# The Reliability of Medical Group Performance Measurement in a Single Insurer's Pay for Performance Program

*Hector P. Rodriguez, PhD, MPH,* Lisa Perry, PhD,† Douglas A. Conrad, PhD, MBA,† Charles Maynard, PhD,†‡ Diane P. Martin, PhD,† and David E. Grembowski, PhD†*

**Background:** Most public reporting and pay for performance (P4P) programs in the United States continue to be organized and implemented by single insurers. Adequate medical group-level reliability on clinical care process measures is possible in multi-stakeholder initiatives because patient samples can be pooled across payers. However, the extent to which reliable measurement is achievable in single insurer P4P initiatives remains unclear.

**Methods:** This study uses 7 years (2001 to 2007) of patient-level clinical care process data from an insurer in Washington State involving 20 medical groups. Eight clinical care process measures were analyzed. We compared the medical group-level reliability and resulting sample size requirements for each of the 8 measures using unadjusted and adjusted binary mixed models. The relation of baseline intraclass correlation coefficients (ICCs) and medical group performance change over time was examined for each clinical care process measure.

**Results:** Only 45% of all medical group measurements (group-years for all observations) had sufficient sample sizes to achieve reliable estimates of group performance. Measures with the largest deficiencies in patient samples per group included appropriate asthma treatment and low-density lipoprotein screening for patients with coronary artery disease. There was an inconsistent relationship between the size of baseline ICCs and medical group performance improvement over time.

**Conclusions:** Unreliable performance measurement is an important consequence of the prevailing organization and implementation of public reporting and P4P programs in the US. Multi-payer collaborations may be an important vehicle for ensuring reliable medical group performance measurement and comparisons on clinical care process measures.

From the *Department of Health Services, School of Public Health, University of California, Los Angeles, Los Angeles, CA; †Department of Health Services, School of Public Health, University of Washington; and ‡VA Health Services Research & Development Center of Excellence, Seattle, WA.

This study was supported by Robert Wood Johnson Foundation's Health Care Financing and Organization (RWJF-HCFO) Program (Grant # 63214).

The authors declare no conflict of interest.

Reprints: Hector P. Rodriguez, PhD, MPH, Department of Health Services, School of Public Health, University of California, Los Angeles, Box 951772, Los Angeles, CA 90095-1772. Email: hrod@ucla.edu.



ISSN: 0025-7079/12/5002-0117

**Key Words:** reliability, HEDIS, physician organizations, pay for performance

(*Med Care* 2012;50: 117–123)

## BACKGROUND

Integrated and network physician organizations (medical groups) are increasingly being persuaded by individual payers to make clinical quality information available to the public and to adopt financial incentives for improving the quality of patient care.[1,2] Medical groups incur high transaction costs when participating in public reporting and pay for performance (P4P) programs because most P4P programs are not coordinated efforts among payers.[2] Consequently, medical groups often respond to diverse reporting requirements and incentives for different payers, requiring costly systems for customized reporting to external parties.

Multistakeholder initiatives that standardize performance measures and reporting requirements are an important strategy for reducing the burden of disconnected payer reporting initiatives for medical groups. However, achieving the cooperation of commercial competitors has been a major challenge for many communities in the United States.[3–5] Some statewide initiatives manage to attain the consensus of commercial health plan and medical groups to support physician practice-level performance measurement and improvement. However, these large-scale performance improvement efforts have failed to produce breakthrough quality improvements.[6,7] Spectators are now skeptical that a natural evolution toward multipayer collaborations will emerge. Most public reporting and P4P programs in the US continue to be organized and implemented by single insurers and payers.

Single insurer public reporting and P4P initiatives assume that the available patient samples per medical group in a reporting year are sufficiently large to achieve stable and reliable estimates of medical group performance. When measuring and reporting clinical quality performance data, it is important to ensure that the appropriate organizational unit (individual physician, practice site, or medical group) is target for feedback and other improvement activities and that patient sample sizes per unit enable reliable measurement. Previous research has revealed that reliable physician performance measurement on clinical quality measures requires

patient sample sizes that often exceed those available in individual primary care physicians' panels,[8–11] medical groups and their practice sites are widely considered the most appropriate organizational units for measuring and reporting clinical quality measures.[12,13] Given that medical groups (vs. individual physicians or practice sites) have been the primary target of single insurer public reporting and P4P initiatives, the availability of adequate patient sample sizes for the reliable measurement and comparison of medical group performance is essential. The few studies that have examined the group-level reliability of widely used ambulatory care quality measures all employ data from multipayer and medical group initiatives[14,15] in which sufficient patient samples per organizational unit (medical group or practice site) are available. Pooled patient samples in multipayer and medical group initiatives can enable reliable measurement and performance comparisons.[14,16–21] To our knowledge, the reliability of medical group performance on clinical care process measures in the context of a single insurer public reporting and P4P initiative has not been examined.

