DOI: 10.1377/hlthaff.2019.01061
HEALTH AFFAIRS 39,
NO. 5 (2020): 852–861

By Amol S. Navathe, Kevin G. Volpp, Amelia M. Bond, Kristin A. Linn, Kristen L. Caldarella, Andrea B. Troxel, Jingsan Zhu, Lin Yang, Shireen E. Matloubieh, Elizabeth E. Drye, Susannah M. Bernheim, Emily Oshima Lee, Mark Mugiishi, Kimberly Takata Endo, Justin Yoshimoto, and Ezekiel J. Emanuel

Assessing The Effectiveness Of Peer Comparisons As A Way To Improve Health Care Quality

Amol S. Navathe (amol.navathe@gmail.com) is a core investigator at the Corporal Michael J. Cresencz Veterans Affairs (VA) Medical Center; and an assistant professor in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, and a senior fellow at the Leonard Davis Institute of Health Economics, University of Pennsylvania, all in Philadelphia.

Kevin G. Volpp is a professor of medicine in the Department of Medicine at the Perelman School of Medicine and of health care management at the Wharton School, vice chair for health policy in the Department of Medical Ethics and Health Policy, and director of the Center for Health Incentives and Behavioral Economics, all at the University of Pennsylvania, and a staff physician at the Corporal Michael J. Crescenz VA Medical Center.

Amelia M. Bond is an assistant professor of health care policy and research at Weill Cornell Medical College, in New York City.

Kristin A. Linn is an assistant professor of biostatistics in the Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania.

Kristen L. Caldarella is a project manager in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania.

ABSTRACT Policy makers are increasingly using performance feedback that compares physicians to their peers as part of payment policy reforms. However, it is not known whether peer comparisons can improve broad outcomes, beyond changing specific individual behaviors such as reducing inappropriate prescribing of antibiotics. We conducted a cluster-randomized controlled trial with Blue Cross Blue Shield of Hawaii to examine the impact of providing peer comparisons feedback on the quality of care to primary care providers in the setting of a shift from fee-for-service to population-based payment. Over 74,000 patients and eighty-eight primary care providers across sixty-three sites were included over a period of nine months in 2016. Patients in the peer comparisons intervention group experienced a 3.1-percentage-point increase in quality scores compared to the control group—whose members received individual feedback only. This result underscores the effectiveness of peer comparisons as a way to improve health care quality, and it supports Medicare's decisions to provide comparative feedback as part of recently implemented primary care and specialty payment reform programs.

In the quest for a health care system that maximizes the value of care, policy makers have used various types of incentives directed at clinicians. However, most direct pay-for-performance incentives that were focused on quality improvement have produced mixed results.[1–5] Consequently, policy makers recently began using behavioral science, including principles from psychology and behavioral economics, to guide the design of financial and nonfinancial interventions. One promising nonfinancial strategy for improving quality is to provide feedback to clinicians about their performance relative to that of their peers.[6–8]

Recent programs introduced by Medicare include peer comparisons feedback together with changes in payments. Examples include the Comprehensive Primary Care Plus program and the Oncology Care Model, both of which provide peer comparisons feedback to physician groups.[9,10] Other payers have also done so as part of payment programs.

Policy makers are enthusiastic about using peer comparisons to stimulate practice change for good reason, which is supported by the psychology of clinician behavior. The psychology principle of relative social ranking states that people care about how they compare to others in close proximity or within the same group.[11–13] Intervention designs that use this principle may be particularly effective among clinicians. Would-be physicians must navigate a rigorous selection process throughout medical training, and the process of selecting students for medical school and residency is based on choosing those who excel in various types of peer comparisons. In addition, professional norms are strong in medicine: Professional societies and certification boards disseminate guidelines that are frequently used to judge clinicians' practice.[12]

However, despite the nationwide rolling out of programs that emphasize broad outcomes of health care costs and quality, peer comparisons have mostly been tested in clinical contexts that focus on improving a single behavior—such as decreasing the inappropriate prescribing of antibiotics, opioids, or antipsychotics.[6–8,14–17] These discrete decisions inherently have a narrow scope, and the application of peer comparisons to broader settings and outcome metrics may yield different results. Furthermore, these interventions were compared to "no feedback" as a control, not to feedback on individual performance. Thus, policy makers seeking to design future programs lack evidence on the impact of peer comparisons on broader measures of quality or cost and their effect relative to that of providing individual feedback alone.

