Misclassification Risk of Tier-Based Physician Quality Performance Systems

John L. Adams and Susan M. Paddock

Objective. There is increasing interest in identifying high-quality physicians, such as whether physicians perform above or below a threshold level. To evaluate whether current methods accurately distinguish above- versus below-threshold physicians, we estimate misclassification rates for two-category identification systems.

Data Sources. Claims data for Medicare fee-for-service beneficiaries residing in Florida or New York in 2010.

Study Design. Estimate colorectal cancer, glaucoma, and diabetes quality scores for 23,085 physicians. Use a beta-binomial model to estimate physician score reliabilities. Compute the proportion of physicians whose performance tier would be misclassified under three scoring systems.

Principal Findings. In the three scoring systems, misclassification ranges were 8.6–25.7 percent, 6.4–22.8 percent, and 4.5–21.7%. True positive rate ranges were 72.9–97.0 percent, 83.4–100.0 percent, and 34.7–88.2 percent. True negative rate ranges were 68.5–91.6 percent, 10.5–92.4 percent, and 81.1–99.9 percent. Positive predictive value ranges were 70.5–91.6 percent, 77.0–97.3 percent, and 55.2–99.1 percent.

Conclusions. Current methods for profiling physicians on quality may produce misleading results, as the number of eligible events is typically small. Misclassification is a policy-relevant measure of the potential impact of tiering on providers, payers, and patients. Quantifying misclassification rates should inform the construction of high-performance networks and quality improvement initiatives.

Key Words. High-performance networks, reliability, misclassification, provider profiling, quality of care

Identifying high-performing physicians is central to many health care quality initiatives. Physician quality is typically assessed by summarizing patient outcomes or the proportion of times recommended clinical processes are delivered by a physician to his/her patients. These performance estimates are often compared against a threshold target to determine which physicians meet or exceed the target, thereby classifying physician performance as either above threshold or below threshold.

One arena where such classifications are routine is health plans' development of high-performance networks (HPNs) of physicians. The physicians participating in an HPN are selected for their relatively strong performance with respect to lower cost and/or higher quality. Patients' access to physicians outside the network is typically disincentivized by factors such as increased copayments (Sinaiko and Rosenthal 2010). HPNs were initially designed to identify providers with relatively low cost (Summer 2014), but they are increasingly targeting quality as well. One example from the commercial health insurance sector is the Cigna Care network, which includes physicians in its HPN if they are in the top 34 percent on quality—or, if cost data are available, the top 40 percent on quality and top 40 percent on cost-efficiency (Cigna Health Corporation 2016).

Physicians who treat Medicare beneficiaries are increasingly being classified into performance groups, such as through Medicare Advantage plans with HPNs (Biles 2014). Many physicians who treat patients covered by traditional Medicare will soon be classified as above (or below) various performance thresholds under the Merit-based Incentive Payment System. Receipt of incentive payments will depend on their classifications (Centers for Medicare and Medicaid Services 2015); physicians and other eligible professionals with below-threshold performance on the composite performance measure will have their payments reduced while those labeled as above threshold will receive neutral or positive payment adjustments.

One challenge to these efforts is that individual physicians can be one of the most problematic groups to profile given their smaller patient populations relative to other larger profiling targets and the reductions in the number of patients eligible for any given measure due to patients switching providers or health plans. Furthermore, summarizing the care provided by physicians often relies not only on attributing patient index visits but also all related visits (Mehrotra et al. 2010). Using data from commercial plans, several authors have found low reliabilities for measuring some aspects of physician quality (Hofer et al. 1999; Scholle et al. 2008; Sequist et al. 2011; Smith et al. 2013) and cost profiles (Hofer et al. 1999; Adams et al. 2010). Reliability is a key metric for evaluating whether a measure is suitable for profiling because it indicates the proportion of signal versus noise contained in a physician's quality profile. The signal in this case is the proportion of the variability in

measured performance that can be explained by real differences in provider performance. A reliability level of zero can be interpreted as meaning that all the variability in the profiles is attributable to measurement or sampling error. A reliability level of 1 can be interpreted as meaning that all the variability is attributable to real differences in performance.

While users appreciate the importance of reliability as a relative measure, few have a good sense of how much better a value of 0.8 is than a value of 0.7 or whether either value is “good enough” for a particular purpose. In fact, the notion of what constitutes “good enough” might vary depending on the objective—for example, whether it is to classify physicians into three bins (e.g., providers who are significantly above or below average or not significantly different from average) (Paddock, Adams, and Hoces de la Guardia 2014) or two bins (e.g., top performers vs. others), along with how top performance is to be defined (e.g., top 10 percent vs. top 25 percent) (Paddock and Louis 2011; Paddock 2014). The purpose of the profiling can guide the choice of a loss function to evaluate the usefulness of a quality measure in absolute terms (Lin et al. 2006). Following this logic, Normand et al. (2007) compute the minimum numbers of patients needed to classify hospitals into performance tiers with high levels of sensitivity for hospital profiling of acute myocardial infarction care. In the context of developing physician cost profiles, Adams et al. (2010) computed rates of misclassifying physicians as having costs in the lowest 25 percent versus not for a combined dataset of four large health plans in Massachusetts. They found provider cost profiles had reliabilities less than 0.7 in 59 percent of cases and physician performance was misclassified 22 percent of the time, thereby supplementing the reliability information with an absolute and policy-relevant metric that raised concerns about characterizing physician costs in this way.

