Appendix 5.2.3a: Additional Reliability Testing Results

Section 5.2.2: Method(s) of Reliability Testing

We performed a signal-to-noise analysis to assess the reliability of the measure's performance rates at the clinician and clinician group levels. The signal-to-noise reliability statistic R (also known as the SNR), which ranges from zero to one, summarizes the proportion of performance variation that is a result of true differences in underlying entity characteristics (such as differences in the quality of clinical care) as opposed to random variation (for example, because of sampling error). If R = 0, there is no true variation on the measure across entities, and all observed variation is a result of random error. In this case, the measure is not useful for distinguishing between entities with respect to differences in the quality of health care provided.

The PQM expects SNR to be at least 0.6 for at least 50%, and preferably at least 80%, of the accountable entities subject to reporting on the measure. Measure developers are now required to report not only median SNR values, but also decile-specific SNR values to highlight the full distribution of reliabilities at the accountable entity level.

CBE #1879 is a binary, non-risk adjusted measure of physician performance with respect to patient adherence to antipsychotic therapy. Therefore, we estimated reliability using estimated parameters from a beta-binomial model, which involves two assumptions:

1. Each clinician/clinician group/state has a true pass rate, p, which varies and follows a beta distribution: p~Beta(α,β).
2. Given p and the sample size n for the accountable entity (where the sample size is considered to be uninformative with regard to performance), the numerator Y for each entity's performance rate is a binomial random variable Y | p ~ Binomial(n, p). Using this model, the reliability of the accountable entity's performance rate $\bar{Y}$= Y/n is defined using the usual variance formula (aka the signal-noise ratio) where the numerator is the between-provider variance in the true pass rate and the denominator is the total variance of Y.

$$\rho = \frac{Var(\,p\,)}{Var(\,\bar{Y}\,)} = \frac{Var(\,p)}{Var(\,p) + E\{\,Var(\,\bar{Y}|p)\}}$$

Although this method is well-suited for application in the context of provider profiling, a standard practice has been to replace the second term in the denominator with a provider-specific Var($\bar{Y}$| p), which is then estimated with a plug-in estimator $\widehat{Var}(\,\bar{Y}|p) = \hat{p}_i(1 - \hat{p}_i)/n$, where $\hat{p}_i$ is an estimate of the provider-specific rate.[1] This standard practice is flawed, though, both in theory and in practice, as provider-specific rate estimates are very unstable, given small provider-specific sample sizes and the frequency of provider rate estimates close to the extremes (0% or 100%). Although other developers have suggested estimating provider-specific variances using more stable plug-in estimates of provider-specific rates (e.g. Bayesian estimates[2] or smoothed estimates from hierarchical regression models[3]), we advocate for an alternative approach recently introduced into the provider-profiling literature by Nieser and Harris.[4] In that approach, which is aligned with foundational work on reliability from the psychometrics literature,[5] the reliability target parameter is defined as above and the provider-specific reliability can be estimated with $\widehat{\rho_i} = n_i / (n_i + \hat{\alpha} + \hat{\beta})$, where n_i is the provider sample size and $\hat{\alpha}$ & $\hat{\beta}$, are estimates of the beta-binomial parameters, such as maximum-likelihood estimates from analyzing large samples of provider numerators and denominators. This approach is theoretically sound in estimating the correct target parameter, and it avoids relying on data-poor estimates of provider-specific rates, instead pooling all of the data to estimate only two parameters, alpha and beta, which can then be plugged in to generate provider-specific reliability estimates that are sample-size specific. In our own simulation studies, this approach outperformed Bayesian approaches, even when the target of inference uses just the provider-specific within-provider variance in the denominator, not the overall expectation of that conditional variance over all providers. This approach also avoids the estimation problem when providers have rates of 0% or 100%, and it can be implemented with equivalent results using either SAS or R.

In particular, to get maximum likelihood estimates of the sum $\hat{\alpha} + \hat{\beta} = \hat{\varphi}$ that can then be used to estimate the provider-specific reliabilities described above, we used SAS PROC FMM to fit an intercept-only beta-binomial regression model on a dataset with one record per provider that contains separate variables for the provider numerator and denominator.

References

1. Adams JL, Mehrotra A, Thomas JW, McGlynn EA. Physician cost profiling--reliability and risk of misclassification. *N Engl J Med*. 2010;362(11):1014-1021. doi:10.1056/NEJMsa0906323
2. Zhou G, Lin Z. Improved beta-binomial estimation for reliability of healthcare quality measures. medRxiv 2023.01.07.22283371. https://doi.org/10.1101/2023.01.07.22283371
3. Hwang J, Adams JL, Paddock SM. Defining and estimating the reliability of physician quality measures in hierarchical logistic regression models. Health Serv Outcomes Res Methodol. 2020;21:111-130.
4. Nieser KJ, Harris AHS. Comparing methods for assessing the reliability of health care quality measures. *Stat Med*. 2024;43(23):4575-4594. https://doi.org/10.1002/sim.10197
5. Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27,* 59–72. https://doi.org/10.1007/BF02289665

Section 5.2.3a: Additional Reliability Testing Results

As described in Section 5.2.3 Reliability Testing Results, Table 5.2.3-1 provides National Provider Identifier (NPI)-level reliability results, and Table 5.2.3-2 provides Tax Identification Number (TIN)-level reliability results. Note that Table 5.2.3-2 corresponds to the data provided in **Table 2** of the Full Measure Submission.

Table 5.2.3-1: NPI-Level Reliability Testing Results

	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability	0.771	0.624	0.646	0.675	0.707	0.733	0.764	0.799	0.833	0.865	0.903	0.978	0.978
Mean Performance Score	0.775	0.171	0.761	0.758	0.781	0.766	0.758	0.784	0.779	0.792	0.790	0.790	1
N of Entities	2,311	1	249	238	188	266	219	241	210	233	234	233	1
N of Persons / Encounters / Episodes	131,490	20	5,202	5,693	5,097	8,234	7,986	10,491	11,280	15,566	21,202	40,739	540

Table 5.2.3-2: TIN-Level Reliability Testing Results

	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability	0.770	0.615	0.637	0.675	0.698	0.731	0.764	0.800	0.834	0.868	0.906	0.991	0.991
Mean Performance Score	0.775	0.171	0.766	0.765	0.769	0.767	0.766	0.779	0.778	0.781	0.788	0.789	1
N of Entities	2,254	1	237	213	239	232	206	215	228	234	224	226	1
N of Persons / Encounters / Episodes	138,508	20	4,959	5,099	6,591	7,420	7,716	9,598	12,693	16,786	22,024	45,622	1,365