In this study we worked with the Hawaii Medical Services Association (HMSA) (that is, Blue Cross Blue Shield of Hawaii) to compare the effectiveness in improving the quality of care provided by primary care providers of individual feedback alone versus such feedback accompanied by peer comparisons.

Study Data And Methods

SETTING The HMSA has a majority market share within Hawaii. The health care market in Hawaii is generally fragmented, with no large integrated delivery systems for primary care that have long experience with risk-based contracting. Years of high annual growth in health care costs led the leaders of the association to seek a new primary care payment system.

We conducted a cluster-randomized controlled trial among eighty-eight primary care providers and four primary care organizations that participated in the new payment system, called the Population-Based Payments for Primary Care system. Participating primary care providers and organizations served patients in the four different plan types offered by the HMSA: Medicare Advantage, Medicaid managed care, commercial health maintenance organizations (HMOs), and non-HMO commercial plans.[18]

PARTICIPANTS Patient inclusion criteria included enrollment in at least nine of the twelve months in 2015 in a commercial HMO, non-HMO commercial plan, or Medicaid managed care plan and attribution to a primary care organization and provider (based on the attribution logic defined in protocol exhibit 1 in the online appendix)[19] that participated in the first wave of the Population-Based Payments for Primary Care system. Patients of primary care providers in the following specialties were eligible for inclusion: internal medicine, general practice, family medicine, pediatrics, advanced practice registered nurses (APRNs), and physician assistants.

STUDY DESIGN Details of the study design and interventions are presented below and in the protocol supplement in the appendix.[19] In summary, we collaborated with the HMSA to test the impact of a peer comparisons intervention on quality within the initial pilot test of the Population-Based Payments for Primary Care system for primary care providers and organizations in Hawaii.[20] All primary care providers and organizations in this study were among the first group of participants in the system.

We randomly assigned patients based on their attributed primary care provider to one of three groups: a control group in which providers received individual feedback alone, an intervention group whose members received peer comparisons feedback (which by necessity also included individual feedback), and an intervention group whose members received peer comparisons feedback and a shared patient–primary care provider incentive for glycemic control among patients with diabetes. The effectiveness of the interventions was assessed after nine months.

This pragmatic study was part of a broader quality improvement initiative led by HMSA leaders. The leaders and the study authors conducted in-person sessions with clinician and administrative leaders from all eligible primary care organizations to describe study procedures and answer questions. The study protocol, including a waiver of informed consent for patients and physicians, was approved by the Institutional Review Board at the University of Pennsylvania (for the study protocol, see the appendix).[19]

INTERVENTIONS

▸ **GROUP 1, INDIVIDUAL FEEDBACK ALONE:** In the first group, primary care providers and organizations received a dashboard with feedback on individual performance on cost and quality metrics, but they were not "pushed" any information by email (see protocol exhibit 2 in the appendix).[19] The dashboard was designed to simplify interpretation of the information, increase its salience, and create a stronger tie between actions and resulting improvements in outcomes. Patients of primary care providers in this group did not participate in any other intervention and served as an "active" control for the combined peer comparisons groups.

▸ **GROUP 2, PEER COMPARISONS FEEDBACK:** In the second group, primary care providers and organizations received peer comparisons in addition to individual feedback. Each organization's and clinician's performance was highlighted on cost and quality metrics included in the Population-Based Payments for Primary Care

Andrea B. Troxel is director of the Division of Biostatistics, New York University School of Medicine, in New York City.

Jingsan Zhu is associate director of data analytics in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania.

Lin Yang is a programmer analyst in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania.

Shireen E. Matloubieh is a research coordinator in the Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania.

Elizabeth E. Drye is a research scientist in the Department of Pediatrics, Yale University School of Medicine, in New Haven, Connecticut.

Susannah M. Bernheim is director of quality measurement at the Center for Outcomes Research and Evaluation at Yale–New Haven Hospital and an assistant clinical professor in the Department of Internal Medicine at Yale University School of Medicine.

Emily Oshima Lee is assistant vice president of health strategy at the Hawaii Medical Services Association (HMSA), in Honolulu.

Mark Mugiishi is the president and CEO of the HMSA.

Kimberly Takata Endo is a health strategist in the Department of Payment Transformation, HMSA.

Justin Yoshimoto is a health strategist in the Department of Payment Transformation, HMSA.