A related challenge for public reporting and P4P initiatives is clarifying which quality of care measures are more likely to improve as a result of focused medical group-level interventions. Performance measures with large baseline intraclass correlation coefficients (ICCs) or intragroup correlations, are often assumed to be most amenable to intervention as a high proportion of explainable variation on the measure is attributable to groups (between-group effects).[21–26] However, there continues to be vigorous debate regarding whether or not low medical-group level reliability (or limited variation between medical groups) for certain clinical quality measures signals that group performance on these measures is unlikely to improve in the context of focused quality improvement initiatives.[17,18,27]

To inform the reliable measurement of medical group performance in the context of the predominate organization of P4P initiatives in the US, this study examines 2 important measurement issues in a single insurer's public reporting and P4P initiative (2001 to 2007). First, we examine the extent to which sufficiently reliable estimates of medical group performance on clinical care process measures can be achieved with the available patient sample sizes per medical group in a single insurer P4P initiative. We estimate the intragroup correlations for each of 8 clinical care process measures and estimate the patient sample sizes required to achieve sufficient medical group-level reliability ($\alpha_{GRP} = 0.70$, $\alpha_{GRP} = 0.80$). Second, we examine the extent to which clinical care process measures with larger baseline medical group effects, that is, larger intragroup correlations, improve more consistently over time compared with measures with smaller baseline group effects in the context of a public reporting and P4P program.

# METHODS

## Study Data and Measures

This study employs 7 years (2001 to 2007) of patient-level clinical care process data for patients of 20 medical groups from a single insurer in Washington State. The insurer's program used medical group-level quality-based financial incentives that ranged from a potential of 1% to 3% incremental contracted revenue from the insurer. Medical groups were all integrated physician organizations and operated a wide-range of practice sites (median = 5, range = 1 to 13), primary care physicians (median = 41.1, SD = 37.6, range = 6 to 123), and advanced practice clinicians (median = 15.0, SD = 20.4, range = 0 to 69) in 2007. In addition, the study insurer accounted for a mean of 19.5% (SD = 13.2) of participating medical groups' 2007 revenue, whereas other private or commercial patients accounted for a mean of 55.2% (SD = 13.2) of group revenue.

Eight clinical quality measures were analyzed for this study, all of which are Healthcare Effectiveness Data and Information Set clinical care process measures.[28] The measures included evidence-based treatment of asthma, coronary artery disease (CAD) and diabetes, screening for breast cancer and cervical cancer, and the provision of comprehensive well child visits (Table 1). A total of 197,905 person-years are represented by these data. The average annual number of patients eligible for 1 or more measurements was 2726 (SD = 1602) per medical group.

## Statistical Methods

### Estimation of Medical Group-Level Reliability

The proportion of variation attributable to medical group effects (or medical group-level reliability) was examined using the ICC for each of the 8 measures, where ICC = variation attributable to differences between physician groups ÷ (variation attributable to differences between physician groups+variation attributable to differences within physician group groups) and ranges from 0 to 1. Reliable discrimination of medical group performance is essential for informing consumer choice of providers and systems,[29,30] for performance feedback to physicians,[6,28] and for P4P initiatives.[31,32] Following Selby et al,[18] we estimated medical group random effects using unadjusted and adjusted binary mixed regression models. These regression models were estimated using the Newton-Raphson algorithm[33] because the calculation is robust to small and uneven cluster sizes, which are an important limitation in many single insurer quality improvement initiatives. For the adjusted models, patient age, sex, a modified Charlson comorbidity index,[34] and the measurement year, were specified as control variables when estimating medical group random effects, so that the ICC estimates for each measure account for patient case-mix differences between the medical groups.

