8-2016

Reliability of Pressure Ulcer Rates: How Precisely Can We Differentiate Among Hospital Units, and Does the Standard Signal-Noise Reliability Measure Reflect This Precision?

Vincent S. Staggs

Emily Cramer

RESEARCH IN NURSING & HEALTH

Reliability of Pressure Ulcer Rates: How Precisely Can We Differentiate Among Hospital Units, and Does the Standard Signal-Noise Reliability Measure Reflect This Precision?

Vincent S. Staggs, Emily Cramer

Correspondence to Vincent S. Staggs
E-mail: vstaggs@cmh.edu

Vincent S. Staggs
Health Services and Outcomes Research
Children's Mercy Hospitals and Clinics
School of Medicine
University of Missouri-Kansas City
2401 Gillham Road
Kansas City, MO 64108

Emily Cramer
School of Nursing
University of Kansas Medical Center
Kansas City, KS

Abstract: Hospital performance reports often include rankings of unit pressure ulcer rates. Differentiating among units on the basis of quality requires reliable measurement. Our objectives were to describe and apply methods for assessing reliability of hospital-acquired pressure ulcer rates and evaluate a standard signal-noise reliability measure as an indicator of precision of differentiation among units. Quarterly pressure ulcer data from 8,199 critical care, step-down, medical, surgical, and medical-surgical nursing units from 1,299 US hospitals were analyzed. Using beta-binomial models, we estimated between-unit variability (signal) and within-unit variability (noise) in annual unit pressure ulcer rates. Signal-noise reliability was computed as the ratio of between-unit variability to the total of between- and within-unit variability. To assess precision of differentiation among units based on ranked pressure ulcer rates, we simulated data to estimate the probabilities of a unit's observed pressure ulcer rate rank in a given sample falling within five and ten percentiles of its true rank, and the probabilities of units with ulcer rates in the highest quartile and highest decile being identified as such. We assessed the signal-noise measure as an indicator of differentiation precision by computing its correlations with these probabilities. Pressure ulcer rates based on a single year of quarterly or weekly prevalence surveys were too susceptible to noise to allow for precise differentiation among units, and signal-noise reliability was a poor indicator of precision of differentiation. To ensure precise differentiation on the basis of true differences, alternative methods of assessing reliability should be applied to measures purported to differentiate among providers or units based on quality.

Keywords: healthcare quality; patient safety; pressure ulcers; quality measurement; reliability

Research in Nursing & Health, 2016, 39, 298–305

Accepted 22 April 2016

DOI: 10.1002/nur.21727

Published online 25 May 2016 in Wiley Online Library (wileyonlinelibrary.com).

Reliability can be defined as the extent to which multiple measurements of the same quantity yield consistent results. Reliable measurements are stable, varying minimally around the "true" value of the quantity being measured. Low reliability means high variability among observed measurements, and as a result, higher likelihood of any given observed measurement not being close to the true value.

In measurement of healthcare quality and safety, we often are looking for differences—either differences among

providers or other care sources or differences across time. The extent to which the differences we observe reflect true differences is determined by the reliability of our measurements. If reliability is low, two observed measurements of the same quantity may be different, and two observed measurements of different quantities may be nearly the same. For example, a healthcare venue with a given true rate of some adverse event (e.g., a hospital's average inpatient fall rate during a year) will likely have different observed rates across measurement occasions, due simply to random variations in patient mix. Similarly, the true adverse event rates for two hospitals may be different, but their observed rates for a given time period may be indistinguishable. Thus, our ability to differentiate among care sources or time periods based on true differences is limited by the reliability of our measurements.

Importance of Reliable Measures of Health Care Quality

Reliability is a known concern in provider profiling (Adams, Mehrotra, Thomas, & McGlynn, 2010; Hofer et al., 1999). In a large study of primary care provider profiling using six quality measures and five measures of resource use, researchers commented that "most measures appear to be driven largely by chance" (Eijkenaar & van Vliet, 2014, p. 192). Nevertheless, public reporting of profile data, such as hospital rankings, is increasingly common, underscoring the need for sound methods of assessing measurement reliability.

In assessing reliability, the ideal is a true score that is relatively stable across time, or repeated measurements carried out in sufficiently brief time to ensure minimal change in the true score between measurements. However, these conditions are rarely, if ever, met in healthcare quality measurement. Quality is a moving target, changing as hospitals carry out improvement efforts and even as staff on duty change from shift to shift. And true replications—repeated measurements taken under identical conditions—are typically unavailable, given continual variation in patient mix and staff composition. This makes it difficult to distinguish temporal changes in true scores from random variability in observed measurements.