Ezekiel J. Emanuel is the Diane V. S. Levy and Robert M. Levy University Professor, chair of the Department of Medical Ethics and Health Policy, and vice provost for global initiatives, all at the University of Pennsylvania.

system, with a histogram or scatterplot used to depict performance relative to that of other organizations and providers (dashboard examples are in protocol exhibits 3 and 4 in the appendix).[19] This feedback was delivered by email weekly for four weeks and then biweekly for thirty-three weeks (for the schedule, see the protocol supplement, section 5, Study Interventions, in the appendix),[19] starting April 1, 2016. The email messages were sent to the primary care providers with performance figures for two to three measures (which varied for each mailing date) and a link to the dashboard log-in page. In this fashion, the peer comparisons intervention also included the "push" of the comparative data to the primary care providers. The peer comparisons dashboard was intended to use the behavioral principle of relative social ranking as a motivator to improve performance on quality and cost metrics.

▶ **GROUP 3, PEER COMPARISONS FEEDBACK PLUS SHARED DIABETES INCENTIVE:** In the third group, in addition to the peer comparisons intervention delivered to the providers in group 2, patients with poorly controlled diabetes and their attributed primary care providers were each eligible for $75 three and six months after enrollment in the program if the patient's hemoglobin A1c went down by at least 0.5 points from baseline or achieved a value of 9.0 or lower (see protocol exhibits 5 and 6 in the appendix).[19]

However, the third group failed to engage patients adequately in baseline HbA1c testing, so participation in the shared patient–primary care provider incentive was low: Only forty patients (13 percent of eligible patients) completed baseline testing and received the associated incentives. Consequently, we combined the two intervention groups and analyzed them together as the combined peer comparisons feedback group (for an analysis using the original design, see the appendix).[19] The group with individual feedback alone served as the control group, highlighting the differences between the group whose primary care providers received individual feedback alone and that whose providers also received peer comparisons feedback. To align with this change, the primary outcome was also changed from improvement in HbA1c among patients with poorly controlled diabetes to a composite quality score among all patients across metrics included in the Population-Based Payments for Primary Care system. Both of these changes to the analytic plan were made to the study protocol after the start of the trial, but before analysis was started (see the protocol supplement in the appendix).[19]

RANDOMIZATION Patients were randomly assigned by attributed primary care provider to the control group or one of the two intervention groups in a 1:1:1 ratio, stratified by primary care organization. Study participants and operational staff members were not blinded to group assignment, because knowledge of the intervention and access to the peer comparisons data were essential to the intervention's mechanism, but the study authors remained blinded until all follow-up data were obtained and primary analyses were finalized.

OUTCOMES The primary outcome was the 2016 composite quality score that indicated the probability of achieving a quality measure for which a patient was eligible. The composite quality score included thirteen pooled individual quality measures based on those in the Healthcare Effectiveness Data and Information Set that had also been incentivized in the HMSA's prior quality program. Thus, we had available the pre- and post-intervention data required to adjust for preintervention performance on measures (see the protocol supplement in the appendix).[19] An improvement in quality would require the mean probability of achievement to increase across all eligible measures, not just a single measure.

Only measures for which the patient was eligible were included in the analysis, and eligibility was defined by patient characteristics and diagnosis. For example, diabetes measures were restricted to patients with diabetes.

Secondary outcomes included the probability of achieving each individual quality measure, spending on primary care services, and HbA1c for eligible patients with poorly controlled diabetes (as defined by the protocol in the appendix).[19]

BASELINE VARIABLES Control variables included characteristics of the primary care provider (age, sex, practice site location in an urban ZIP code, specialty, location of residency in Hawaii versus another US state, and medical school in Hawaii versus another US state or another country), primary care provider panel (plan type mix, sex mix, number of attributed patients, average patient age, and average Episode Risk Group score), and patient characteristics (age, sex, Episode Risk Group score, plan type, residence in an urban ZIP code or one whose population had low education or low income levels, interactions between age and sex, and the proportion of eligible quality measures achieved in 2015).[21] A commercially available risk score, the Episode Risk Group score is intended to stratify individuals based on their predicted health care use and spending.[21] In this study the score was calculated using baseline 2014–15 data.

