© Health Research and Educational Trust
DOI: 10.1111/1475-6773.12086
RESEARCH ARTICLE

Blending Group and Practice Site Scores to Increase the Reliability of Physician Quality Information

William H. Rogers, Kristy Thornton, and Ted von Glahn

Objective. To use an empirical Bayesian approach, blending practice, and group quality data with physician results to increase the accuracy of quality of care measures.

Data Sources. Performance data on diabetes glycemic screening for 8,357 physicians collected from multiple payers as part of a statewide physician performance reporting initiative.

Study Design. A variance components analysis assessed the strength of group, practice, and physician effects compared with random error. We derived formulas to describe reliability and measurement error variances and calculated the optimal blend of physician, practice, and group data. We constructed a simulation to show what various methods can achieve. The value of blending strategies was assessed by simulating a common pay-for-performance criterion—performance in the top 25 percent. We estimated the proportion of physicians whose true percentage would place them in the top 20 percent but who would not receive payment based on the observed success rate.

Principal Findings. Blending reduced the error rate from 29.7 to 22.7 percent. Simpler empirical Bayes estimates using shrinkage alone produced no gains over simple doctor percentages.

Conclusions. When good structural data about physician groups and practices exist, gains from blending can be substantial.

Key Words. Quality of care, patient safety, variance components modeling, multilevel, shrinkage

Quality of care measures for individual physicians, practices, and groups are in widespread use by health care purchasers in the public and private sectors. Health insurance companies, large employers, and the federal government use physician quality profiles to create programs aimed at changing the behavior of both health care providers and consumers (Rosenthal et al. 2004, 2006; Med-Vantage Inc. 2011). These initiatives include provider quality improvement programs, value-based contracting, and public reporting of quality measure results.

A lack of consensus exists over whether the appropriate unit of accountability for quality measures is the individual physician or an organizational structure, such as practice site or medical group. Individual physician-level quality profiles are most useful to support consumer decision making; consumers are most interested in selecting an individual physician, as opposed to a medical group (Milstein and Lee 2007; Mehrotra et al. 2010). In addition, strong evidence exists of significant variations in quality performance at the level of individual physicians (Goodell and Harris 2008; Holmboe et al. 2010). These variations, coupled with the central role of individual physicians in clinical decision making, create a strong leverage point for quality improvements. However, many stakeholders favor health care quality profiles that focus on physician groups, and a key aspect of their argument is the issue of reliability (Scholle et al. 2008; Sequist et al. 2011).

Reliability is a measure assessing the ratio of signal to noise; its value can range from 0 (all variability arises from random sampling considerations) to 1 (all variability arises from true differences). In the case of physician quality scores, reliability reflects the proportion of the variability in physician performance that can be explained by real performance differences. High reliability means that one can confidently distinguish the performance of one physician from another and the risk of misclassifying a provider is low (Adams, Mehrotra, and McGlynn 2010). Low reliability means that the data are essentially random and useless for most quality measurement purposes, such as informing consumers, and that classification mistakes are likely. As noise is inversely related to sample size, adequate sample size is a major consideration. With the exception of some preventive care measures, the sample size of quality events available at the individual physician level often is too small to result in high reliability when data are supplied by a health plan or even through pooling data across multiple plans (Nyweide et al. 2009).

The quest to reduce noise has led to numerous suggestions that can be divided into two groups. The first is to create composite measures that combine data across similar metrics. A second option is to “shrink” the estimates toward the expected value using empirical Bayes logic (Efron and Morris 1977).

Address correspondence to Kristy Thornton, B.S., The Pacific Business Group on Health, 221 Main Street, Suite 1500, San Francisco, CA 94105; e-mail: kthornton@pbgh.org. William H. Rogers, Ph.D., is with The Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA. Ted von Glahn, M.S., is with The Pacific Business Group on Health, San Francisco, CA.

Composite measures might be constructed by combining measures that have elements of similarity in analytically derived ways, such as factor analysis. These methods reduce information clutter; however, in our experience, they do not substantially increase reliability. Larger available sample sizes are offset by weak relationships between measures and multiple uses of data from the same patients.