An often overlooked reliability issue is the effect that lower reliability can have on estimating the cut points used for classifying physician performance. If means or percentiles are used as cut points, they must be estimated from the physician score distribution. Lower reliability estimates of physician scores will have an effect on the precision of the estimates of these cut points. A physician may be misclassified by either an unreliable score or an imprecisely estimated cut point.

There is the potential to miss gains from recognizing and encouraging high-quality care when provider performance is misclassified, such as including physicians in an HPN who are erroneously classified as above a performance threshold while excluding those performers with true

strong performance who are incorrectly classified below the threshold (Kassirer 1994; Werner and Asch 2005). Despite these risks, misclassification rates for quality measures at the physician level have yet to be assessed. We thus examine rates of misclassification of provider performance in this paper, highlighting the importance of evaluating misclassification risk given the context of identifying high-performing physicians. We illustrate using three HEDIS[®] process-of-care measures applied to Medicare fee-for-service (FFS) beneficiaries from the states of Florida and New York.

METHODS

Study Population and Measures

The study population is composed of primary care physicians with at least one attributed quality measure who served Medicare FFS beneficiaries residing in New York or Florida in 2010. Beneficiary care was attributed to physicians according to the majority of the payments for evaluation and management visits (Mehrotra et al. 2010). We used Medicare enrollment, inpatient, outpatient, and carrier FFS claims files from 2010 to calculate HEDIS measure scores for 23,085 primary care physicians (11,374 in Florida and 11,711 in New York) providing care to FFS beneficiaries eligible for any of three care measures: colorectal cancer screening (502,628 beneficiaries, with 284,503 in Florida and 218,125 in New York), glaucoma screening (921,894 beneficiaries, with 509,885 in Florida and 412,009 in New York), or a composite constructed from four diabetes measures (71,831 beneficiaries, with 42,030 in Florida and 29,801 in New York). The colorectal cancer screening and glaucoma screening measures apply to a large proportion of FFS beneficiaries; the HEDIS specifications stipulate the measures that apply to those aged 67 or older for glaucoma and aged 50–75 for colorectal cancer. For each physician, the pass rate is computed as the number of screenings among those eligible for screening divided by the denominator, which is the number of patients eligible for the particular screening. The diabetes composite reflects opportunities for passing the component indicators, and it was constructed for each physician as the sum of the number of passes on the diabetes measures (Hb-A1c screening, eye test, LDL test, nephropathy screening) divided by the sum of the number of opportunities to pass each measure for eligible patients (Reeves et al. 2007).

Calculation of Reliability

Both misclassification rates and reliabilities are calculated by fitting the appropriate hierarchical model. In many health care provider profiling settings, a hierarchical linear model (HLM) is used, assuming normal errors and normal mixing distributions. Given the true provider performance rate is between 0 and 1, we will instead use the beta-binomial model, which is a more natural fit for estimating the reliability of provider performance based on simple pass/fail rate measures such as colorectal cancer screening, glaucoma screening, and most other HEDIS measures. Applying the standard HLM to reliability calculations in this context would require an assumption of asymptotic normality of physician-level score estimates, which might not be tenable for physician profiling when the number of opportunities for physicians to pass quality indicators is small.

Our approach to computing reliabilities of provider performance estimates follows (Adams 2009). We assume that each physician has a true pass rate, p . This true pass rate could vary from physician to physician if there is variation in how physicians deliver the care indicated by the quality measure. The observed pass rate, $\hat{p}$, will vary because the number of events is small and subject to random variation around the true rates. Under the beta-binomial model, the physician's score is modeled as a binomial random variable conditional on the physician's true value:

$$x_i | p_i \sim \text{Binomial}(n_i, p_i),$$

and a physician's true performance is assumed to come from a beta distribution:

$$p_i | \alpha, \beta \sim \text{Beta}(\alpha, \beta),$$

where α and β parameters characterize the distribution of performance among all providers in the population, such that the mean and variance of p_i are μ and $\sigma_{\text{provider}}^2$:

$$\mu = \frac{\alpha}{(\alpha + \beta)}$$

$$\sigma_{\text{provider}}^2 = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}$$

Figure 1 illustrates the flexibility of the beta distribution for modeling pass rates from 0 to 1 for various values of α and β . The distribution can be symmetric or skewed left or right to represent populations of physicians with

Figure 1: Examples of Beta Distributions

low or high averages, or even U-shaped to represent populations that nearly separate into high and low subpopulations. The beta distribution has the advantage of great flexibility on the unit interval with a parsimonious two parameters.

The reliability of the estimate of physician i 's pass rate is estimated as:

$$\text{reliability} = \frac{\sigma_{\text{provider-to-provider}}^2}{\sigma_{\text{provider-to-provider}}^2 + \frac{\hat{p}_i \times (1 - \hat{p}_i)}{n_i}}$$

The variance of physician i 's pass rate is the right-hand quantity in the denominator of the reliability calculation. It is estimated as the usual binomial variance for the error, where $\hat{p}_i$ is the observed pass rate for physician i based on n_i opportunities. As the number of opportunities a physician has to pass the quality indicator and the estimated pass rate vary across physicians, each physician's quality profile has a different reliability.

We will explore three scoring systems based on a high-performance or two-category network (Draper, Liebhaber, and Ginsburg 2007). The first system will be a simple assignment of a provider to above or below the mean.

