_{up} is the corresponding patient characteristic(s), μ_u is an effect for unit u, e_{up} is an error term, and β is a coefficient. The first term of this equation captures the effects of individual characteristics x on outcomes, among patients of the same unit (all sharing the same value of μ_u). Quality differences among units are captured in the unit-specific effects μ_u, which determines differences among patients with the same characteristics x_{up} treated at different units. In some simple situations, regression analysis is closely related to direct standardization [7]. Note that while the model controls for unit effects through the dummy variables μ_u, characteristics of units do not appear in the model and their effects cannot be estimated.

This analysis is illustrated (for a single casemix variable) in Figure 1. The dots indicate mean values of casemix and outcome in two units, and the lines through them indicate the distribution of the variables and their relationship. Unit A is slightly below unit B in raw scores, but higher after adjustment.

These models can be fitted using standard regression software. Adjusted scores for unit i can be obtained by calculating model predictions for a standard population (such as the pooled sample from all the units), or equivalently by inserting the mean value of the covariates x. In either case the scores obtained can be interpreted as a prediction of the mean outcomes if the same population (or the same 'average' individual patient) had been treated at each of the different units.

In some cases a non-linear regression (such as logistic regression for a dichotomous outcome, or linear modeling of log-transformed costs) is suggested by the structure of the data. With non-linear models, substituting mean predictors will differ from averaging predictions over a population, but the ordering of the adjusted values will be the same.

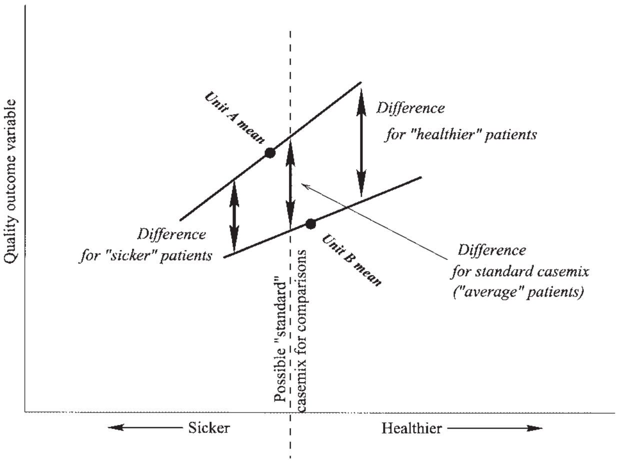

Figure 2 Casemix adjustment when effects of covariates are not the same at every unit.

Regression models accommodate a large number of adjustor variables using simplifying assumptions about how their effects combine (e.g. that they are additive and linear if so specified in the model), hence depending to some degree on the validity of the assumptions. For small adjustments, sensitivity to departures from models is modest. Very large adjustments (where the distributions of casemix in different units are distinct) use the model to extrapolate beyond the range of the data at any unit, and therefore are critically dependent on model specification. Direct standardization signals when the units have distinct distributions, because some cells will have no cases at some units. Regression allows the careless analyst to attempt extreme adjustments that are overly dependent on model assumptions. Sensitivity of results to the details of the adjustment method may also be of concern [8,9].

An interesting situation arises when the effect of casemix variables on outcomes is not the same at each unit [10,11]. For example, the difference in outcomes between 'sicker' and 'healthier' patients may be larger at hospital A than at hospital B. In that case, the comparison of the two hospitals for 'healthier' patients will differ from that for 'sicker' patients. With direct standardization, this will be reflected in systematically different comparisons in the various strata. In a regression analysis, the regression lines will have different slopes (Figure 2).

Under these circumstances, no single value fully characterizes the differences in performance between units, although summary values still can be calculated. Direct standardization reports the average of the differences between hospitals for 'healthier' and 'sicker' patients, weighted according to their prevalences in the reference population. Regression adjustment using the average coefficient across units may be acceptable if both differences in casemix and differences in regression coefficients among units are modest. Better yet, comparisons can be made at a standard value of the covariates (the vertical dotted line in Figure 2), adjusting each unit to that value using its own coefficient.