In this study, we focus on solutions that come from empirical Bayes logic applied to a single measure. The starting point is the usual mean estimate for a physician, the “best linear unbiased prediction” or BLUP. We begin by evaluating the value of shrinking toward the average of the measure across all physicians and then discuss a more complex form of shrinking that uses the group and practice structure of the measurement system, a method we call “blending.” Physicians joining together to form a practice may share compatible clinical views and styles and become increasingly similar over time under the influence of practice leadership. Multiple physician practices that are part of a medical group may have information systems or other mechanisms that impose a degree of uniformity. Physician practice sites and, to an even greater degree, physician groups have more quality events available than do individual physicians, typically resulting in higher reliability (Scholle et al. 2008; Mehrotra et al. 2010; Sequist et al. 2011).

METHODS

To illustrate the properties and challenges of physician-level performance reporting, we used a dataset of medical and pharmacy claims information from three California commercial health insurance plans submitted for the California Physician Performance Initiative (CPPI). CPPI is a multistakeholder collaborative demonstration project, sponsored by the California Cooperative Healthcare Reporting Initiative, to measure and report the quality of patient care provided by individual physicians. Physicians agreeing to participate in the project were affiliated with one or more medical groups in the California Department of Managed Health Care system and reported being in one or more practice locations. In 2009, CPPI scored more than 12,000 California physicians on 17 administrative clinical quality measures.

For this study, the measurement year spanned a 12-month period from October 1, 2007 to September 30, 2008. Using a deidentified secondary dataset, we examined a single diabetes care measure from the Healthcare Effectiveness and Data Information Set (HEDIS): the percentage of patients

between the ages of 18 and 75 years with a hemoglobin A1c screening test (HBA) performed during the measurement year. The designated processor of HEDIS data (Thompson-Reuters Healthcare, now Truven Healthcare Analytics) attributed patient data to the primary care physician (PCP) seen most frequently (in case of a tie, the one seen most recently) and to any endocrinologist who had at least one encounter/claim for evaluation and management services during the measure's attribution period. Using a beta-binomial analysis, 26 patients per physician would be needed for this measure to achieve 70 percent reliability. Only 13 percent of the PCPs in the sample had that many patients with diabetes; the median number of patients with diabetes per physician was nine.

The dataset contained information on physician practice and group affiliations, which we used to infer group-practice-doctor relationships at the patient level. Patients who were eligible for the HBA measure (between 18 and 75 years of age with type 1 or type 2 diabetes) were divided into four groups depending on the available data about the doctor's practice and group.

- Analyzable practice, group, and doctor: A hierarchical, nested arrangement in which there is more than one doctor in the practice and more than one practice in the group.
- Analyzable practice and doctor: An independent practice with more than one doctor and with multiple or no group affiliations.
- Analyzable group and doctor: A group with multiple practice locations, a single doctor in a practice affiliated with a group, or an unknown practice.
- Analyzable doctor only: Instances in which data about structure were missing or combinations not listed above, for example, a doctor practicing in multiple settings.

Within each group, we estimated hierarchical models to describe the allocation of variance among group, practice location, doctor, and residual (including patient) or the available subset of these levels. A large amount of variance suggests that the corresponding hierarchical level has substantial influence over whether the HBA test is conducted or not for eligible patients.

Blending

The blending method calculates a weighted average of physician, practice, and/or group data. The weights depend on the relative amounts of variance at each level, as well as the standard error of these estimates and the sample sizes.

If the three-level hierarchy is not available, a two-level hierarchy should be used if possible.

Blending is an empirical Bayes or “shrinkage” method because the physician score “shrinks” toward the practice and group scores. A biased estimate, blending is not based solely on the physician’s data. In return for a relatively small bias, there is a larger reduction in the standard error of the estimate. However, the blended estimate differs from common empirical Bayes shrinkage methods because it does not attempt to shrink toward the overall physician average but rather a better estimate based on group and practice data (Efron and Morris 1977).

To understand how the blending weights are used, we will denote them by w_G , w_P , and w_D for group, practice, and doctor weights, respectively, where $w_G + w_P + w_D = 1$. If the group or practice information is not available or if there are multiple groups or practices, then the corresponding weight is 0. If the group, practice, and doctors averages are similarly denoted as y_G , y_P , and y_D , then the weighted average is $w_G y_G + w_P y_P + w_D y_D$.