The second system will identify a narrow HPN by identifying providers with performance above the 75th percentile. The third system will define a broad network with exclusions of providers with performance at or below the 25th percentile. For each system, we will create two sets of thresholds that will define the two performance categories. The first will be the simple mean or the simple percentile of the observed pass rates for providers. The second will be reliability-adjusted thresholds that are estimated using the beta-binomial model.

Calculation of Misclassification

Fundamentally, reliability is the measure of whether it can be determined that a physician is different from his or her peers. One concern is that, for most users, interpreting reliability is not intuitive. Additionally, there is no agreement on a gold standard level of reliability for quality profiling. We will illustrate the relationship between misclassification and reliability of physician quality measures using the example scoring systems. In this illustrative application, the physicians who are above the cut point are labeled “above threshold” and the remaining labeled “below threshold.” In such a categorization system, there are two types of misclassification errors: (1) flagging an above-threshold physician as being below the threshold (equivalent to a false negative) and (2) flagging a below-threshold physician as above threshold (equivalent to a false positive). We will also examine the overall misclassification rate, which is the population-weighted sum of these two error rates. Positive predictive values and negative predictive values are other metrics we present. Positive predictive value characterizes the probability that a physician flagged as above threshold actually has above-threshold performance. Positive predictive value may be the most relevant summary from a plan enrollee’s point of view in terms of trusting whether the physicians included in an HPN are actually better-performing than others (Sinaiko and Rosenthal 2010). To estimate the misclassification probabilities for primary care physicians, we performed the following steps:

1. Estimate the mean or percentile of the observed quality score from the data.
2. Estimate the physician-to-physician distribution using the beta-binomial model.
3. Determine where the mean or percentile of the observed quality score falls in the estimated beta scale.

- 4. For each physician, calculate the misclassification probability under the beta distribution assumption.
- 5. Sum the misclassification probabilities.

We will also examine whether misclassification risk varies according to the choice of using a simple mean or percentile versus a reliability-adjusted cut point. We will also explore the possibility of reducing misclassification with sample size minima, albeit for a reduced fraction of the providers profiled.

RESULTS

Table 1 presents summary statistics of the measure pass rates and number of attributed denominator events for the physician study population. Note that even for these relatively high prevalence measures that apply to a broad subset of the Medicare FFS beneficiary population, many physicians will not meet minimum sample size requirements of 11 or 30 assigned measures that are used in some quality profiling efforts (National Committee for Quality

Table 1: Summary Statistics of Physician Denominators and Pass Rates for CRC Screening, GSO, and the Diabetes Composite Measure

<i>State Measure</i>	<i>Florida</i>			<i>New York</i>		
	<i>CRC</i>	<i>GSO</i>	<i>Diabetes</i>	<i>CRC</i>	<i>GSO</i>	<i>Diabetes</i>
Number of providers	9,889	10,225	7,388	11,097	11,374	7,687
Denominator						
Maximum	362	602	312	281	612	260
90%	73	133	52	49	95	32
75%	40	69	28	27	50	20
50%	16	25	16	12	19	12
25%	5	6	8	4	5	4
10%	2	2	4	1	2	4
Minimum	1	1	4	1	1	4
Measure pass rate						
Maximum	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
90%	50.0%	75.6%	100.0%	50.0%	75.0%	100.0%
75%	40.3%	66.3%	90.6%	40.0%	64.3%	93.8%
50%	29.7%	54.5%	78.6%	27.3%	50.0%	78.6%
25%	16.7%	40.0%	71.9%	13.8%	33.3%	72.5%
10%	0.0%	16.7%	58.3%	0.0%	4.1%	50.0%
Minimum	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

CRC, colorectal cancer; GSO, glaucoma screening.

Assurance 2014). The diabetes composite applies to a smaller subset of the population but is based on four indicators per beneficiary.

Table 2 presents the beta-binomial parameter estimates for α , β , μ , and $\sigma^2_{\text{provider}}$. The average reliability-adjusted pass rate is lowest for colorectal cancer (0.32 in Florida and 0.30 in New York), near 0.50 in both states for glaucoma screening, and highest for the diabetes composite (0.79 in both states), as indicated by the μ estimates. The smallest provider-to-provider variance in the pass rates is for colorectal cancer, as indicated by comparing the estimates of $\sigma^2_{\text{provider}}$ across measures; for a given denominator size and $\hat{p}_i$, the measurement reliability will be highest for glaucoma screening and lowest for colorectal cancer screening. Figure 2 presents the estimated beta distributions based on the parameter estimates of Table 2. The distributions across states are similar within measure. The thin set of vertical lines show the cut points based on estimated means or percentiles from a simple calculation on physician-level score (e.g., number of passes/number of eligible events). The thick set of vertical lines in Figure 2 show the model-based cut points, estimated as μ or percentiles under the beta-binomial model. The latter cut points are adjusted for varying provider reliabilities, thereby giving more weight to providers with

Table 2: Beta-Binomial Parameter Estimates for α , β , and μ , and the Resulting Provider-to-Provider Variance, $\sigma^2_{\text{provider}}$

State	Measure	Parameter	Estimate	Standard Error	t-value	Pr> t	$\sigma^2_{\text{provider}}$
FL	CRC	α	10.18	0.361	28.21	<.0001	0.0066
		β	21.68	0.757	28.65	<.0001	
		μ	0.32	0.001	231.94	<.0001	
FL	GSO	α	10.27	0.257	40.01	<.0001	0.0126
		β	8.38	0.199	42.17	<.0001	
		μ	0.55	0.002	350.57	<.0001	
FL	Diabetes	α	11.38	0.382	29.77	<.0001	0.0108
		β	3.07	0.101	30.58	<.0001	
		μ	0.79	0.002	459.77	<.0001	
NY	CRC	α	8.91	0.333	26.78	<.0001	0.0070
		β	20.47	0.757	27.03	<.0001	
		μ	0.30	0.001	213.34	<.0001	
NY	GSO	α	5.71	0.125	45.85	<.0001	0.0205
		β	5.48	0.114	48.01	<.0001	
		μ	0.51	0.002	283.61	<.0001	
NY	Diabetes	α	8.76	0.31	27.97	<.0001	0.0139
		β	2.39	0.08	29.19	<.0001	
		μ	0.79	0.00	406.65	<.0001	

CRC, colorectal cancer; GSO, glaucoma screening.