### Calculation of Sample Size Requirements

Next, the Spearman-Brown Prophecy formula[35] was used to calculate the sample sizes required to achieve group-level reliability ($\alpha_{GRP}$) of 0.70 and 0.80, commonly used standards for unit-level reliability. Sample size estimates were compared against the mean annual available sample for each measure using the 3 different ICC estimation methods. Patient sample deficits/excess for each quality measure at $\alpha_{GRP} = 0.70$, $\alpha_{GRP} = 0.80$ are reported. The proportion of medical group measurements (group-years) that included sufficient observations for

**TABLE 1.** Medical Group Sample Sizes and Performance on Clinical Care Process Measures

| Measure | Inclusion Criteria | No. Groups Reporting Measure | Annual Eligible Patients Per Group | | Group Performance Level Across Years (Percentage of Patients Receiving Recommended Care) | |
|---|---|---|---|---|---|---|
| | | | Mean | SD | Mean | SD |
| Appropriate asthma medications | Children and adults identified with asthma who received prescription for long term control of asthma (inhaled corticosteroids, cromolyn sodium, nedocromil, leokotriene modifiers, methylxanthines)* | 19 | 65.0 | 49.5 | 0.85 | 0.07 |
| Breast cancer screening | Mammogram in the measurement year or one year prior for women 40–69† | 19 | 796.6 | 609.1 | 0.79 | 0.06 |
| Cervical cancer screening | PAP test within the measurement year or prior 2 y for women ages 21–64‡ | 19 | 919.2 | 681.3 | 0.83 | 0.04 |
| ACE inhibitor or ARB for diabetic patients with hypertension | Patients with diabetes screened during prior 12 mo | 20 | 192.5 | 147.1 | 0.50 | 0.05 |
| Glycosylated hemoglobin (HbA1c) screening for patients with diabetes | Patients with diabetes screened during prior 12 mo | 20 | 203.5 | 154.7 | 0.63 | 0.07 |
| LDL screening for patients with diabetes | Patients with diabetes screened during prior 12 mo | 20 | 203.5 | 154.7 | 0.76 | 0.08 |
| Comprehensive well child examination | Annual well-care visit for children ages 3–18. Must show evidence of health and development history (physical and mental), physical exam, health education/anticipatory guidance | 16 | 60.1 | 59.6 | 0.59 | 0.20 |
| LDL screening for patients with CAD | Patients with cardiovascular conditions screened during prior 12 mo | 6 | 86.9 | 75.7 | 0.83 | 0.07 |

*Exclusion: Patients with diagnosis of emphysema or chronic obstructive pulmonary disease (COPD).
†Exclusion: Women who have had bilateral mastectomy (may occur on the same or separate dates).
‡Exclusion: Women who have had a complete hysterectomy with no residual cervix.
ACE indicates angiotensin-converting enzyme; ARB, angiotensin receptor blockers; CAD, coronary artery disease; LDL, low-density lipoprotein.

obtaining acceptably reliable medical group performance estimates were calculated for each measure.

### Baseline Group Effects and Change Over Time

Finally, we examined the extent to which the clinical care process measures with larger medical group effects (higher ICCs) at baseline were, in fact, the measures that improved most in the context of the single-payer P4P initiative. Accordingly, we estimated the extent of change over time (2001 to 2007) for each group and measure. To do this, for each medical group and measure, we estimated logistic regression models that accounted for the patient age, sex, a modified Charlson comorbidity index,[34] and a continuous measure of time. Resulting continuous time coefficients from these regression models were classified as (1) worse performance over time and statistically significant, (2) worse performance over time and not significant, (3) better performance over time and not statistically significant, and (4) better performance over time and statistically significant. The relation of medical group improvements over time and baseline ICCs was examined qualitatively.

## RESULTS

Of the 8 clinical care process measures examined, comprehensive well child examinations (ICC = 0.06), glycosylated hemoglobin (HbA1c) screening (ICC = 0.03) for patients with diabetes, and low-density lipoprotein (LDL) screening (ICC = 0.03) for patients with diabetes had the largest ICCs. Most other measures had relatively smaller estimated medical group effects (ICCs < 0.01; Table 2). Most ICCs were larger in the adjusted models that accounted for patient case-mix compared with the unadjusted models. For example, the ICC estimates for several clinical care process measures (asthma medications, HbA1c and LDL screening for patients with diabetes, comprehensive well child examinations) increased substantially (by >50%) when adjusted models were used to estimate ICCs compared with unadjusted models that did not account for patient age, sex, and comorbidities. By contrast, the angiotensin-converting enzyme (ACE) inhibitor or angiotensin receptor blockers (ARB) measure for hypertensive diabetics had a significantly lower ICC once patient case-mix adjustments (ICC = 0.034 vs. 0.004) were implemented.