The weights are computed by optimizing them so that the weighted average has minimum mean square error relative to the true doctor average. Let e_G stand for a group effect, e_P stand for the practice effect, and e_D stand for the doctor effect. Then the expected values

$$E y_D = \mu + e_D + e_P + e_G$$

$$E y_P = \mu + e_P + e_G$$

$$E y_G = \mu + e_G$$

This is a standard hierarchical summed effects model used in analysis of variance. In doctor performance evaluation, the object of estimation is $E y_D$. A practice or group mean is a biased representation of $E y_D$. The bias can be obtained by subtraction; for y_P the bias is e_D , and for y_G the bias is $e_D + e_P$. Based on a variance components model, we can compute the variance of those effects. After accounting for the bias contributed by parts of the structure (e.g., other doctors in the practice), we can compute squared bias estimates B_G , B_P , and a doctor effect variance Σ_D . On average, the squared biases in the practice and group averages are Σ_D and $B_P + \Sigma_D$, respectively.

The mean square error depends on the average weighted squared bias and the squared standard errors of the estimates, denoted by S^2 . We are trying to minimize

$$\text{MSE} = w_D^2 S_D^2 + w_P^2 (S_P^2 + \Sigma_D) + w_G^2 (S_G^2 + \Sigma_D + B_P)$$

This becomes a Lagrange multiplier problem in calculus, which leads to the equations:

$$\begin{aligned}\tau_D w_D &= \zeta \\ (\tau_P + \Sigma_D) w_P &= \zeta \\ (\tau_G + \sigma_D + \Sigma_D B_P) w_G &= \zeta\end{aligned}$$

The value of ζ is resolved by the fact that the three weights must sum to 1. We show more details of the solution for a two-level structure (omitting group) in the appendix, which is available online.

The idea of actually shrinking toward the mean, when described in empirical Bayes references (Efron and Morris 1977), has appealed to statisticians as it was originally proposed. In this application, it would lead to using a fourth component in the equation, a mean of all the observed data. The mathematical solution is similar to that shown in this study. In addition, it would be simpler to ignore the group and practice data and just shrink the doctor averages, if the sample size is insufficient. A simulation illustrating this method follows.

Simulation

We wondered how blending would work, with and without shrinkage, so we conducted an experiment to find out. Due to several possible configurations of data and the complexity of discerning truth from observed data, we elected to run a simulation based on the effects we found in the variance components analysis of HBA.

We generated a set of known means based on doctor and practice effects and then simulated patient values. To avoid complications related to binary distributions, we based the simulations on normal random variables. The simulation study compares the performance of four different types of estimates on the same data: the original doctor-level mean (no shrinkage), classical empirical Bayes shrinkage toward the grand mean based on the doctor mean, the proposed shrinkage based on a combination of doctor and practice means including the grand mean; and the latter without including the grand mean. The true distribution of the data was generated from the practice and doctor variance components equation, which was fit using the *xtmixed* procedure in Stata. We started with a simplified simulation directed at small-sample inference with 12 patients per doctor and 4 doctors per practice. For a more realistic experiment, we used a design with mixed sample sizes, as found in the CPPI data.

Evaluation

We considered two calculations that led us to want greater precision: the probability of error and reliability. Probability of error is represented by questions such as: If a doctor’s performance lies in the upper 20 percent of the distribution, what is the chance that the data will not put him in the top 25 percent? From the doctor’s perspective, the question is “If I am truly deserving of a pay-for-performance bonus, what is the chance I won’t get it?” If a doctor’s performance is exactly at the 75th percentile, there is an even chance that data will put him on the wrong side of the boundary, a situation that statistics cannot help. We also evaluated whether a doctor in the upper 45 percent of performance would have a score greater than the median.

Reliability is harder to understand. In the introduction to this article, we associated reliability with the signal to noise ratio. For this purpose, we should consider squared bias to be part of the noise, but this may not fully describe the impact of bias. To help readers understand the results of the simulation in reliability terms, we computed the reliability that would yield equivalent gains in the probability of error.

RESULTS

Table 1 illustrates the complex structure imposed by varied configurations of California medical care organizations and the nature of the CPPI demonstration project. This structure includes detail to help understand the context of

Table 1: Number of Doctors, Patients, and Average HBA Performance, by Type of Doctor

<i>Type of Doctor</i>	<i>No. of Doctors</i>	<i>No. of Patients</i>	<i>Average HBA Performance (%)</i>
Analyzable practice and group	1,204	16,879	75.9
Analyzable practice, multigroup	3,070	46,341	76.4
Analyzable practice, group unavailable	1,205	15,520	72.3
Analyzable practice, unique group*	0	0	
Multipractice, analyzable group	508	7,359	75.1
Practice unavailable, analyzable group	72	874	77.3
One-person practice, analyzable group	999	14,870	73.2
Multipractice, multigroup	904	12,031	74.3
Other, analyzable by doctor only	395	4,823	71.6
Total	8,357	118,697	74.9

*Due to the way that practice was assessed in the CPPI, practices were enumerated within groups.

our results imposed by the nature of California medical care and the nature of the CPPI demonstration project. The largest group of patients has accessible practice and doctor information, but the practice is associated with multiple groups.