A simple approach to assessing reliability in this context is to define a provider's true score as the average over a period of time and examine variability in observed measurements within this time period. For example, taking nursing units as providers, we can define a unit's true adverse event rate as its average rate during a year and examine how monthly or quarterly rates vary around this average. Defining the average for a longer period (e.g., several years) provides more data for estimating the true rate but makes the issue of temporal variability in the true rate more problematic. On the other hand, defining the average for a shorter time period (e.g., one quarter) results in fewer repeated measurements with which to estimate random variability around the true rate. Some middle ground must be chosen.

Signal-to-Noise Ratio as a Reliability Indicator

Reliability is often quantified in terms of signal relative to noise, where signal is the amount of variance among the true quantities being measured, and noise is the amount of error variance due to randomness. The size of the signal is determined by the amount of variance among providers in the population, typically estimated in practice by variance among providers in a sample. The amount of noise is determined in part by the variability caused by unmeasured factors and in part by sample size. Error variance can come from a number of sources. In repeated measurements of adverse event rates on nursing units, for example, within-unit changes over time in patient mix or staffing result in sampling error, and differences in reporting accuracy result in measurement error. Sample size is important because larger within-unit samples—that is, larger numbers of patients—provide more measurements and reduce error variance, thus improving the reliability of measurement. This is shown for the case of pressure ulcer rate reliability treated below.

Defining reliability as $signal/(signal + noise)$, we have a measure taking values in the interval [0, 1]. This signal-noise definition is appealing for its simplicity and has been used to assess reliability of a variety of measures, including physician cost profile scores, physician diabetes care measures, hospital surgical site infection rates, and rates of inpatient falls (Adams et al., 2010; Hofer et al., 1999; Kao, Ghaferi, Ko, & Dimick, 2011; Staggs & Gajewski, 2015). However, there is no consensus on the threshold that must be met for acceptable reliability, nor is it straightforward to interpret a signal-noise reliability measure in terms of our ability to differentiate among providers or time periods.

Reliability of Pressure Ulcer Rates

Hospital-acquired pressure ulcers have been a focus of hospitals, researchers, and policymakers for some time. Inter-rater consistency in pressure ulcer risk assessment and classification has received some research attention (Bergquist-Beringer, Gajewski, Dunton, & Klaus, 2011; Kottner & Dassen, 2010; Kottner, Halfens, & Dassen, 2009; Kottner, Raeder, Halfens, & Dassen, 2009; Waugh & Bergquist-Beringer, 2016), but researchers have not assessed nursing unit pressure ulcer prevalence rates for reliability in terms of between-unit differences (signal) relative to error variance (noise), nor have they examined the extent to which these rates allow for precise differentiation among nursing units.

We can define a unit's true hospital-acquired pressure ulcer rate as the average proportion of patients on the unit who develop a pressure ulcer, or equivalently, as the probability that a randomly chosen patient will develop a pressure ulcer. The observed prevalence rate for any given day will likely differ from the true rate, due to the unique mix of patients, the patient volume and staffing ratio, and

the combination of nurses on duty—all of which are sources of sampling error—and possibly due to measurement errors in identifying pressure ulcers. When observed pressure ulcer rates are ranked for comparison reporting (a routine practice) or identification of high or low performers, the conversion from absolute to relative units introduces additional measurement error.

Our twofold purpose in the present analysis was to describe and apply methods for assessing the reliability of rates of hospital-acquired pressure ulcers, noting the methodological challenges involved, and to evaluate a standard signal-noise reliability measure as an indicator of our ability to differentiate among units or time periods using ranked pressure ulcer rates.

Methods

Sample and Data Preparation

The National Database of Nursing Quality Indicators (NDNQI) provided 2013 data on pressure ulcers. Participating hospitals submitted nursing unit-level data on hospital-acquired pressure ulcers quarterly to the NDNQI, which collected the data with oversight from a university IRB. We limited the sample to nursing units of five types (critical care, step-down, medical, surgical, and medical-surgical) with pressure ulcer data for all four quarters in 2013. There were 8,199 units from 1,299 hospitals in the sample.