Table 3 details the annual sample size requirements for each of the 8 study measures, calculated using the unadjusted and adjusted ICC estimates. The estimated sample sizes requirements were compared with the median annual available patient sample size for medical groups. Sample size deficits were most pronounced for the appropriate asthma medication measure, the ACE inhibitor or ARB measure for hypertensive diabetics, and LDL screening for patients with CAD, where median sample size deficits ranged from 230 to

**TABLE 2.** Intraclass Correlation Coefficients Estimated Using Binary Mixed Regression Models Using the Newton-Raphson Algorithm, Unadjusted Versus Adjusted

| | Intraclass Correlation Coefficients | | |
|---|---|---|---|
| | A | B | |
| Measure | Mixed Models, Unadjusted | Mixed Models, Adjusted | Percent Change in ICC (B-A)/A |
| Appropriate asthma medications | 0.002 | 0.006 | 200 |
| Breast cancer screening | 0.009 | 0.011 | 22 |
| Cervical cancer screening | 0.010 | 0.014 | 40 |
| ACE inhibitor or ARB for diabetic patients with hypertension | 0.034 | 0.004 | −88 |
| Glycosylated hemoglobin (HbA1c) screening for patients with diabetes | 0.016 | 0.030 | 88 |
| LDL screening for patients with diabetes | 0.018 | 0.031 | 72 |
| Comprehensive well child examination | 0.038 | 0.062 | 63 |
| LDL screening for patients with CAD | 0.005 | 0.007 | 40 |

Adjusted models control for patient age, sex, a modified Charleson comorbidity score, and measurement year.
ACE indicates angiotensin-converting enzyme; ARB, angiotensin receptor blockers; CAD, coronary artery disease; LDL, low-density lipoprotein.

826 patients per year, depending on the reliability requirement ($\alpha_{GRP}=0.70$ vs. $\alpha_{GRP}=0.80$) and ICC estimation method (unadjusted vs. adjusted binary mixed models).

Only 45% of all medical group measurements (group-years for all observations) had sufficient sample sizes to achieve reliable estimates of group performance, irrespective of case-mix adjustment (Fig. 1). However, 70% or more of group measurements had sufficient patient samples for the cervical cancer screening (91%), breast cancer screening (77%), and LDL (71%) and HbA1c (70%) measures for patients with diabetes in adjusted models. By contrast, large patient sample deficits for medical groups included appropriate asthma medications (reliable estimates for 0% of medical group measurements), the ACE inhibitor or ARB measure for hypertensive diabetics (reliable for only 2% of medical group measurements), and LDL screening for patients with CAD (reliable estimates for only 3% of medical group measurements).

The largest baseline ICCs were observed for the LDL and HbA1c screening measures for patients with diabetes and the comprehensive well child examination measure (Table 4). Over the course of the study period (2001 to 2007) medical groups' performance improved most consistently for the diabetes care process measures. For example, HbA1c and LDL screening for patients with diabetes improved significantly over time for 6 and 13 of the 20 groups, respectively (Table 4). Importantly, the ACE inhibitor or ARB measure for hypertensive diabetics improved for 19 of 20 groups over the study period despite having a small baseline ICC.

**TABLE 3.** Sample Size Requirements for Clinical Quality Measures Estimated at Various Levels of Medical Group-Level Reliability