Figure 2: Distribution of Physician Performance Derived from the Beta-Binomial Models of Table 2 for Colorectal Cancer (CRC), Glaucoma Screening (GSO), and Diabetes Measures (DIABETES) for Florida and New York

Note. Vertical lines represent the mean, 25th, and 75th percentile cut points under the mean-based high-performing, broad network with exclusions, or narrow high-performing network scoring systems, respectively.

higher reliabilities. For colorectal cancer and glaucoma screening, the beta-binomial model-based estimate for the mean is higher than the empirical estimate, though the model and empirical means are almost identical for the diabetes measures. This would not necessarily be true in other settings or for other measures. In most cases, the 25th and 75th empirical percentiles are farther from the mean than the model-based percentile estimates. These graphs illustrate that accounting for reliability in cut point estimation can yield different results.

Table 3 presents the misclassification rates for the six state-measure combinations for a mean-based high-performance system (Table 3A), broad

network with exclusions of providers below the 25th percentile of performance (Table 3B), and the narrow HPN of providers above the 75th percentile of performance (Table 3C). The true positive rate is the probability that a physician who is actually above the mean or percentile will be classified as above threshold. The true negative rate is the probability that a physician who is actually below the mean or percentile will be classified as below threshold. The misclassification rate combines these two error types weighted by the true proportions. The positive and negative predictive values are the probabilities of getting an actual above-threshold or below-threshold physician when selecting physicians labeled as above threshold or below threshold, respectively. The empirical cut points section of Table 3 presents the results for an HPN as it would be typically implemented using the observed means or percentiles of the provider scores as the cut point. The reliability-adjusted cut points section shows the mean or percentiles from the beta-binomial model as the cut points. The general reduction in misclassification rates for the model-based cut point analysis relative to the empirical cut point analysis illustrates the influence of imprecisely estimated cut points on misclassification. However, the performance in the model-based cut points system in the table is overly optimistic as the model-based cut points are assumed to be true in the misclassification calculations.

A 50 percent misclassification rate is equivalent to a coin toss and reflects a no-information tiering scenario. All misclassification rates in Table 3 are less than 50 percent, with a range of 4.5 percent to 25.7 percent, indicating there is information provided for each state-measure-denominator restriction combination. The denominator restrictions column in Table 3 breaks out alternative misclassification calculations restricting to the providers with sample sizes above various minima. As expected, misclassification improves by discarding the smaller sample size physicians. However, this improvement comes at a cost of reducing the fraction of the physicians that can be scored (percent of physicians profiled column). Even with the relatively low standard of denominator sizes of 11 or more, the denominator restriction can exclude as many as 50 percent of the physicians from tiering. The implications of the misclassification rates for designing HPN can be illustrated using results from the Florida data analysis in Table 3B. If the network designer's goal was to classify providers on diabetes care as above threshold for a broad network with exclusions only if the misclassification rate was less than 10 percent, Table 3B shows the minimum denominator size required would be 30, and the median reliability of providers in such a system would be 0.775. However, other measures may have different minimum denominator size requirements.

Table 3: Misclassification Rates for the Six State-Measure Combinations and Three Scoring Systems