The NDNQI computes pressure ulcer rates based on surveys carried out by each participating unit on 1 day each quarter. A trained survey team carries out a skin inspection of each patient on the floor, classifies each pressure ulcer as hospital- or community-acquired (i.e., present on admission), and categorizes each ulcer according to NPUAP-EPUAP (2009) guidelines as Stage I–IV, unstageable, suspected deep tissue injury, or indeterminable. Hospitals report to the NDNQI the number of patients on the unit who were assessed for pressure ulcers and the count and category of pressure ulcers observed. The NDNQI uses these data to compute unit pressure ulcer rates and, based on these rates, each unit's percentile ranking among units of the same type. More details on the NDNQI pressure ulcer data, including inter-rater reliability of ulcer identification, staging, and risk assessment, are available elsewhere (Bergquist-Beringer et al., 2011; Waugh & Bergquist-Beringer, 2016).

We summed across quarters to compute each unit's annual count of patients assessed, annual count of patients with at least one pressure ulcer, and annual count of patients with at least one pressure ulcer stage II or above. The pressure ulcer rates of interest in this study were the proportion of patients assessed as having at least one hospital-acquired pressure ulcer (the total pressure ulcer rate) and the proportion assessed as having at least one pressure ulcer stage II or above (the stage II+ pressure ulcer rate).

Analytical Framework

In the following paragraphs, we describe our analyses for a single pressure ulcer rate. We carried out these analyses twice: once for the total pressure ulcer rate and once for the stage II+ pressure ulcer rate. Thus, data for these two rates were modeled separately.

In this context, signal is the variation among units' true pressure ulcer rates, or between-unit variance. These true rates can be thought of as average rates, determined by the unit's average quality of care and average patient mix. Noise is the within-unit variance we would expect in a unit's observed pressure ulcer rates due to randomness, including random variation in patient mix. Defining reliability in terms of signal and noise we have

$$\text{signal}/(\text{signal} + \text{noise}) = \text{between-unit variance}/(\text{between-unit variance} + \text{within-unit variance}) \quad (1)$$

We modeled the pressure ulcer rates by fitting a beta-binomial model for each unit type (Adams, 2009). Let p_i denote the true pressure ulcer rate on the ith unit. In the beta-binomial framework the p_is are assumed to follow a beta-distribution, and the unit's count of pressure ulcers is modeled as a binomial(n_i, p_i) random variable, where n_i is the number of patients assessed for pressure ulcers and p_i is the probability of a patient developing at least one pressure ulcer.

In terms of signal and noise, the between-unit variance (signal) for a given unit type is the variance of the corresponding β-distribution, and the within-unit variance (noise) for a given unit is simply the variance of the unit's binomial distribution, $p_i(1-p_i)/n_i$. This within-unit variance, and thus reliability measure (1), depend on both the true pressure ulcer rate (which must be estimated) and the number of patients assessed (n_i). Units with a pressure ulcer rate of zero (or one) have zero within-unit variance and thus perfect reliability according to measure (1), regardless of the value of the between-unit variance. For pressure ulcer rates between zero and one, within-unit variance decreases as n_i increases, so all else being equal, units with more patients assessed have higher reliability scores.

Estimation

To estimate the beta-distributions' parameters, we fit the beta-binomial models using the SAS BETABIN macro (Wakeling, 2005) in SAS 9.4. In a typical analysis, we might take units' observed pressure ulcer rates as estimates of their true pressure ulcer rates (the p_is). However, under this frequentist approach, units with an observed ulcer rate of zero have an estimated within-unit variance of zero, implying perfect reliability. This is not only theoretically problematic—surely we cannot conclude based on data from 4 days out of a year that the true probability of a

patient developing a pressure ulcer on these units is exactly zero—but it also makes it impossible in practice to differentiate among these zero-pressure ulcer units or meaningfully assess their measurement reliability.

To avoid these problems, we computed empirical Bayes estimates of the pressure ulcer rates (Gajewski, Mahnken, & Dunton, 2008). In the empirical Bayes framework, the beta-distribution for p_i is a conjugate prior distribution, and the posterior distribution of p_i given the pressure ulcer count is also beta. The posterior mean, taken as the estimate of p_i, is a weighted average of the *i*th unit's observed pressure ulcer rate and the mean of the prior distribution (i.e., the mean rate among all units of that type). This estimator "shrinks" each unit's observed pressure ulcer rate toward the prior mean, yielding non-zero pressure ulcer rate estimates for all units. This shrinkage effect depends on the unit's number of patients assessed; the observed pressure ulcer rate is given less weight for units with fewer patients assessed, resulting in more shrinkage toward the prior mean.