Nonetheless, these summaries do not completely represent the experiences of subgroups of patients. In Figure 2, 'healthy' patients would find hospital A much better than hospital B, while 'sicker' patient might expect little difference. Extrapolating to the left, some patients might even find hospital B superior to hospital A. If these differences are large enough, it may be worthwhile to perform a stratified analysis, calculating and reporting separate comparisons for various subpopulations. In practice, however, samples may be too small to support subgroup analyses, and consumers of quality data may not be prepared for the additional complexity of reports that provide several comparisons for each measure. Stratified analysis is worthwhile only when i) stratified results can be calculated with adequate reliability and ii) the reports for different strata are likely to have substantially different implications for users of the reports. The reports might, for example, lead to different decisions about choice of provider or identify quality problems for specific subpopulations.

Casemix adjustment can be controversial when the clinical reasons for the relationship between a casemix variable and the quality outcome are not apparent, especially if the variable distinguishes a vulnerable population. For example, outcomes on the Health Plan Employer Data and Information Set (HEDIS®) clinical measures are negatively related to the concentration of racial/ethnic minorities and poverty in the patient's area of residence [12]. Casemix adjustment would raise the scores of health plans drawing large numbers of members from such areas. Whether the results are adjusted or not, the poorer quality of care received by the more vulnerable members within each plan is invisible in aggregated plan-level results, even knowing which plans had more members from these groups. If each plan's results are stratified by neighborhood poverty, the poorer quality of care received by poverty-area residents is detectable only with a careful analysis to summarize the differences between groups across plans. Yet if the difference between quality for low- and high-income patients is approximately the same at each plan, the same information could be presented more efficiently using a two-part summary, consisting of i) casemix-adjusted ratings for each plan, and ii) coefficients from the casemix model, representing average quality differences among socio-demographic subgroups. The latter component powerfully summarizes inequities affecting all plans. Arguably, casemix coefficients are most interesting and worth reporting when they represent quality variations that are not clinically inevitable, but rather are potentially correctable.

A comment on the HEDIS® finding [13] argued that adjustment would reward plans that are failing to provide good service to members of vulnerable populations, excusing them for inequities in quality of care. If the magnitude of the inequities is similar across units, however, then all units are equally responsible, even those that enroll relatively few patients from the underserved groups. Such uniform differences are better identified and addressed as systemic problems, rather than by penalizing units that serve these patients. For example, the fact that less-educated patients more often miss screening tests may reflect a weakness in the way providers communicate information regardless of which health plan covers the patients.

When quality differences among patient groups vary across units, on the other hand, stratified reporting may be worthwhile. Alternatively, reports could summarize overall plan performance using an average casemix adjustment, and then show subgroup differences for each unit separately, distinguishing between overall quality and its equitable distribution.

When the subgroups of interest are small relative to the population (e.g. minority ethnic groups, patients with chronic conditions) stratified reporting may be impractical. At least, analysts should estimate the amount of variation in subgroup effects on quality across units. When subgroup differences are large for some units and smaller for others, adopting 'best practices' on equity of care may help to close the quality gap.

Statistical variation and small samples

Whenever a measure is based on a sample from a larger population, random variation is introduced by sampling; by chance, a larger or smaller than average rate of successful outcomes will appear in the sample than in the population. The amount of variation in the measure is related to the sample size by well-known statistical formulae. Specifically, the standard error (SE) of the measure is inversely proportional to the square root of sample size, so multiplying sample size by four halves the SE.

When the sample contains all relevant cases during the period of measurement, then there is no sampling variation in the estimate for that period (although there may be variation from other sources such as measurement error). Historical facts about outcomes during a specific period are important, though, primarily to predict the likely outcomes for a larger hypothetical population of patients. We rate health plans to help potential members to decide where to enroll, and we rate surgeons to predict their performance with their future patients, not to reward them for their past achievements. For this purpose, the fraction of the population that is measured is irrelevant. Our predictions for the future are more precise for a surgeon who performed 200 operations of which we have data for 100, then for a surgeon who performed 20 operations of which we have data for all 20. (Hence the 'finite population correction' for variance estimates from descriptive surveys is not relevant to quality measurement.)