Table 2 shows variance components results for several different structural possibilities in the CPPI, compressed into sets according to the maximally detailed model that makes sense for each patient. For example, one-doctor practices are mapped into the group-doctor estimate, and practice is ignored because in these situations our weighting formula assigns zero weight to the practice mean.

For the low-sample-size simulation experiment (12 patients per doctor and 4 doctors per practice), we evaluated four analysis scenarios characterized by whether doctor–practice blending is used and whether the overall mean is used as a point of shrinkage. The results are shown in Table 3. Shrinkage lowers the mean squared error dramatically, but it does nothing to lower the error rates based on percentile scores. Blending weights the practice and doctor

Table 2: Standard Deviation of Random Effects in HBA Performance, Percent, by Type of Doctor

<i>Configuration</i>	<i>Group</i>	<i>Practice</i>	<i>Doctor</i>	<i>Residual</i>	<i>Doctor Only (Ignoring Group, Practice)</i>
Group, practice, doctor	9.88	7.58	6.67	40.99	12.63
Practice, doctor		12.08	7.42	40.75	14.19
Group, doctor	9.51		11.01	41.89	13.59
Doctor only			13.88	42.04	13.88
All doctors				41.19	13.84

Table 3: Weighting and Performance of Methods for the Small Sample Size Scenario*

<i>Method</i>	<i>Doctor Weight</i>	<i>Practice Weight (Excluding Doc. Pats.)</i>	<i>Grand Mean Weight</i>	<i>SD of Scores</i>	<i>MSE</i>	<i>Top 20% Error Rate</i>	<i>Top 45% Error Rate</i>
Simple means	1	0	0	18.4	138.8	29.7	19.4
Doctor shrinkage	0.592	0	0.408	10.9	82.1	29.7	19.4
Practice and doctor blend	0.463	0.537	0	14.4	64.4	22.7	15.2
Practice and doctor shrinkage	0.351	0.407	0.242	10.9	59.5	22.7	15.2

*4 physicians per group, 12 patients per physician, practice SD = 12.08, doctor SD = 7.42, and residual SD = 40.75.

results about equally, and lowers the error rates, but combining blending and shrinkage does not further lower the rates over blending alone. Shrinkage fails to improve relative performance; due to equal sample sizes, all doctors shrink by the same proportion in this simulation, leaving the doctors in the same rank as they would be without the shrinkage.

In this example, blending reduced the error probability from 29.7 to 22.7 percent. This could have been achieved by using simple means and sampling 20 patients per doctor instead of 12. The error probabilities for 20 patients with no blending and 12 patients with blending are the same. A design with 20 patients per doctor with simple doctor means would have a reliability of 71 percent, compared to 59 percent for 12 patients per doctor using unblended (BLUP) means.

We wondered how we might set absolute performance standards for data with shrinkage, as shrinkage does not change the relative performance when sample sizes are constant. This problem is not trivial because shrinkage always reduces observed variance. The standard deviation of the true scores is 14.2 percent. Table 3 shows that the ordinary means have a larger standard deviation (18.4 percent), but the shrinkage estimates are too tight (SD 10.9 percent), and they are not an appropriate basis for estimating the true performance distribution. Shrinkage underestimates how bad the worst performers are, and how good the best performers are, achieving its original goal of lowering the mean square error as much as possible. The original scores are also inappropriate for measuring the true distribution because they are more spread out.