More formally, let y_i be the count of pressure ulcers for the *i*th unit. We let $p_i \sim \text{beta}(\alpha, \beta)$ and $y_i|p_i \sim \text{binomial}(p_i, n_i)$. The empirical Bayes estimator of p_i is the posterior mean of $p_i|y_i$, which can be expressed as follows:

$$[n_i/(n_i+\alpha+\beta)]\{y_i/n_i\} + [(\alpha+\beta)/(n_i+\alpha+\beta)]\{\alpha/(\alpha+\beta)\}$$

where y_i/n_i is the observed pressure ulcer rate and $\alpha/(\alpha+\beta)$ is the prior mean. We computed each unit's pressure ulcer rate estimate using estimates of α and β from the beta-binomial models, then computed its score on reliability measure (1).

Simulation Study 1

We carried out a simulation study to assess how well observed pressure ulcer rates allow us to differentiate among units on the basis of their true pressure ulcer rates. Total pressure ulcer rate and stage II+ pressure ulcer rate data were simulated and analyzed separately. The description that follows is in terms of a single pressure ulcer rate, but all steps were carried out twice, once for each pressure ulcer rate.

The idea of the simulation study was as follows: given known pressure ulcer rates (p_is) we have known within-unit variability (the binomial variance, $p_i[1-p_i]/n_i$) and can simulate multiple sets of pressure ulcer data reflecting this variability. With known pressure ulcer rates, the rank of each unit's true rate among units of its type is also known, and we can easily compute this rank in each simulated data set. By comparing the unit's true rank to its rank across simulated data sets, we can assess the precision with which a given set of observed pressure ulcer rates allows us to rank the unit.

Treating each unit's number of patients assessed as known, and treating its empirical Bayes estimate as its true pressure ulcer rate, we used SAS 9.4 to generate 1,000 random binomial pressure ulcer counts for each unit based on its number of patients assessed. This gave us 1,000 simulated data sets, each the same size as the original data set. For each simulated data set, we fit five beta-binomial models, one per unit type, and used the resulting parameter estimates to compute the empirical Bayes estimate of each unit's true pressure ulcer rate. Units were percentile-ranked by unit type within the original data set and within each simulated data set, based on their estimated pressure ulcer rates.

In each simulated data set, we counted the number of units with percentile rank within five and within ten percentiles of their true percentile rank. The proportions, across units and simulated data sets, of units ranked within five and within ten percentiles of true rank were computed as estimates of the probability of a randomly chosen unit's observed pressure ulcer rate rank falling within five (ten) percentiles of true rank. In addition, we identified units with true pressure ulcer rates in the highest quartile and highest decile, counted how many were correctly classified as such in each simulated data set, and took the proportions (across units and data sets) of correctly classified highest-quartile and highest-decile units as estimates of the probability of a randomly chosen highest-quartile (-decile) unit being identified as such based on its observed pressure ulcer rate.

In addition to computing these global indicators of reliable differentiation, we assessed reliability measure (1) as an indicator of our ability to rank a unit precisely among its peers, as follows. For each unit, we computed the proportion of simulated data sets in which it was ranked within five and within ten percentiles of its true rank, and for each highest-quartile and highest-decile unit we computed the proportion of simulated data sets in which the unit was correctly classified as such. Thus, we obtained an estimate of each unit's probability of being ranked within five (ten) percentiles of its true rank, and an estimate of each highest-decile (-quartile) unit's probability of being identified as such, in a given set of observed data.

We then measured the strength of association between measure (1) and these probability estimates using Spearman rank correlations. A non-parametric alternative to the Pearson correlation coefficient, Spearman's correlation measures monotonic (not just linear) association and is robust against outliers. If signal-noise reliability is an indicator of precision of differentiation, units with higher scores on measure (1) should tend to have higher probabilities of being ranked precisely and, if in the top decile or quartile, of being correctly classified as such, resulting in strong, positive correlations between signal-noise reliability and these probability estimates.

Simulation Study 2

In a second simulation study, we examined the effect of substantially increasing sample size (number of patients

Table 1. Descriptive Statistics for Observed Pressure Ulcer Rates in NDNQI Sample

Ulcer Rate	Unit Type	*n*	Mean ± *SD*	Min	25th Percentile	Median	75th Percentile	Max
Total	Critical care	2118	5.7 ± 5.6	0.0	1.4	4.4	8.3	42.9
	Step-down	1424	2.6 ± 3.4	0.0	0.0	1.7	3.8	28.8
	Medical	1813	2.0 ± 2.6	0.0	0.0	1.3	2.9	50.0
	Surgical	1247	1.5 ± 2.1	0.0	0.0	1.0	2.3	17.1
	Medical-surgical	2197	1.7 ± 2.6	0.0	0.0	1.1	2.4	37.3
Stage II+	Critical care	2118	4.9 ± 5.0	0.0	0.0	3.8	7.4	35.7
	Step-down	1424	2.0 ± 2.9	0.0	0.0	1.1	2.8	24.2
	Medical	1813	1.5 ± 2.2	0.0	0.0	1.1	2.2	50.0
	Surgical	1247	1.2 ± 1.7	0.0	0.0	0.6	1.8	17.1
	Medical-surgical	2197	1.2 ± 1.9	0.0	0.0	0.8	1.8	22.5