To protect the consumer from being misled by chance fluctuations we commonly report both the estimates and some measure of uncertainty. One popular graphical presentation uses error bars, representing a 'confidence interval' about the estimated value (Figure 3, left side), constructed so that the probability that the interval will contain the population value equals a prespecified level. The specific formula for the confidence interval depends on the estimand and perhaps on features of the sample design, such as stratification and clustering. For given data, we can be more confident that the interval contains the true value if we make the interval wider. The conventional 95% level represents a

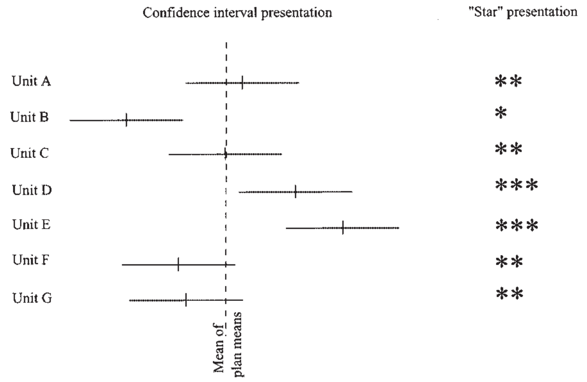

Figure 3 Two graphical displays for comparisons of quality.

Table 2 Probability that a plan 1 SD above mean will be ‘significantly above average’ (3-star rating)

F[1]	IUR[2]	Probability (%) of 3-star classification
2	0.500	16.9
3	0.667	29.3
4	0.750	41.0
5	0.800	51.6
6	0.833	60.9
7	0.857	68.8
8	0.875	75.4
9	0.889	80.7
10	0.900	85.1
15	0.933	96.3
20	0.950	99.2

[1] F statistic from 1-way ANOVA test of differences among unit means.
[2] Inter-Unit Reliability.

compromise that assures a reasonable level of coverage (95% probability that the interval includes the population value) without making the interval too wide (±2 SE, or altogether 4 SE wide).

A graphic like Figure 3 directs the viewer’s attention to comparisons among units. If a reference line is included for the mean of all units, it emphasizes the comparison between each unit and that mean. If the error bar does not cross the reference line, the unit is significantly better or worse than the average of all units. (The SE of the difference from the overall mean is slightly different from that for the unit mean itself, usually smaller, but with a fairly large number of units, the difference is minimal.) A simple summary of this comparison [standard for Consumer Assessments of Health Plans (CAHPS®) displays] [14] reports only whether the unit is significantly below, significantly above, or not significantly different from the mean of all units (signified by 1, 3 or 2 stars respectively), as illustrated on the right-hand side of Figure 3. This style of report removes much of the detail in an ‘error bar’ report; for example, it does not indicate the difference in quality between units D and E in Figure 3. Furthermore, the cutoffs are determined by criteria of statistical, rather than clinical, significance, and therefore affected by sample sizes. Nonetheless, this report classifies plans simply for consumers.

The estimated mean (or proportion) for each unit is based on data from some sample of patients, members, or cases. We can interpret the estimate as the sum of a population mean (for all patients of that unit, or a larger potential populations of patients as described earlier) and random error due to sampling. Consequently the variation among the means for different units is the sum of variation because the units actually differ from each other (the ‘signal’ that we wish to measure), and variation due to sampling error (‘noise’). The usefulness of the report is determined in part by the relative magnitude of these components, as results that are mainly determined by random noise are highly misleading and encourage unwarranted inferences about quality.

When the standard errors of estimates for each unit are similar (primarily determined by whether units have similar sample sizes), an overall summary of the relative magnitude of variation due to signal and noise is useful. Let F be the usual statistic from the F-test of equality of the unit effects (readily obtained from standard regression programs such as SAS PROC GLM). Then F-1 is an estimate of the ratio of signal to noise, and $1–1/F$ estimates the fraction of total variance that is due to signal (real variation among units), also known as the ‘interunit reliability’ (IUR) [15]. When F is large (IUR close to 1), the measure distinguishes reliably among units; conversely, when F is small (IUR tending toward 0), no reliable distinctions can be made.

What, then, is a ‘large enough’ value of F? While there is no single standard for all applications, the following argument suggests the implications of various values of the F statistic. Suppose that a large number of units are being compared and that the distributions of both the true means and sampling errors are approximately normal (a typical statistical assumption for such problems). Consider a unit that is moderately [one standard deviation (SD)] above average, hence better than about 84% of the other units. The probability that this unit will be reported as above average (‘3 stars’ by the criteria described above) depends on the reliability of the measure. If $F=10$ (IUR$=0.90$), then this probability is 85%, but if $F=4$ (IUR$=0.75$), the probability falls to 41% (Table 2 and Appendix). In either case there is highly significant evidence of variation among the units (as indicated by the F test), but only with the larger value of reliability will there be a good chance that a plan that is moderately better or worse than average can actually be declared so by the conventional test. In effect, the test is not very sensitive unless F is fairly large. Quality reports that are dominated by noise are useless and potentially counterproductive [16].