State	Measure	Denominator Restriction	Percent of Physicians Profiled	Median Reliability	Empirical Cut Points					Reliability-Adjusted Cut Points				
					True Positive Rate	True Negative Rate	Misclassification Rate	Positive Predictive Value	Negative Predictive Value	True Positive Rate	True Negative Rate	Misclassification Rate	Positive Predictive Value	Negative Predictive Value
(A) Mean-based high-performance scoring system														
FL	CRC	11	59.9%	0.538	81.8%	68.5%	25.1%	70.5%	80.4%	74.8%	76.8%	24.1%	74.8%	76.8%
FL	CRC	30	33.8%	0.637	86.8%	70.7%	21.6%	73.2%	85.3%	79.2%	80.3%	20.2%	78.8%	80.7%
FL	CRC	100	4.9%	0.790	93.0%	74.7%	16.5%	77.2%	92.1%	85.3%	85.9%	14.4%	84.8%	86.4%
FL	GSO	11	67.2%	0.728	90.0%	70.6%	19.5%	76.0%	87.3%	82.3%	82.3%	17.9%	82.7%	81.6%
FL	GSO	30	46.3%	0.801	93.2%	73.1%	16.7%	78.1%	91.3%	85.5%	85.4%	14.6%	85.8%	85.1%
FL	GSO	100	16.3%	0.891	97.0%	75.4%	13.7%	80.3%	96.0%	89.5%	89.6%	10.5%	89.9%	89.2%
FL	DIAB	11	63.5%	0.656	86.9%	80.5%	16.0%	84.0%	84.0%	86.1%	81.6%	16.0%	84.6%	83.3%
FL	DIAB	30	24.0%	0.775	89.6%	84.1%	12.9%	86.9%	87.3%	88.9%	84.9%	12.9%	87.4%	86.7%
FL	DIAB	100	2.0%	0.890	93.0%	88.2%	9.2%	90.2%	91.4%	92.3%	89.0%	9.2%	90.8%	90.8%
NY	CRC	11	52.4%	0.498	78.6%	70.1%	25.7%	72.0%	77.1%	72.9%	76.4%	25.3%	75.1%	74.3%
NY	CRC	30	22.7%	0.616	84.1%	73.4%	21.4%	75.5%	82.5%	77.1%	81.2%	20.8%	80.0%	78.4%
NY	CRC	100	1.1%	0.800	91.5%	78.2%	15.2%	80.4%	90.4%	84.1%	87.5%	14.2%	86.8%	85.0%
NY	GSO	11	62.4%	0.786	89.6%	77.8%	16.3%	80.3%	88.0%	84.3%	84.6%	15.5%	84.7%	84.2%
NY	GSO	30	39.1%	0.850	93.0%	79.9%	13.5%	82.4%	91.9%	87.7%	87.5%	12.4%	87.6%	87.5%
NY	GSO	100	9.1%	0.923	96.5%	83.3%	10.1%	85.4%	95.9%	91.3%	91.6%	8.6%	91.6%	91.2%
NY	DIAB	11	50.7%	0.685	84.7%	75.7%	19.4%	80.6%	80.7%	83.6%	77.1%	19.4%	81.3%	79.8%
NY	DIAB	30	12.1%	0.804	89.5%	81.3%	14.3%	85.0%	86.6%	87.4%	83.9%	14.2%	86.6%	84.9%
NY	DIAB	100	0.5%	0.916	94.1%	87.2%	9.1%	89.7%	92.5%	93.0%	88.8%	8.9%	90.8%	91.4%
(B) Broad network with exclusions scoring system														
FL	CRC	11	59.9%	0.538	96.0%	27.6%	21.1%	79.9%	69.7%	89.9%	72.2%	14.5%	90.7%	70.5%
FL	CRC	30	33.8%	0.637	98.8%	22.5%	20.3%	79.3%	85.8%	84.6%	75.6%	17.6%	91.2%	62.1%
FL	CRC	100	4.9%	0.790	99.9%	15.0%	21.3%	77.9%	98.8%	89.8%	81.8%	12.2%	93.7%	72.8%
FL	GSO	11	67.2%	0.728	96.0%	48.0%	16.0%	84.7%	80.2%	86.8%	78.6%	15.3%	92.4%	66.5%

Continued

Table 3. Continued

State	Measure	Denominator Restriction	Percent of Physicians Profiled	Median Reliability	Empirical Cut Points						Reliability-Adjusted Cut Points					
					True Positive Rate	True Negative Rate	Misclassification Rate	Positive Predictive Value	Negative Predictive Value	True Positive Rate	True Negative Rate	Misclassification Rate	Positive Predictive Value	Negative Predictive Value	True Positive Rate	True Negative Rate
FL	GSO	30	46.3%	0.801	98.3%	46.9%	14.6%	84.7%	90.0%	89.8%	81.9%	12.2%	93.7%	72.8%		
FL	GSO	100	16.3%	0.891	99.7%	44.3%	14.1%	84.4%	97.9%	93.0%	86.6%	8.6%	95.4%	80.6%		
FL	DIAB	11	63.5%	0.656	90.7%	79.6%	12.1%	93.0%	74.0%	89.9%	81.0%	12.3%	93.4%	72.8%		
FL	DIAB	30	24.0%	0.775	92.9%	82.9%	9.6%	94.2%	79.6%	92.1%	84.6%	9.8%	94.7%	78.1%		
FL	DIAB	100	2.0%	0.890	94.6%	85.7%	7.6%	95.2%	84.1%	93.7%	87.6%	7.8%	95.8%	82.2%		
NY	CRC	11	52.4%	0.498	96.6%	24.0%	21.5%	79.2%	70.3%	79.0%	71.6%	22.8%	89.3%	53.2%		
NY	CRC	30	22.7%	0.616	99.1%	18.3%	21.1%	78.5%	87.3%	83.4%	76.1%	18.4%	91.3%	60.5%		
NY	CRC	100	1.1%	0.800	100.0%	10.5%	22.4%	77.0%	99.4%	89.5%	83.0%	12.1%	94.0%	72.5%		
NY	GSO	11	62.4%	0.786	97.4%	49.3%	14.6%	85.3%	86.3%	89.6%	79.3%	13.0%	92.8%	71.8%		
NY	GSO	30	39.1%	0.850	98.9%	77.8%	6.4%	93.0%	95.9%	92.1%	82.8%	10.2%	94.2%	77.7%		
NY	GSO	100	9.1%	0.923	99.8%	48.5%	13.0%	85.3%	98.9%	94.9%	87.4%	7.0%	95.8%	85.2%		
NY	DIAB	11	50.7%	0.685	85.5%	77.8%	16.4%	92.0%	64.2%	86.8%	75.9%	15.9%	91.5%	65.7%		
NY	DIAB	30	12.1%	0.804	88.0%	86.2%	12.5%	95.0%	70.5%	91.3%	81.1%	11.2%	93.5%	75.7%		
NY	DIAB	100	0.5%	0.916	92.0%	92.4%	7.9%	97.3%	79.4%	94.3%	88.8%	7.1%	96.2%	83.9%		
(C) Narrow high-performance network scoring system																
FL	CRC	11	59.9%	0.538	57.5%	89.1%	18.8%	63.7%	86.3%	69.8%	82.6%	20.6%	57.3%	89.1%		
FL	CRC	30	33.8%	0.637	59.1%	92.8%	15.6%	73.2%	87.2%	85.5%	86.6%	13.7%	68.0%	94.7%		
FL	CRC	100	4.9%	0.790	60.9%	97.0%	12.0%	87.2%	88.2%	79.2%	91.5%	11.6%	75.7%	93.0%		
FL	GSO	11	67.2%	0.728	64.2%	92.3%	14.7%	73.6%	88.5%	77.8%	86.9%	15.4%	66.5%	92.2%		
FL	GSO	30	46.3%	0.801	65.0%	95.5%	12.1%	82.8%	89.1%	80.9%	90.1%	12.2%	73.1%	93.4%		
FL	GSO	100	16.3%	0.891	66.6%	98.1%	9.7%	92.3%	99.8%	85.5%	93.4%	8.6%	81.1%	95.1%		
FL	DIAB	11	63.5%	0.656	55.2%	95.1%	14.9%	78.9%	86.4%	82.8%	87.1%	14.0%	68.1%	93.8%		
FL	DIAB	30	24.0%	0.775	54.3%	97.6%	13.3%	88.1%	86.5%	85.6%	90.1%	11.0%	74.2%	95.0%		