Note. *SD*, standard deviation; min, minimum; max, maximum, NDNQI, National Database of Nursing Quality Indicators.

assessed per unit) on our ability to rank units precisely using their observed pressure ulcer rates. It is clear from the binomial variance formula (above) that increasing the number of patients assessed for a given unit results in more precise estimates of the unit's pressure ulcer rate; it is less clear to what extent such an increase improves differentiation among units. We simulated annual pressure ulcer rates based on 52 weekly (rather than four quarterly) pressure ulcer surveys per unit, by multiplying each unit's number of patients assessed by 13 (13 weeks × 4 quarters = 52 weeks) and repeating the data simulation described under Simulation Study 1. We also re-computed each unit's binomial variance using its new count of patients assessed and substituted this value for within-unit variance in expression (1), thereby obtaining reliability scores corresponding to the increased counts of patients assessed. The analyses carried out in Simulation Study 1 were repeated in Simulation Study 2.

Results

Units reported on 674,640 patients assessed during 2013. Of these patients, 16,689 (2.5%) had at least one pressure ulcer, and 13,348 (2.0%) had at least one pressure ulcer stage II or above. Descriptive statistics for observed pressure ulcer rates are shown by unit type in Table 1. Total pressure ulcer rates ranged from 1.5% on surgical units to 5.7% on critical care units. Stage II+ rates ranged from 1.2% on surgical and medical-surgical units to 4.9% on critical care units. Of the 8,799 units in the sample, 2,987 (34%) reported a total pressure ulcer rate of zero, and 3,503 (40%) reported a stage II+ rate of zero.

Means and standard deviations (*SD*s) for reliability scores by unit type are provided in Table 2, along with results of Simulation Study 1. Average scores on reliability measure (1) varied across unit types, from .53 to .67 for the total pressure ulcer rate and from .44 to .63 for the stage II+

Table 2. Reliability of Pressure Ulcer Rate Rankings Based on Four Quarterly Pressure Ulcer Surveys

			Estimated Ranking Probability				Spearman Correlation With Reliability Score			
			Prob(|Observed Rank − True Rank| < δ)		Prob(Correct Assignment)		Prob(|Observed Rank − True Rank| < δ)		Prob(Correct Assignment)	
Ulcer Rate	Unit Type	Reliability Score (Mean ± *SD*)	δ = 5 Percentiles	δ = 10 Percentiles	Highest Quartile	Highest Decile	δ = 5 Percentiles	δ = 10 Percentiles	Highest Quartile	Highest Decile
Total	Critical care	.56 ± .15	.28	.39	.60	.50	−.43	−.20	.10	.43
	Step-down	.67 ± .18	.36	.45	.64	.57	−.12	−.18	−.07	−.02
	Medical	.58 ± .15	.33	.41	.59	.49	−.05	−.11	−.23	.05
	Surgical	.53 ± .16	.38	.46	.56	.45	.09	−.03	−.08	.10
	Medical-surgical	.62 ± .16	.39	.47	.61	.53	.04	−.07	−.34	−.20
Stage II+	Critical care	.54 ± .15	.28	.39	.59	.49	−.42	−.29	.09	.37
	Step-down	.63 ± .19	.38	.47	.62	.53	−.07	−.17	−.10	.16
	Medical	.50 ± .15	.36	.42	.54	.44	.06	−.03	−.13	.08
	Surgical	.44 ± .15	.43	.49	.51	.40	.14	.04	.11	.17
	Medical-surgical	.51 ± .15	.42	.49	.54	.45	.13	.03	−.16	−.01

Note. *SD*, standard deviation; prob, probability.

rate. Individual unit reliability scores varied widely around these means, as indicated by the *SD*s (all .15 or larger).

Estimated probabilities of ranking a randomly chosen unit within five or within ten percentiles of its true pressure ulcer rate rank were comparable for the two pressure ulcer rates, and all were less than .50. Estimated probabilities of correct top-quartile classification were around .60 for the total pressure ulcer rate and slightly lower for the stage II+ rate. Top-decile probability estimates were generally around .10 lower than the corresponding top-quartile estimates.