If the variation among units is large enough to be worth measuring at all, we might want to distinguish units fairly sensitively that are as high as the 84th percentile of units in true quality (or as low as the 16th percentile). Thus, F values closer to 10 than to four are desirable. The F value depends on several factors, including the sizes of the differences among the units, the amount of within-unit variation in responses and the sample size per unit. Of these, the sample size is most likely to be under the control of those designing the quality measurement program; the required sample size depends on the other factors.

Dealing with small sample sizes: indirect estimation and shrinkage

A 'small sample size' is one that does not yield adequately precise estimates for the intended purposes. In practice, considerations of cost, burden on patients or staff, or limited numbers of relevant cases may deny us the sample sizes required for desirable levels of precision, at some or all units. At worst, useful reports become impossible. When samples are small, however, statistical methodology may make valuable contributions. We now consider two relevant concepts: indirect estimation and hierarchical modeling.

Broadly speaking, direct estimation bases the estimate for each domain (e.g. for a unit's performance during a specific time period) only on information from that domain. An indirect estimate combines information from many domains to improve estimation for all of them. Direct estimation is simpler than indirect estimation, and requires less justification; the preceding discussion has assumed that direct estimators are used. When direct estimators are insufficiently precise, indirect estimators become an attractive alternative.

Indirect estimation may use relationships with 'auxiliary variables' that are well measured and have a systematic relationship with the outcome of interest. For example, in estimating mortality rates in cardiac care, we might use information about whether the hospital is urban or rural, whether it is an academic center, and its size, each of which is related to mortality rates [17]. Because relationships are quantified using data from all units in the study, estimates using them are indirect.

Relationships of outcomes with a continuous auxiliary variable or with more than one variable are characterized using regression modeling. The interpretation, however, is opposite to that for casemix adjustment. Here, the covariates are characteristics of the unit (or for which the unit is responsible) and their effects are part of the predicted measure for the unit, while in casemix adjustment the covariates are characteristics of the patients and their effects are removed from the reported measure. The differences in interpretation depend to some degree on the face validity of the relationships and whether the characteristics in question are intrinsic to the institution or brought to it by the patients. For example, to decide whether to use rural location as a predictor of quality, we might investigate whether rural patients tend to have worse outcomes at all hospitals (an individual-level effect) or conversely whether patients from the same area of residence have different outcomes at urban versus rural hospitals.

Volume-quality relationships have been established for some types of cancer care, but the relationships are often weak and the processes underlying them are not well understood [18]. Because of this, their use in predicting quality for individual hospitals is somewhat controversial. Furthermore, to use volume as an important predictor of quality could also have serious policy implications, such as reducing access to care in remote areas by closing low-volume treatment centers. Nonetheless, even though the volume-quality relationship does not conclusively prove inferior quality at any particular small hospital, it still might pass the 'grandmother test' ('would this influence your informed decision on where to send your loved one for treatment?'). In this case, more direct measures of hospital processes would be desirable but are not readily available. Preferably, relationships used in predicting quality for reporting should be scientifically well established and face-valid. Furthermore, only fairly strongly related auxiliary variables can contribute substantially to precision.

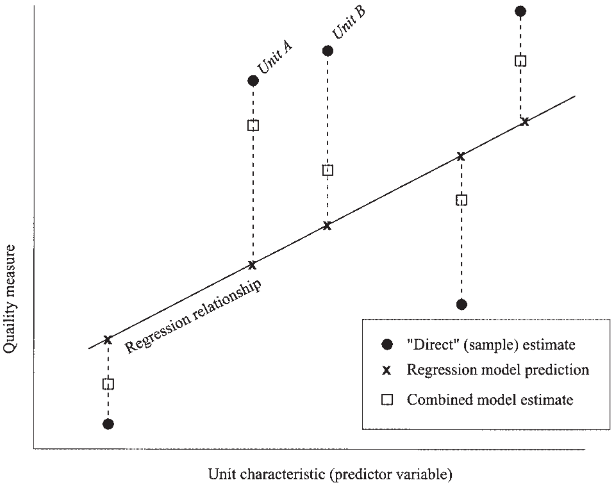

Figure 4 Combining regression predictions and direct estimates using a hierarchical model.