In the mixed sample size scenario, the performance of the simple doctor mean and the simple shrinkage estimate would be the same in two-doctor and four-doctor practices, depending on the number of patients per doctor. Over-

Table 4: Weighting and Performance of Methods for the Mixed Sample Size Scenario*

<i>Configuration</i>	<i>Doctors per Practice</i>	<i>Patients per Doctor</i>	<i>Simple Mean Error Rate</i>	<i>Shrinkage Error Rate</i>	<i>Blended Error Rate</i>
A	4	12	28.1	31.0	23.0
B	4	20	24.2	21.4	19.1
C	2	12	28.2	31.3	23.5
D	2	20	24.6	21.3	19.5
Total	0.463	0.537	26.2	26.5	21.2

*Practice SD = 12.08, doctor SD = 7.42, residual SD = 40.75, and doctors per practice and patients per doctor in four equal probability configurations.

all, the correlation of the scores between these two methods was .986, and they have approximately equal performance overall. Using the pay-for-performance criterion of the 75th percentile and above, high-performing doctors are at a disadvantage relative to their peers who have larger samples. Unfortunately, this problem is accentuated by shrinkage. The error rates decrease more uniformly for blending, although not quite as much when there are only two doctors in the practice. For both doctor means and blending, the disadvantage that small-sample doctors have in achieving the 75th percentile when they are actually above the 80th percentile is offset by the advantage, in the form of a random bonus, given to doctors who are achieving lower than the 80th percentile. However, with simple shrinkage, doctors with small sample sizes have a difficult time of ever receiving a bonus unless they are top performers (data not shown).

DISCUSSION AND CONCLUSIONS

Empirical Bayes was originally designed to minimize mean square error and predict future performance by shrinking the observed results toward an expected value. If implemented as shrinkage toward a common mean, this method has nothing to offer for physician performance evaluation purposes. If sample sizes do not vary, the results give an identical ranking to the original data because all doctors shrink by the same amount. If sample sizes vary, doctors with smaller sample sizes will shrink more toward the mean, making it hard for them to obtain pay-for-performance incentives offered to high-performing practices.

Blending is also based on empirical Bayes theory but uses the practice mean or a combination of group and practice means as the point of attraction, rather than the overall means. If large amounts of the doctor's own data are available, the estimate is dominated by that data. However, if relatively little data are available, the estimate uses an optimal amount of information from the practice and/or group to increase accuracy. For doctors with smaller sample sizes, blending can substantially improve efficiency. Blending makes sense if the practice or group has a substantial influence on the performance, as evidenced by contributing a large component of variance. The data from the CPPI demonstration project showed that, for HBA and other diabetes performance measures, practice did indeed contribute a major portion of the variance. Our CPPI data suggested that these gains might be available for other

HEDIS measures, but other work suggests that gains would be less for patient experience measures.

Several limitations deserve mention. One of the groups in the state with a substantial laboratory practice did not give permission for their lab data to be used in this analysis, which may have raised the practice SD. Physician corrections were permitted for the primary purposes of the project, and financial constraints precluded audit for this study; we used the original, uncorrected data. The effect of these factors on our analysis remains unknown. Our simulation is based on normal random variables, but performance data are binomial. Empirical Bayes methods can be used in binomial data using the beta-binomial conjugate distribution pair, and it is also possible to use probit models based on underlying normal effects (Teixeira-Pinto and Normand 2008). We verified that using these techniques for binomial data increase complexity of analysis without materially affecting our results.

Applying blending takes some effort. Various practice arrangements must be considered, including doctors who have solo practices and those who are members of multiple practices and multiple groups. The CPPI data were not ideally suited to investigate these subtleties because we were not given the group and practice structure for each patient. We could only study cases with clear practice and group affiliation, which may have been better organized than other arrangements. This raises an important consideration, which is what how to address solo practitioners. Doctors who share an office but are legally solo practitioners might be grouped together. Beyond that, a practice could be defined by the patients of other doctors who are on call for a particular doctor, or a group might be defined by hospital admitting privileges. The only thing we can assert with authority is that it would be a bad idea to combine all the solo practitioners into a group.

In California, many practices are affiliated with more than one group. Without specific data on what patients are treated by which groups and practices, it is difficult to say how blending should work. It is likely that some structures will always be ambiguous, in which case we suggest avoiding blending or shrinking physician performance data. Blending should be used in cases where the structure is clear.

Although we can calculate the variance and mean squared, error for any blending configuration, we could not relate these to fair calculations of a reliability statistic. We can, however, calculate error rates using simulation techniques and compare error rates with similar experiments using simple doctor means, for both of which we can calculate reliability.

Using practice and group data to buttress scores for individual physicians may be resisted by physicians who prefer that their performance be assessed independent of their colleagues. Bias is introduced, but it is more than offset by lower standard errors if appropriate weights are used. For within-practice quality improvement, the group and practice structure will be fixed; the gains from blending will not improve the precision of internal comparisons. For this reason, blending might not be universally perceived as helpful.