Spearman correlations between reliability measure (1) and the estimated probability of unit ranking within five or within ten percentiles of true rank did not exceed .14 for either pressure ulcer rate, and over half were negative. Correlations between reliability measure (1) and the estimated probabilities of correct highest-quartile and highest-decile classification were less than .20, with two exceptions (.37 and .43), and half were negative. Thus, overall, higher signal-noise reliability, as defined by measure (1), was not a strong indicator of better differentiation among units.

Not surprisingly, increasing the counts of patients assessed in Simulation Study 2 resulted in dramatic increases in signal-noise reliability scores as well as reductions in their variability among units (see Table 3). Scores averaged .90 or higher for both pressure ulcer rates on all five unit types. There were also increases, some quite large, for both pressure ulcer rates in the estimated probabilities of unit ranking within five and within ten percentiles of true rank. However, these probabilities remained rather low: the former ranged from .31 to .48 and the latter from .50 to .73. Probabilities of correct assignment of highest-quartile and highest-decile units improved substantially, exceeding .80 in most cases. As shown by the Spearman correlations in Table 3, most of which were negative, increasing the counts of patients assessed did not improve the overall performance of reliability measure (1) as an indicator of precision of differentiation.

Discussion

For hospital-acquired pressure ulcer rates computed from four quarterly pressure ulcer surveys, average signal-noise reliability scores ranged across unit types from .53 to .67 for total pressure ulcer rates and from .44 to .63 for stage II+ pressure ulcer rates. Based on a cutoff in the 0.7–0.9 range, these scores were unacceptably low. Increasing counts of patients assessed (sample sizes) to reflect 52 weekly pressure ulcer surveys raised these average reliability scores (though not each individual unit's score) to .90 or higher and reduced their variability. Similarly, simulation-based estimates of precision of ranking and classification indicated poor differentiation with four quarterly surveys and better differentiation with 52 weekly surveys.

What the signal-noise scores mean is not entirely clear. They have no straightforward interpretation in terms of how well we can differentiate among providers or time periods, and thresholds for acceptable reliability vary (Adams, 2009; Adams et al., 2010; Centers for Medicare & Medicaid Services, 2012). Even if there were a consensus threshold, researchers would need to decide for their specific contexts whether the threshold only needs to be met on average, or should be applied to individual provider or unit reliabilities.

This analysis made clear that for the hospital-acquired pressure ulcer rates we analyzed, the signal-noise measure was a poor indicator of the precision with which we can differentiate among nursing units. Units with higher reliability scores did not have higher probability of being

Table 3. Reliability of Pressure Ulcer Rate Rankings Based on 52 Weekly Pressure Ulcer Surveys

			Estimated Ranking Probability				Spearman Correlation With Reliability Score			
			Prob(|Observed Rank − True Rank| <δ)		Prob(Correct Assignment)		Prob(|Observed Rank − True Rank| <δ)		Prob(Correct Assignment)	
Ulcer Rate	Unit Type	Reliability Score (Mean ± *SD*)	δ = 5 Percentiles	δ = 10 Percentiles	Highest Quartile	Highest Decile	δ = 5 Percentiles	δ = 10 Percentiles	Highest Quartile	Highest Decile
Total	Critical care	.93 ± .04	.43	.67	.84	.82	.25	.26	−.12	.15
	Step-down	.95 ± .04	.48	.73	.88	.85	−.09	−.08	−.26	−.25
	Medical	.94 ± .05	.40	.65	.85	.81	−.13	−.15	−.38	−.22
	Surgical	.92 ± .05	.35	.58	.82	.78	−.19	−.20	−.35	−.21
	Medical-surgical	.95 ± .05	.40	.64	.85	.83	−.29	−.29	−.47	−.38
Stage II+	Critical care	.93 ± .05	.41	.64	.84	.81	.22	.24	−.10	.10
	Step-down	.94 ± .05	.43	.67	.86	.82	−.16	−.18	−.24	−.12
	Medical	.92 ± .05	.34	.57	.81	.79	−.17	−.17	−.35	−.23
	Surgical	.90 ± .06	.31	.50	.78	.74	−.08	−.16	−.29	−.19
	Medical-surgical	.92 ± .06	.35	.55	.81	.80	−.15	−.27	−.40	−.35

Note. *SD*, standard deviation; prob, probability.

ranked within five or within ten percentiles of their true rank, nor did highest-quartile or highest-decile units with higher reliability scores generally have a higher probability of being correctly identified as such. This held in both of the simulation studies, despite the second study involving much larger counts of patients assessed.