| Estimated Group-Level Reliability | | $\alpha_{GRP}$ = 0.70 | | | | $\alpha_{GRP}$ = 0.80 | | | |
|---|---|---|---|---|---|---|---|---|---|
| Intraclass Correlation Coefficient Estimation Method | | Mixed Models, Unadjusted | | Mixed Models, Adjusted | | Mixed Models, Unadjusted | | Mixed Models, Adjusted | |
| Clinical Care Process Measure | Annual Eligible Subjects per Group (Mean) | Sample Size Required | Deficit/ Excess | Sample Size Required | Deficit/ Excess | Sample Size Required | Deficit/ Excess | Sample Size Required | Deficit/ Excess |
| Appropriate asthma medications | 65 | 1207 | **−1142** | 409 | **−344** | 2068 | **−2003** | 701 | **−636** |
| Breast cancer screening | 797 | 270 | 527 | 213 | 584 | 463 | 334 | 364 | 432 |
| Cervical cancer screening | 919 | 227 | 692 | 164 | 756 | 389 | 530 | 280 | 639 |
| ACE inhibitor or ARB for diabetic patients with hypertension | 193 | 66 | 127 | 594 | **−401** | 113 | 80 | 1018 | **−826** |
| Glycosylated hemoglobin (HbA1c) screening for patients with diabetes | 204 | 142 | 61 | 76 | 127 | 244 | **−41** | 131 | 73 |
| LDL screening for patients with diabetes | 204 | 130 | 74 | 74 | 130 | 222 | **−19** | 127 | 77 |
| Comprehensive well child examination | 60 | 59 | 1 | 35 | 25 | 101 | **−41** | 61 | **−1** |
| LDL screening for patients with CAD | 87 | 441 | **−354** | 317 | **−230** | 756 | **−669** | 543 | **−456** |

Sample size deficits are noted in **bold**. Adjusted models control for patient age, sex, modified Charleson comorbidity score, and measurement year.
ACE indicates angiotensin-converting enzyme; ARB, angiotensin receptor blockers; CAD, coronary artery disease; LDL, low-density lipoprotein.

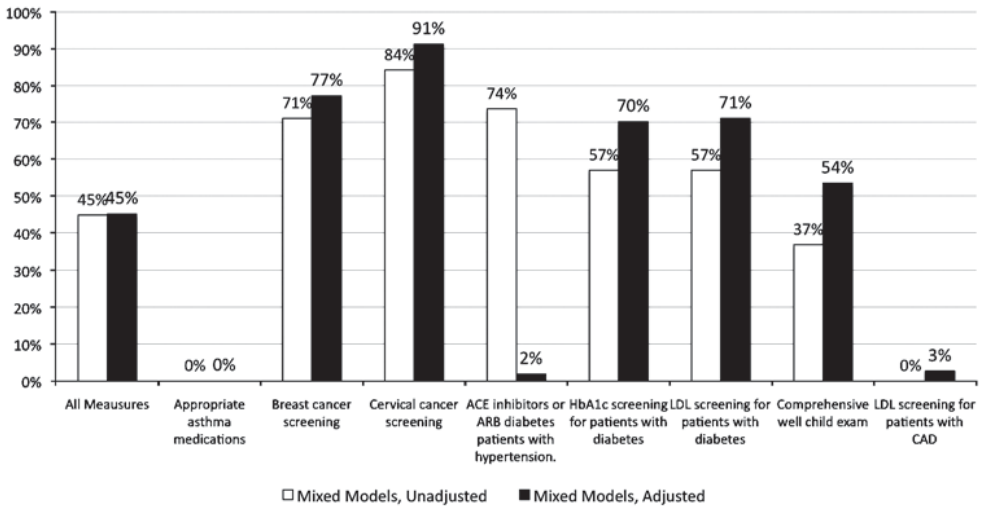


**FIGURE 1.** Proportion of clinic-years with sufficient patient samples for reliable medical group estimates, by measure. Note: Medical group-level reliability estimated at 0.70.

## DISCUSSION

The study results underscore that, even for a large insurer with substantial market share, reliable measurement of medical group performance on high volume clinical care processes is challenging. Only 45% of the clinic-year observations across measures had sufficient patient counts for acceptable ($MD_{GRP} = 0.70$) medical group reliability using robust ICC estimation procedures (adjusted binary mixed models). Our results indicate that some clinical care process measures with larger numbers of available patients per group—including breast cancer screening, cervical cancer screening, and diabetes care process measures—can reasonably be measured in a single insurer initiative measuring and comparing medical group performance. In contrast, this may not hold true for the other measures examined in the study.

Given the challenges with measurement precision, multistakeholder collaborations of payers, medical groups, and key community partners may be an important vehicle for ensuring reliable medical group performance measurement and comparisons on clinical care process measures because patient samples can be pooled. Previous research has underscored that physicians are concerned about the accuracy and reliability of data used in public reporting and P4P programs[36,37] and our study results corroborate this concern. Although initiatives are easier to implement for insurers compared with multistakeholder collaborations, the high administrative costs associated with developing separate reporting systems and requirements for payers' public reporting and P4P requirements can be burdensome for medical groups.[4,38] Our study provides evidence that unreliable measurement of medical group performance for clinical care process measures is another important consequence of the prevailing organization and implementation of public reporting and P4P programs in the US.