Blending can be used with composites as well as individual measures. In these cases, one needs to be able to calculate the variance of the composite at the doctor level, for which a specialized variance components method could be needed.

The value of blending is context dependent in other ways. In the project from which our data were drawn, PCP screening and heart care measure performance data were also available. The PCP screening measures had accurate results and large sample sizes, which would lead to little emphasis on practice or group. An alternative to using group and practice data is to compel health plans to report quality data statewide. “All-payer claims” databases are gaining traction in some states, although years may elapse before all necessary linkages can be made while protecting individual privacy. The diabetes care measures also could have been enhanced by larger sample sizes, which could be attained through compulsory reporting. The heart measures from the CPPI demonstration had such small sample sizes that, even with blending and compulsory participation, it is hard to know that they would be reliable enough to report at least at the physician level. However, as quality assessment continues to evolve, we may find new measures and composites for which blending is helpful. The prescription for the success of blending is the presence of a strong group or practice structure—for example, the existence of uniformly imposed information systems that materially affects performance—and the availability of good data about physician practice structures at the patient level.

In summary, we found that, by itself, shrinkage to a common mean has little utility for performance measurement purposes. It does nothing to enhance relative ranking. On the other hand, doctor, practice, and group data can be blended to produce better estimates and more accurate rankings of individual doctor performances than would be obtained from doctor means, which are the usual estimates.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This manuscript was supported by a grant from the Robert Wood Johnson Foundation (#68275). Jennifer Green provided editorial support. William H. Rogers consults with the Pacific Business Group on Health, Massachusetts Health Quality Partners, and Blue Cross, where the methods described in this report could be applied. The authors thank Wylina Guan for proofreading.

Disclosures: None.

Disclaimers: None.

REFERENCES

- Adams, J. L., A. Mehrotra, and E. A. McGlynn. 2010. *Estimating Reliability and Misclassification in Physician Profiling*. Santa Monica, CA: RAND Corporation.
- Efron, B., and C. Morris. 1977. "Stein's Paradox in Statistics." *Scientific American* 236: 119–27.
- Goodell, S., and K. M. Harris. 2008. *Choosing a Health Care Provider: The Role of Quality Information*. The Synthesis Project. Princeton, NJ: Robert Wood Johnson Foundation.
- Holmboe, E. S., W. Weng, G. K. Arnold, S. H. Kaplan, S. L. Normand, S. Greenfield, S. Hood, and R. S. Lipner. 2010. "The Comprehensive Care Project: Measuring Physician Performance in Ambulatory Practice." *Health Services Research* 45 (6 Pt 2): 1912–33.
- Med-Vantage Inc. 2011. "Our Resources" [accessed on November 21, 2011]. Available at <http://www.medvantage.com>.
- Mehrotra, A., J. L. Adams, J. W. Thomas, and E. A. McGlynn. 2010. "Cost Profiles: Should the Focus be on Individual Physicians or Physician Groups?" *Health Affairs (Millwood)* 29 (8): 1532–8.
- Milstein, A., and T. H. Lee. 2007. "Comparing Physicians on Efficiency." *New England Journal of Medicine* 357 (26): 2649–52.
- Nyweide, D. J., W. B. Weeks, D. J. Gottlieb, L. P. Casalino, and E. S. Fisher. 2009. "Relationship of Primary Care Physicians' Patient Caseload with Measurement of Quality and Cost Performance." *Journal of the American Medical Association* 302 (22): 2444–50.
- Rosenthal, M. B., R. Fernandopulle, H. R. Song, and B. Landon. 2004. "Paying for Quality: Providers' Incentives for Quality Improvement." *Health Affairs (Millwood)* 23 (2): 127–41.
- Rosenthal, M. B., B. E. Landon, S. L. Normand, R. G. Frank, and A. M. Epstein. 2006. "Pay for Performance in Commercial HMOs." *New England Journal of Medicine* 355 (18): 1895–902.

- 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." *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.
- Teixeira-Pinto, A., and S. L. Normand. 2008. "Statistical Methodology for Classifying Units on the Basis of Multiple-Related Measures." *Statistics in Medicine* 27: 1329–50.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Optimal Weighting Formulas for a Practice–Doctor Blend.

Appendix SA2: Author Matrix.