In assessing reliability in terms of ranked-based differentiation among providers, be they hospitals, nursing units, physicians, or other entities, it may not be enough to consider a single provider's signal-noise reliability in isolation. A particular unit's measurements may be highly reliable, but the unit's rank may be unstable because of low reliability for other providers' measurements. As a hypothetical example, suppose our measurements for one unit are perfectly reliable and that this unit's true rank is the 25th percentile. Although any measurement for this unit will be equal to the true quantity of interest (the definition of perfect reliability), the measurements of other units in the ranking will deviate from their target quantities, and as a result the observed rank of our 25th-percentile unit may take some other value. In statistical terms, the issue is independence: a unit's reliability score may be independent of other units' scores, but its rank in a given data set is not independent of other units' ranks.

Furthermore, as demonstrated in Simulation Study 2, it is not always enough for a set of units to have high average signal-noise reliability scores if the goal is precise differentiation among units. Even with unit reliabilities averaging .90 or higher, the probability of a randomly chosen unit being ranked within five percentiles of its true rank was less than one-half, and probabilities of ranking within ten percentiles of true rank were generally no higher than two-thirds. Moreover, the estimated probability of failing to identify a randomly chosen unit with a pressure ulcer rate in the highest decile was as high as .26.

If reliability is defined in terms of our ability to rank units precisely, then the results of this study demonstrate that signal-noise measures can perform poorly as measures of reliability. Staggs and Gajewski (2015) issued a similar caution based on their findings in a study of inpatient fall rates. It should also be noted that reliability scores themselves are subject to error from the parameter estimates on which they are based. Staggs and Gajewski used Bayesian methods to estimate precision of reliability score estimates, but one could also use the simulation approach described here.

Therefore, given its limitations, we would caution against over-reliance on the traditional signal-noise reliability measure in contexts involving rank-based differentiations among providers. If the goal is precision of ranking or accurate classification of poor performers, a simulation approach like the one demonstrated in this study, or the Bayesian methods described by Staggs and Gajewski (2015), are better alternatives.

We would also caution hospital administrators and staff against taking a handful of quarterly pressure ulcer rates based on 1-day prevalence surveys too seriously as a measure of quality of care, especially for units that do not have high patient volume. In the absence of adequate risk adjustment, there seems to be too much noise in observed rates based on such limited sample sizes to draw reliable conclusions about the underlying true rate, much less about the extent to which these observed rates are attributable to quality. A year's worth of data from weekly or daily surveys would be much more valuable. This same caveat applies to comparisons based on ranked pressure ulcer rates, which are less reliable than the pressure ulcer rates themselves.

Although the importance of adequate sample size for reliable measurement is well-known, in practice it can be hard to achieve. Measures that are used for public reporting or quality improvement purposes are limited both by the available patient population, which typically cannot be increased as a matter of practice, and by the burden associated with collecting data too frequently. In order to balance the need for reliable data with the costs and availability of resources for collecting data, it may be important to investigate not only the reliability of quality measures but also how much data are needed to establish reliable measurement.

With over one-third of units in this study reporting total pressure ulcer rates of zero, empirical Bayes estimation of pressure ulcer rates was critical. Under a frequentist approach (taking the observed rates as estimates of the true rates), these units would have perfect signal-noise reliability scores, resulting in artificially high unit type averages. In addition, the simulation study would be unworkable, as these units would have binomial variance of zero and thus zero pressure ulcer counts in every simulated data set. In this context, the empirical Bayes approach is both more realistic and more practical. A fully Bayesian approach would share these merits.

It is important to note that this was not a study of the validity of pressure ulcer rates as measures of quality (i.e., the extent to which they measure quality as opposed to other phenomena), nor an evaluation of the pressure ulcer assessments used by the NDNQI. Validity, potential bias in observed pressure ulcer counts, and related issues are beyond the scope of this study.

It is also worth noting that the pressure ulcer rates we studied are not risk-adjusted (beyond stratification by unit type), and within-unit variability in patient mix is a factor contributing to noise in observed measurements. An effectively risk-adjusted measure would presumably be less subject to random error and thus more reliable. Controlling for other sources of noise, such as inter-rater differences in pressure ulcer identification and staging, would also potentially improve the measure's reliability.

Pressure ulcer rates observed in other samples of nursing units would be subject to the same sources of randomness, although the amount of error variance would differ across units and unit types, as it did in this study.