**TABLE 4.** The Relation of Intraclass Correlation Coefficients and Group Performance Change Over Time

| | | | No. Medical Groups | | | |
|---|---|---|---|---|---|---|
| **Clinical Care Process Measure** | **Baseline Adjusted ICC*** | **Total Groups Reporting Measure** | **Worse Performance Over Time, Statistically Significant** | **Worse Performance Over Time, Not Significant** | **Improved Performance Over Time, Not Significant** | **Improved Performance Over Time, Statistically Significant** |
| Appropriate asthma medications | 0.011 | 19 | 9 | 4 | 4 | 2 |
| Breast cancer screening | 0.011 | 19 | 5 | 6 | 5 | 3 |
| Cervical cancer screening | 0.002 | 19 | 6 | 3 | 4 | 6 |
| ACE inhibitor or ARB for diabetic patients with hypertension | 0.004 | 20 | 0 | 0 | 1 | 19 |
| Glycosylated hemoglobin (HbA1c) screening for patients with diabetes | 0.040 | 20 | 0 | 3 | 11 | 6 |
| LDL screening for patients with diabetes | 0.11 | 20 | 1 | 1 | 5 | 13 |
| Comprehensive well child examination | 0.14 | 16 | 5 | 1 | 1 | 9 |
| LDL screening for patients with CAD | 0.007 | 6 | 0 | 0 | 3 | 3 |

*Intraclass correlation coefficients (ICC)s reflect baseline (2001) year estimates using adjusted binary mixed models.
ACE indicates angiotensin-converting enzyme; ARB, angiotensin receptor blockers; CAD, coronary artery disease; LDL, low-density lipoprotein.

Our results also indicate that substantial improvements in measurement precision can be achieved through statistical adjustment. With the exception of the ACE inhibitor or ARB measure for hypertensive diabetics, adjusting the regression models for patient age, sex, and comorbidities improved the precision of medical group performance measurement on clinical care process measures. Previous studies have underscored the importance of case-mix adjustment for quality of care comparisons[16,39] and our results are consistent with these analyses. Many initiatives, particularly single insurer quality improvement efforts, face challenges with small and/or uneven observations across medical groups, so the use of scoring methods that are robust to these challenges, that is, adjusted binary mixed models, is recommended.

Finally, there was an inconsistent relationship between baseline ICCs and medical group performance improvements over time across the study measures. A recent study in an integrated delivery system found that quality of care measures with relatively smaller facility-level ICCs (<0.005), including mammography screening and patient care experience measures, significantly improved over time in the context of facility level improvement efforts.[18] In our study, the ACE inhibitor or ARB measure for hypertensive diabetics improved for 19 of 20 medical groups over the study period despite having a low baseline ICC (0.004). The diabetes care process measures (with higher ICCs) also improved consistently, whereas improvements for other clinical care process measures (which had a wide range of baseline ICCs) were inconsistent. Taken together, the results suggest that focused diabetes care improvement efforts made by medical groups may have resulted in overall improvement on the diabetes care process measures despite a low baseline ICCs for one of the measures. Importantly, other levels of organization (practice site or individual physician) may account for a higher proportion of variability and change over time compared with the medical level and these organizational units might be the more appropriate target for measurement, reporting, and quality improvement initiatives. However, pooled patient samples across multiple insurers are necessary achieve reliable performance measurement of smaller organizational units (practice sites or individual physicians).

Our study results should be considered in light of important limitations. Our sample consists of patients from 1 insurer in Washington state. Thus, testing consistency of results across insurers is not possible. Nevertheless, public reporting and P4P initiatives continue to be organized and implemented by single insurers. Consequently, our results have substantial practical utility. Second, our medical group sample includes 20 integrated medical groups and no independent practice associations (network organizations). ICC estimates for the study measures might even be smaller for independent practice association samples, given their more limited influence on practice organization.[40–42] The sample size requirements to achieve reliable group comparisons would likely be even greater for less integrated physician organizations. Third, we limited our investigation to clinical care process measures because the measures were subject to public reporting and performance-based financial incentives. Intermediate outcomes tend to have lower group-level reliability, so the inclusion of intermediate outcome measures would not likely alter our conclusion that reliable measurement of medical group performance in a single insurer initiative is challenging.