The performance of the same unit in earlier time periods may be regarded as an auxiliary variable for predicting current performance; hence, a moving average of a quality measure is a time-indirect estimator. Like any indirect estimator, its properties must be evaluated in relation to the objectives of estimation. A moving average adds precision when samples are small in each time period, but is insensitive to changes in quality from one period to the next.

Optimally, indirect estimates (such as regression predictions) and direct estimates (data from the unit in question) should be used together. Each of these sources provides some information but with limited precision, due to the imprecision of the relationships underlying indirect estimates and the sampling variability of the direct estimates. Hierarchical modeling provides a framework within which the two types of information can be combined, weighting each in proportion to its precision ('credibility'). While a methodological exposition of hierarchical modeling [19–21] is beyond the scope of this article, Figure 4 expresses some of the key ideas graphically. Predictions from the regression model (the solid line) are combined with direct estimates for each unit (dots) to obtain improved estimates combining both. For unit A, with a large sample, the final estimates are close to the direct estimates. For unit B, with a small sample, the combined estimates are relatively close to the regression prediction, since the sample data provide relatively little information specific to the unit.

Even if there were no auxiliary variables (and the line in Figure 4 was horizontal), estimates from the hierarchical model would be pulled toward that line, reflecting the observed degree of similarity among units. Because the model estimates are closer together than the direct estimates, the former are sometimes called 'shrinkage estimates'. Units with small samples are 'shrunk' toward the mean more than those with larger samples, reflecting the weaker evidence that they differ from the mean.

If all units have large samples or samples of equal sizes, the effects of shrinkage are not informative because they are negligible or uniform. Shrinkage is important, however, when some sample sizes are small. Imagine, for example, that a surgical procedure has a mean success rate of 50%. Naively using direct estimates, the 'best' and 'worst' units might each have a single case, in either a success (100% success) or a failure (0% success). With shrinkage estimation, these estimates would be pulled strongly toward the mean, and a unit with 100 cases and 70% success might emerge as the more likely leader. This conclusion, despite the technical statistics that underlies it, passes the grandmother test, and similar results may be found whenever sample sizes vary greatly [17,22].

Conclusion

Collecting quality data is often difficult and expensive. Hence investing in statistical methodology to improve the analysis and reporting of these data is worthwhile.

Although the methods described here are of varying technical difficulty, the limiting factor often will not be the mechanics of the methodologies, but building consensus around the choice of methodology for an intended purpose. In some areas the methods described here are widely used, e.g. for rating schools [23] and for distributing funds to local governments [24], and applications to health quality measurement are also being developed [17,22,25–27]. Improved understanding of the rationale for these analytic methods will facilitate their wider adoption in quality reporting, making quality reports more useful and accurate.

Acknowledgement

This article was prepared with support from the Agency for Healthcare Research and Quality (grants HS09473 and HS09205) and the Health Care Financing Administration (contract 500-95-007).