Continued

Table 3. Continued

State	Measure	Denominator Restriction	Percent of Physicians Profiled	Median Reliability	Empirical Cut Points						Reliability-Adjusted Cut Points					
					True Positive Rate	True Negative Rate	Misclassification Rate	Positive Predictive Value	Negative Predictive Value	True Positive Rate	True Negative Rate	Misclassification Rate	Positive Predictive Value	Negative Predictive Value		
FL	DIAB	100	2.0%	0.890	53.4%	98.9%	12.4%	94.4%	86.4%	88.2%	92.3%	8.8%	79.2%	95.9%		
NY	CRC	11	52.4%	0.498	53.2%	90.0%	19.2%	63.8%	85.2%	69.9%	81.1%	21.7%	55.2%	89.0%		
NY	CRC	30	22.7%	0.616	54.3%	94.0%	15.9%	75.2%	86.1%	74.5%	85.3%	17.4%	62.8%	90.9%		
NY	CRC	100	1.1%	0.800	55.5%	98.0%	12.6%	90.4%	86.9%	82.1%	90.6%	11.5%	74.4%	93.8%		
NY	GSO	11	62.4%	0.786	69.1%	93.4%	12.7%	77.7%	90.1%	79.4%	93.4%	10.1%	80.0%	93.1%		
NY	GSO	30	39.1%	0.850	70.9%	96.0%	10.3%	85.4%	90.8%	82.4%	96.0%	7.4%	87.2%	94.2%		
NY	GSO	100	9.1%	0.923	73.1%	98.4%	8.0%	93.7%	91.6%	86.8%	98.4%	4.5%	94.6%	95.7%		
NY	DIAB	11	50.7%	0.685	40.1%	96.0%	17.9%	77.1%	82.8%	74.5%	84.9%	17.7%	62.2%	90.9%		
NY	DIAB	30	12.1%	0.804	38.3%	98.7%	16.4%	90.9%	82.8%	80.1%	89.9%	12.6%	72.5%	93.1%		
NY	DIAB	100	0.5%	0.916	34.7%	99.9%	16.4%	99.1%	82.1%	88.6%	93.2%	8.0%	81.2%	96.1%		

CRC, colorectal cancer; DIAB, diabetes; GSO, glaucoma screening.

The median reliability column in Table 3 facilitates comparing reliabilities to misclassification rates. Although there is no agreement on a level of reliability that is considered acceptable across all physician profiling applications, 0.70 has been suggested as a minimum and levels of 0.85–0.90 for high-stakes applications (Roland et al. 2009). Even median reliabilities in these broadly accepted ranges can produce misclassification. Table 3 shows that median physician reliabilities of greater than 0.90 (e.g., glaucoma screening in NY, minimum denominator size of 100) can produce misclassification rates of more than 10 percent, while those greater than 0.70 can still result in misclassification rates of 20 percent or more. Unlike reliabilities, which are identical across Table 3A–C, the misclassification rates vary depending upon the stringency of the hurdle, highlighting the importance of considering both reliability and misclassification.

There is an interesting and perhaps surprising simple cut point effect for some of the scenarios. Consider the pattern of the 25th percentile results for CRC in New York (Table 3B). For the reliability-adjusted cut point, the true positive rate increases from 79.0 percent to 89.5 percent monotonically as the minimum sample size is increased. Similarly, the true negative rate increases monotonically from 71.6 percent to 83.0 percent as the sample size minimum is increased. Contrast this with the corresponding empirical cut point case where the true positive rate increases from 96.6 percent to 100.0 percent. However, the true negative rate decreases from 24.0 percent to 10.5 percent. This is explained by providers whose true scores are between the empirical and reliability-adjusted cut points. The higher their reliability, the more likely they are to be misclassified as the probability of their observed scores landing above the empirical cut point increases. This phenomenon also explains the higher true positive rate compared to the model adjusted cut point case.