Precision of differentiation in any sample depends on the between- and within-unit variability in the pressure ulcer rates and numbers of patients assessed, and analyses of different samples might yield somewhat different results. However, the range of unit types and observed pressure ulcer rates in our study gives us some confidence in the generalizability of our broad conclusions.

Conclusion

The standard signal-noise reliability measure was a poor indicator of precision of rank-based differentiation of pressure ulcer rates among units. Alternative methods of assessing reliability need to be applied to measures purported to differentiate among units on the basis of quality of care, to ensure that they allow for precise differentiation based on true differences.

References

Adams, J. L. (2009). *The reliability of provider profiling: A tutorial*. Santa Monica, CA: RAND Corporation.

Adams, J. L., Mehrotra, A., Thomas, J. W., & McGlynn, E. A. (2010). Physician cost profiling—reliability and risk of misclassification. *New England Journal of Medicine*, *362*, 1014–1021. doi: 10.1056/NEJMsa0906323

Bergquist-Beringer, S., Gajewski, B., Dunton, N., & Klaus, S. (2011). The reliability of the National Database of Nursing Quality Indicators pressure ulcer indicator: A triangulation approach. *Journal of Nursing Care Quality*, *26*, 292–301. doi: 10.1097/NCQ.0b013e3182169452

Centers for Medicare & Medicaid Services. (2012). *Memorandum: Results of reliability analysis from Mathematica Policy Research*. Baltimore, MD: US Department of Health & Human Services. Retrieved from http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/Download/HVBP_Measure_Reliability-.pdf

Eijkenaar, F., & van Vliet, R. C. J. A. (2014). Performance profiling in primary care: Does the choice of statistical model matter? *Medical Decision Making*, *34*, 192–205. doi: 10.1177/0272989X13498825

Gajewski, B. J., Mahnken, J. D., & Dunton, N. (2008). Improving quality indicator report cards through Bayesian modeling. *BMC Medical Research Methodology*, *8*(1), 1. doi: 10.1186/1471-2288-8-77

Hofer, T. P., Hayward, R. A., Greenfield, S., Wagner, E. H., Kaplan, S. H., & Manning, W. G. (1999). The unreliability of individual physician report cards for assessing the costs and quality of care of a chronic disease. *JAMA*, *281*, 2098–2105.

Kao, L. S., Ghaferi, A. A., Ko, C. Y., & Dimick, J. B. (2011). Reliability of superficial surgical site infections as a hospital quality measure. *Journal of the American College of Surgeons*, *213*, 231–235. doi: 10.1016/j.jamcollsurg.2011.04.004

Kottner, J., & Dassen, T. (2010). Pressure ulcer risk assessment in critical care: Interrater reliability and validity studies of the Braden and Waterlow scales and subjective ratings in two intensive care units. *International Journal of Nursing Studies*, *47*, 671–677. doi: 10.1016/j.ijnurstu.2009.11.005

Kottner, J., Halfens, R., & Dassen, T. (2009). An interrater reliability study of the assessment of pressure ulcer risk using the Braden scale and the classification of pressure ulcers in a home care setting. *International Journal of Nursing Studies*, *46*, 1307–1312. doi: 10.1016/j.ijnurstu.2009.03.014

Kottner, J., Raeder, K., Halfens, R., & Dassen, T. (2009). A systematic review of interrater reliability of pressure ulcer classification systems. *Journal of Clinical Nursing*, *18*, 315–336. doi: 10.1111/j.1365–2702.2008.02569.x

National Pressure Ulcer Advisory Panel & European Pressure Ulcer Advisory Panel. (2009). *Prevention and treatment of pressure ulcers: Clinical practice guidelines*. Washington DC: National Pressure Ulcer Advisory Panel.

Staggs, V. S., & Gajewski, B. J. (2015). Bayesian and frequentist approaches to assessing reliability and precision of health-care provider quality measures. *Statistical Methods in Medical Research* [advance online publication]. doi: 10.1177/0962280215577410

Wakeling, I. (2005). *MACRO BETABIN Version 2.2. Qi Statistics*. Retrieved from http://www.qistats.co.uk/BetaBinomial.html

Waugh, S. M., & Bergquist-Beringer, S. (2016). Inter-rater agreement of pressure ulcer risk and prevention measures in the National Database of Nursing Quality Indicators® (NDNQI). *Research in Nursing & Health 39*, 164–174. doi: 10.1002/nur.21717

Acknowledgments

Partial support for this work was provided by the American Nurses Association and by Press Ganey Associates through the National Database of Nursing Quality Indicators.