Collaborative efforts between multiple payers and insurers are 1 important strategy for improving the reliability of medical group performance measurement on quality of care measures. The challenges of interorganzational collaboration in a competitive environment[6,7] coupled with other practical implementation barriers impede the development of multistakeholder initiatives in the US. Our results indicate that the fragmented organization and implementation of public reporting and P4P initiatives runs a high risk of unreliable measurement of medical group performance. As a natural evolution toward coordinated multistakeholder efforts has not occurred, future research should examine the extent to which composite measures of care quality can improve group-level reliability in single insurer initiatives so that patient sample size requirements might be reduced. Measurement advances that enable reliable performance measurement might also enhance the feasibility of multipayer and multiorganizational collaboration.

## ACKNOWLEDGMENT

*The authors thank Justin W. Timbie for his helpful feedback.*

## REFERENCES

1. Marshall MN, Shekelle PG, Leatherman S, et al. The public release of performance data: what do we expect to gain? A review of the evidence. *J Am Med Assoc*. 2000;283:1866–1874.
2. Conrad DA, Perry L. Quality-based financial incentives in health care: can we improve quality by paying for it? *Ann Rev Public Health*. 2009;30:357–371.
3. Rosenthal MB, Frank RG. What is the empirical basis for paying for quality in health care? *Med Care Res Rev*. 2006;63:135–157.
4. Conrad DA, Saver BG, Court B, et al. Paying physicians for quality: evidence from the field. *Jt Comm J Qual Patient Saf*. 2006;32:443–451.
5. Friedberg MW, Safran DG, Coltin K, et al. Paying for performance in primary care: potential impact on practices and disparities. *Health Aff (Millwood)*. 2010;29:926–932.
6. Damberg CL, Raube K, Teleki SS, et al. Taking stock of pay-for-performance: a candid assessment from the front lines. *Health Aff (Millwood)*. 2009;28:517–525.
7. Pearson SD, Schneider EC, Kleinman KP, et al. The impact of pay-for-performance on health care quality in Massachusetts, 2001-2003. *Health Aff (Millwood)*. 2008;27:1167–1176.
8. Scholle SH, Roski J, Adams JL, et al. Benchmarking physician performance: reliability of individual and composite measures. *Am J Manag Care*. 2008;14:833–838.
9. Hofer TP, Hayward RA, Greenfield S, et al. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *J Am Med Assoc*. 1999;281:2098–2105.
10. Greenfield S, Kaplan SH, Kahn R, et al. Profiling care provided by different groups of physicians: effects of patient case-mix (bias) and physician-level clustering on quality assessment results. *Ann Intern Med*. 2002;136:111–121.
11. Scholle SH, Roski J, Dunn DL, et al. Availability of data for measuring physician quality performance. *Am J Manag Care*. 2009;15:67–72.
12. Shortell SM, Casalino LP. Health care reform requires accountable care systems. *J Am Med Assoc*. 2008;300:95–97.
13. Rittenhouse DR, Casalino LP, Gillies RR, et al. Measuring the medical home infrastructure in large medical groups. *Health Aff (Millwood)*. 2008;27:1246–1258.

 