References

1. Zaslavsky AM. Issues in case-mix adjustment of measures of the quality of health plans, Proceedings, Section on Social Statistics, 1998. Alexandria, VA: American Statistical Association.
2. Iezzoni LI. The risks of risks adjustment. *J Am Med Assoc* 1997; **278:** 1600–1607.
3. Berlowitz D, Ash A, Hickey E *et al.* Profiling outcomes of ambulatory care: casemix affects perceived performance. *Med Care* 1998; **36:** 928–933.
4. Iezzoni LI, Ayanian JZ, Bates D, Burstin H. Paying more fairly for Medicare capitated care. *N Engl J Med* 1998; **339:** 1933–1937.
5. Kuttner R. The risk adjustment debate. *N Engl J Med* 1998; **339:** 1952–1956.
6. Keyfitz N. Sampling variance of the standardized mortality rates. *Hum Biol* 1966; **38:** 309–317.
7. Little, RJ. Direct standardization: a tool for teaching linear models for unbalanced data. *Am Stat* 1982; **36:** 38–43.
8. Iezzoni LI, Ash AS, Shwartz M *et al.* Judging hospitals by severity-adjusted mortality rates: the influence of the severity-adjustment method. *Am J Public Health* 1996; **86:** 1379–1387.
9. Landon BE, Iezzoni LI, Ash AS *et al.* Judging hospitals by severity-adjusted mortality rates: the case of CABG surgery. *Inquiry* 1996; **33:** 155–166.
10. Zaslavsky AM, Zaborski L, Cleary PD. Does the effect of respondent characteristics on consumer assessments vary across health plans? *Med Care Res Rev* 2000; **57:** 379–394.
11. Elliott MN, Swartz R, Adams J *et al.* Case-mix adjustment of the National CAHPS Benchmarking Data 1.0: a violation of model assumptions? *Health Serv Res* 2001; **36:** 409–427.
12. Zaslavsky AM, Hochheimer JN, Schneider EC *et al.* Impact of sociodemographic case mix on the HEDIS measures of health plan quality. *Med Care* 2000; **38:** 981–992.
13. Romano PS. Should health plan quality measures be adjusted for case mix? *Med Care* 2000; **38:** 977–980.
14. AHCPR. CAHPS 2.0 Survey and Reporting Kit. Washington, DC: Agency for Health Care Policy and Research, 1998.
15. Fleiss J. *Design and Analysis of Clinical Experiments*. New York: John Wiley and Sons, 1966: pp. 11–15.
16. Hofer T, Hayward R, Greenfield S *et al.* The unreliability of individual physician 'report cards' for assessing the costs and quality of care of a chronic disease. *J Am Med Assoc* 1999; **281:** 2098–2105.
17. Normand SL, Glickman ME, Gatsonis C. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc* 1997; **92:** 803–814.
18. Hewitt M, Petitti D. *Interpreting the Volume-Outcomes Relationship in the Context of Cancer Care*. Washington, DC: National Academies Press, 2001.
19. Goldstein H. *Multilevel Statistical Models*. London: Edward Arnold Publishers Ltd, 1995.
20. Bryk AS, Raudenbush SW. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications, Inc., 1992.
21. Snijders T, Bosker R. *Multilevel Analysis*. London: Sage Publications, 1999.

22. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann Intern Med* 1997; **127:** 764–768.

23. Goldstein H, Spiegelhalter D. League tables and their limitations: Statistical issues in comparisons of institutional performance. *J R Stat Soc, Series A, Gen* 1996; **159:** 385–409.

24. National Research Council. Small-area Income and Poverty Estimates: Priorities for 2000 and Beyond. In Citro C, Kalton G, eds. Washington, DC: National Academy Press, 2000.

25. Gatsonis C. Profiling providers of medical care. In: Armitage P, Colton T (eds), *Encyclopedia of Biostatistics*. Chichester: Wiley, 1998.

26. Daniels M, Gatsonis C. Hierarchical generalized linear models in the analysis of variations in health care utilization. *J Am Stat Assoc* 1999; **94:** 29–42.

27. Burgess J, Christiansen C, Michalak S, Morris C. Medical profiling: improving standards and risk adjustments using hierarchical models. *J Health Econ* 2000; **19:** 291–309.

Appendix

To illustrate the performance of a conventional display of significantly above average units with various values of reliability, we assume a simple random effects model. Imagine that unit effects are normally distributed, $\mu_u \sim N(\mu_0,(F-1)\sigma^2)$, and the sample means have normal errors, $y_u \sim N(\mu_u, \sigma^2)$. For convenience of notation and without loss of generality, center and scale the measure so $\mu_0 = 0$ and $\sigma^2 = 1$. The 'between' mean square, or variance of unit means, is then approximately F and the 'within' mean square is ~1, so F represents the approximate value of the F statistic from the analysis of the data. The SD of the unit effects is $\sqrt{(F-1)}$, while to be significantly above average (two-sided 0.05-level test), assuming a large number of units, a unit's sample mean must be at least $\Phi^{-1}(.975) \approx 1.96$, where Φ is the normal cumulative distribution function. Hence the probability of significance for a unit whose unit effect (population mean) is 1 SD above average is the probability that a normal variable with unit variance and mean $\sqrt{(F-1)}$ will exceed 1.96, or $1 - \Phi(1.96 - \sqrt{(F-1)})$. This quantity is tabulated in Table 2.

Accepted for publication 10 August 2001