DISCUSSION

The examination of misclassification risk for quality measures is timely given the burgeoning interest in health care value, as reflected by constructing HPNs using quality information and pending changes in Medicare physician payments. Understanding misclassification risk supports a more principled approach to determining whether the measures and sample sizes are adequate for network construction. It provides a target for balancing the needs of purchasers and providers with respect to how an HPN performs.

It is a more direct and policy-relevant measure of the potential impact of network restrictions on providers and patients than reliability alone. This is especially important as misclassification rates are defined according to either absolute or relative thresholds or performance tiers, and there is a need for greater understanding among network constructors of the interaction between establishing performance thresholds and the likelihood of misclassification.

It is important to evaluate the precision of tiering approaches with respect to how providers are ultimately classified. Tiering might need to balance the preferences of other stakeholders, such as purchasers, who might be willing to tolerate higher misclassification risk in exchange for being able to provide adequate market coverage, or providers, who might prefer lower probabilities of being misclassified into lower performance tiers (Normand et al. 2007). Reducing misclassification risk might also enhance the trust consumers have in their health plans; in a survey of individuals in health plans in Massachusetts, the majority of respondents did not trust or did not know whether they trusted that the tiers could indicate which physicians were better than others (Sinaiko and Rosenthal 2010).

We should not conclude that the misclassification results presented here are the best possible. There are several potential ways to improve reliability and misclassification. One approach would be to favor tiering approaches with lower misclassification rates; for example, in the choice between the scoring systems in Table 3A versus C, based on our data analysis Table 3C would be preferable given the lower misclassification rates for most scenarios. The use of statistical “buffer zones” around cut points might reduce misclassification for those providers whose scores are in the zone and for whom there is relatively large uncertainty about their correct tier assignment (Friedberg and Damberg, 2011). Approaches that increase sample size are promising, such as increasing the size of the measure sets and developing composite measures of performance (Reeves et al. 2007). Examining data for physician groups and provider organizations (and other higher level entities) or combining years of data on physicians are possibilities. Combining data sources (e.g., commercial and Medicare) would not only increase sample size but would also increase the completeness of the characterization of a providers practice, addressing the limitation of our analysis being restricted to only Medicare FFS beneficiaries. However, the policy relevance of such data aggregations would need to be assessed given the goals of particular tiering efforts, such as examining quality for Medicare beneficiaries by coverage type (Ayanian

et al. 2013). Also, it is important to note that reliability is not just a function of sample size. Combining populations may alter the provider-to-provider variation as well, possibly offsetting or even reversing sample size gains. Whether any given health plan would be better off using its own smaller sample size versus a larger dataset that includes other plans members is an open question. Statistical testing is potentially useful, though not a panacea. Statistical testing can be a tradeoff of true positive versus true negative rates that may or may not serve some system design goals. Statistical testing versus the mean can be a strong standard of evidence of a difference resulting in relatively few providers being identified as above or below average. Finally, classifications of providers into performance tiers are typically relative and clinical guidance is often lacking whether above-threshold performance is clinically optimal or below-threshold performance is still adequate care.

The misclassification rates we found might be pessimistic relative to profiling physicians on a composite constructed from numerous inputs or for other ways to select thresholds when designing HPNs. Regardless of how physician quality profiles are derived and categorized, the framework presented here for evaluating misclassification risk can be applied. Alternatively, our findings might be optimistic for several reasons. First, the provider-to-provider variation we found may be relatively large for binary quality measures, making distinguishing among providers easier. Nearly “topped out” measures with scores clustered near one are unlikely to have as much provider-to-provider variation. Second, Florida and New York are relatively large states with larger sample sizes to aid in estimating the cut points. Empirical cut points estimated for smaller states are likely to have higher variation, thus contributing to a greater proportion of the misclassification rate. Third, we have used measures that apply to large proportions of the Medicare population. Measures restricted to narrow subsets of the population may have much smaller denominators. Fourth, there is no case mix adjustment used. While this is traditional for these measures, there is recent discussion of whether such measures should be adjusted for sociodemographic characteristics (National Quality Forum 2014). Adjusting for beneficiary characteristics that vary across providers and that are associated with pass rates could reduce provider-to-provider variation with a corresponding increase in misclassification. Misclassification rates will also vary based on the structure of the tiering system (e.g., how and where the cut points are made). The ultimate benefit from assessing misclassification risk will be to strengthen the credibility of tiering systems, incentivizing the production of high-quality health care.

ACKNOWLEDGMENTS

Joint Acknowledgement/Disclosure Statement: This work is supported by grant 1 R21 HS021860 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

Disclosures: None.

Disclaimers: None.