14. Sequist TD, Schneider EC, Li A, et al. Reliability of Medical Group and physician performance measurement in the primary care setting. *Med Care*. 2010;49:126–131.
15. Safran DG, Karp M, Coltin K, et al. Measuring patients' experiences with individual primary care physicians. Results of a statewide demonstration project. *J Gen Intern Med*. 2006;21:13–21.
16. Huang IC, Dominici F, Frangakis C, et al. Is risk-adjustor selection more important than statistical approach for provider profiling? Asthma as an example. *Med Decis Making*. 2005;25:20–34.
17. Fung V, Schmittdiel JA, Fireman B, et al. Meaningful variation in performance: a systematic literature review. *Med Care*. 2011;48: 140–148.
18. Selby JV, Schmittdiel JA, Lee J, et al. Meaningful variation in performance: what does variation in quality tell us about improving quality? *Med Care*. 2011;48:133–139.
19. Littenberg B, MacLean CD. Intra-cluster correlation coefficients in adults with diabetes in primary care practices: the Vermont Diabetes Information System field survey. *BMC Med Res Methodol*. 2006;6:20.
20. Fuhlbrigge A, Carey VJ, Finkelstein JA, et al. Are performance measures based on automated medical records valid for physician/ practice profiling of asthma care? *Med Care*. 2008;46:620–626.
21. Huang IC, Diette GB, Dominici F, et al. Variations of physician group profiling indicators for asthma care. *Am J Manag Care*. 2005;11:38–44.
22. Hayward RA, Manning WG Jr. McMahon LF Jr., et al. Do attending or resident physician practice styles account for variations in hospital resource use? *Med Care*. 1994;32:788–794.
23. Katon W, Rutter CM, Lin E, et al. Are there detectable differences in quality of care or outcome of depression across primary care providers? *Med Care*. 2000;38:552–561.
24. Tuerk PW, Mueller M, Egede LE. Estimating physician effects on glycemic control in the treatment of diabetes: methods, effects sizes, and implications for treatment policy. *Diabetes Care*. 2008;31: 869–873.
25. O'Connor PJ, Rush WA, Davidson G, et al. Variation in quality of diabetes care at the levels of patient, physician, and clinic. *Prev Chronic Dis*. 2008;5:A15.
26. Krein SL, Hofer TP, Kerr EA, et al. Whom should we profile? Examining diabetes care practice variation among primary care providers, provider groups, and health care facilities. *Health Serv Res*. 2002;37: 1159–1180.
27. Holmboe ES, Weng W, Arnold GK, et al. The comprehensive care project: measuring physician performance in ambulatory practice. *Health Serv Res*. 2010;45:1912–1933.
28. Friedberg MW, Coltin KL, Safran DG, et al. Associations between structural capabilities of primary care practices and performance on selected quality measures. *Ann Intern Med*. 2009;151:456–463.
29. Fanjiang G, von Glahn T, Chang H, et al. Providing patients web-based data to inform physician choice: If you build it, will they come? *J Gen Intern Med*. 2007;22:1463–1467.
30. Fung CH, Elliott MN, Hays RD, et al. Patients' preferences for technical versus interpersonal quality when selecting a primary care physician. *Health Serv Res*. 2005;40:957–977.
31. Roland M, Elliott M, Lyratzopoulos G, et al. Reliability of patient responses in pay for performance schemes: analysis of national General Practitioner Patient Survey data in England. *Br Med J*. 2009;339:b3851.
32. Browne K, Roseman D, Shaller D, et al. Measuring patient experience as a strategy for improving primary care. *Health Aff (Millwood)*. 2010;29:921–925.
33. Kaplan SH, Griffith JL, Price LL, et al. Improving the reliability of physician performance assessment: identifying the "physician effect" on quality and creating composite measures. *Med Care*. 2009;47:378–387.
34. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol*. 1992;45:613–619.
35. Nunnelly J, Bernstein I. *Psychometric Theory*. New York: McGraw-Hill; 1994.
36. Meterko M, Young GJ, White B, et al. Provider attitudes toward pay-for-performance programs: development and validation of a measurement instrument. *Health Serv Res*. 2006;41:1959–1978.
37. Bokhour BG, Burgess JF Jr. Hook JM, et al. Incentive implementation in physician practices: a qualitative study of practice executive perspectives on pay for performance. *Med Care Res Rev*. 2006;63:73S–95S.
38. Damberg CL, Raube K, Williams T, et al. Paying for performance: implementing a statewide project in California. *Qual Manag Health Care*. 2005;14:66–79.
39. Zaslavsky AM, Epstein AM. How patients' sociodemographic characteristics affect comparisons of competing health plans in California on HEDIS quality measures. *Int J Qual Health Care*. 2005;17:67–74.
40. Weeks WB, Gottlieb DJ, Nyweide DE, et al. Higher health care quality and bigger savings found at large multispecialty medical groups. *Health Aff (Millwood)*. 2010;29:991–997.
41. Mehrotra A, Epstein AM, Rosenthal MB. Do integrated medical groups provide higher-quality medical care than individual practice associations? *Ann Intern Med*. 2006;145:826–833.
42. Casalino LP. Which type of medical group provides higher-quality care? *Ann Intern Med*. 2006;145:860–861.


 