REFERENCES

- Adams, J. L. 2009. *The Reliability of Provider Profiling: A Tutorial*. TR-653-NCQA. Santa Monica, CA: RAND Corporation.
- Adams, J. L., A. Mehrotra, J. W. Thomas, and E. A. McGlynn. 2010. "Physician Cost Profiling—Reliability and Risk of Misclassification." *The New England Journal of Medicine* 362 (11): 1014–21.
- Ayanian, J. Z., B. E. Landon, A. M. Zaslavsky, R. C. Saunders, L. G. Pawlson, and J. P. Newhouse. 2013. "Medicare Beneficiaries More Likely to Receive Appropriate Ambulatory Services in HMOs than in Traditional Medicare." *Health Affairs* 32 (7): 1228–35.
- Biles, B. 2014. *Medicare Advantage: Changing Networks and Effects on Consumers*. Washington, DC: Congressional Testimony for the Committee on Aging, United States Senate.
- Centers for Medicare and Medicaid Services. 2016. "The Medicare Access and CHIP Reauthorization Act of 2015: The Merit-Based Incentive Payment System (MIPS)" [accessed August 27, 2016]. Available at <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/Quality-Payment-Program-MIPS-NPRM-Powerpoint.pdf>.
- Cigna Health Corporation. 2016. "Cigna Care Designation and Physician Quality and Cost-Efficiency Displays 2015 Methodologies Whitepaper for Health Care Professionals" [accessed on May 26, 2016]. Available at <http://www.cigna.com/pdf/2015-cigna-care-designation-methodology.pdf>
- Draper, D. A., A. Liebhaber, and P. B. Ginsburg. 2007. "High-Performance Health Plan Networks: Early Experiences." *Issue Brief* 111.
- Friedberg, M. W., and C. L. Damberg. 2011. *Methodological Considerations in Generating Provider Performance Scores for Use in Public Reporting: A Guide for Community Quality Collaboratives*. Rockville, MD: Agency for Healthcare Research and Quality. AHRQ Pub. No. 11-0093.
- Hofer, T. P., R. A. Hayward, S. Greenfield, E. H. Wagner, S. H. Kaplan, and W. G. Manning. 1999. "The Unreliability of Individual Physician "Report Cards" for

- Assessing the Costs and Quality of Care of a Chronic Disease.” *Journal of the American Medical Association* 281 (22): 2098–105.
- Kassirer, J. P. 1994. “The Use and Abuse of Practice Profiles.” *The New England Journal of Medicine* 330 (9): 634–6.
- Lin, R. H., T. A. Louis, S. M. Paddock, and G. Ridgeway. 2006. “Loss Function Based Ranking in Two-Stage, Hierarchical Models.” *Bayesian Analysis* 1 (4): 915–46.
- Mehrotra, A., J. L. Adams, J. W. Thomas, and E. A. McGlynn. 2010. “The Effect of Different Attribution Rules on Individual Physician Cost Profiles.” *Annals of Internal Medicine* 152 (10): 649–54.
- National Committee for Quality Assurance. 2014. *HEDIS & Quality Measurement Certified Survey Vendors/Auditors/Software Vendors: HEDIS Compliance Audit Program*. Washington, DC: National Committee for Quality Assurance.
- National Quality Forum. 2014. *Risk Adjustment for Socioeconomic Status or Other Sociodemographic Factors*. Washington, DC: National Quality Forum.
- Normand, S.-L. T., R. E. Wolf, J. Z. Ayanian, and B. J. McNeil. 2007. “Assessing the Accuracy of Hospital Clinical Performance Measures.” *Medical Decision Making* 27 (1): 9–20.
- Paddock, S. M. 2014. “Statistical Benchmarks for Health Care Provider Performance Assessment: A Comparison of Standard Approaches to a Hierarchical Bayesian Histogram-Based Method.” *Health Services Research* 49 (3): 1056–73.
- Paddock, S. M., J. L. Adams, and F. Hoces de la Guardia. 2014. “Better-than-Average and Worse-than-Average Hospitals May Not Significantly Differ from Average Hospitals: An Analysis of Medicare Hospital Compare Ratings.” *British Medical Journal Quality & Safety* 24: 128–134. doi:10.1136/bmjqs-2014-003405.
- Paddock, S. M., and T. A. Louis. 2011. “Percentile-Based Empirical Distribution Function Estimates for Performance Evaluation of Healthcare Providers.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 60 (4): 575–89.
- Reeves, D., S. M. Campbell, J. Adams, P. G. Shekelle, E. Kontopantelis, and M. O. Roland. 2007. “Combining Multiple Indicators of Clinical Quality: An Evaluation of Different Analytic Approaches.” *Medical Care* 45 (6): 489–96.
- Roland, M., M. Elliott, G. Lyratzopoulos, J. Barbiere, R. A. Parker, P. Smith, P. Bower, and J. Campbell. 2009. “Reliability of Patient Responses in Pay for Performance Schemes: Analysis of National General Practitioner Patient Survey Data in England.” *British Medical Journal* 339.
- Scholle, S. H., J. Roski, J. L. Adams, D. L. Dunn, E. A. Kerr, D. P. Dugan, and R. E. Jensen. 2008. “Benchmarking Physician Performance: Reliability of Individual and Composite Measures.” *The American Journal of Managed Care* 14 (12): 833–8.
- Sequist, T. D., E. C. Schneider, A. Li, W. H. Rogers, and D. G. Safran. 2011. “Reliability of Medical Group and Physician Performance Measurement in the Primary Care Setting.” *Medical Care* 49 (2): 126–31.
- Sinaiko, A. D., and M. B. Rosenthal. 2010. “Consumer Experience with a Tiered Physician Network: Early Evidence.” *The American Journal of Managed Care* 16 (2): 123–30.

- Smith, K. A., J. B. Sussman, S. J. Bernstein, and R. A. Hayward. 2013. "Improving the Reliability of Physician "Report Cards"." *Medical Care* 51 (3): 266–74.
- Summer, L. 2014. "Health Plan Features: Implications of Narrow Networks and the Trade-Off between Price and Choice" [accessed on August 27, 2016]. *Academy Health Research Insights*. Available at <http://academyhealth.org/files/HCFO/RIBrief0315.pdf>.
- Werner, R. M., and D. A. Asch. 2005. "The Unintended Consequences of Publicly Reporting Quality Information." *Journal of the American Medical Association* 293 (10): 1239–44.