STATISTICAL ANALYSIS Although randomization occurred at the level of the primary care provider, the unit of analysis was the patient-measure (that is, each measure for which each patient was eligible) for all primary and secondary outcomes. We used 2015 baseline attribution in the 2016 intervention year to remove the chance that changes in patient attribution to primary care providers could confound the analysis (that is, ours was an intention-to-treat design). The primary analysis used a linear probability model to estimate the probability of achieving quality measures for which each patient was eligible (of the thirteen individual measures included in the study) in 2016.[20–23] The model adjusted for the baseline variables described above and for fixed-effects indicators for each quality measure, the baseline proportion of eligible measures achieved in 2015 by the patient (to increase statistical power), and the treatment-group indicator (to give the primary effect of interest). Standard errors were clustered by primary care provider to account for repeated measures from cluster randomization at the level of the primary care provider and used the Huber-White correction with an independent working correlation structure.[20–22,24,25] All hypothesis tests used an α of 0.05 and were two-sided. We adjusted for multiple testing in a secondary analysis of individual measure performance that used the Holm-Bonferroni correction.[26]

Secondary outcomes for primary care spending and HbA1c results were not adjusted for multiple testing. We estimated the risk-standardized proportion of evidence-based measures achieved to display findings on the original scale of the data.

We conducted additional analyses that examined the effect of the intervention on patients enrolled in Medicare Advantage or Medicaid managed care (relative to those enrolled in commercial HMO or non-HMO plans) and patients whose primary care providers had lower versus higher baseline quality (by adding an interaction between the intervention arm and baseline quality scores). These additional analyses were not prespecified in the study protocol (see the study protocol in the appendix).[19]

We used multiple imputation to impute individual quality measure values for approximately 9 percent of patients with missing follow-up data on quality measures[27] (see appendix exhibit 1).[19] We also conducted sensitivity analyses that used complete case data, an exchangeable covariance structure, and pairwise group comparisons in the original three-group design to examine the robustness of the results.

The pragmatic trial included eighty-eight primary care providers and sixty-three sites across the three groups and had 80 percent power to detect a difference of 3 percentage points (appendix exhibits 1 and 2).[19] Analyses were conducted using SAS, version 9.4.

LIMITATIONS This study had several limitations. First, given the pragmatic design in which primary care providers in the same organization were randomly assigned to either an intervention or a control group, there may have been contamination between the groups. However, we accepted this risk since peer comparisons interventions work best with "local" comparators with whom individuals identify. Furthermore, had there been contamination, it would have biased the effect toward the null.

Second, our initial study design used a third group with a shared patient–primary care provider incentive, but that group did not receive enough participation. Thus, we analyzed this group together with the peer comparisons group. This could have introduced bias, since a few patients did receive incentives. However, our primary study design retained randomization because members were still cluster-randomized to individualized feedback alone or to that and peer comparisons feedback, and the pairwise analysis results using the original design were similar to those in the final study. Although the study authors remained blinded, the imbalance in group size because of the third group's inclusion may have effectively unblinded the analysis. However, the analysis did not reveal differences in the HbA1c outcome by group.

Third, while the thirteen individual quality measures were selected to reduce choice overload, they did not comprehensively assess quality. They might not have fully reflected quality improvements if primary care providers improved their documentation of preexisting adherence instead of increasing their quality measure achievement.

Fourth, time-varying member eligibility for measures could have introduced bias, though our sensitivity analysis did not reveal this.

Fifth, this study did not evaluate the cost-effectiveness of implementing the peer comparisons intervention.

Sixth, the study did not measure the heterogeneity of intervention effects by all dimensions of type of practice or clinician or physician organization characteristics.

Finally, the study results might not be generalizable to other settings or other peer comparisons interventions.

Study Results

SAMPLE CHARACTERISTICS Of the 74,778 patients and 88 primary care providers in our sample, 28,249 patients and 29 primary care providers were in the control group, and 46,529 patients and 59 providers were in the peer comparisons intervention combined group (exhibits 1 and 2) (see also appendix exhibit 2).[19] The groups exhibited small differences in average age, sex, plan type (Medicare Advantage, Medicaid managed care, or commercial HMO or non-HMO), Episode Risk Group scores, utilization characteristics, and socioeconomic characteristics of ZIP codes of residence (exhibit 1). They exhibited few differences in characteristics of primary care providers—mainly geography, size of the primary care organization, and panel mix by health plan type (exhibit 2).

QUALITY An adjusted analysis indicated that the peer comparisons intervention group had a 3.1-percentage-point higher composite quality score than the control group did ($p = 0.048$) (exhibit 3).

A secondary analysis of individual measures that adjusted for multiple comparisons indicated that compared to the control group, the combined peer comparisons intervention group did significantly better at meeting the breast cancer screening (3.7 percentage points), diabetes care eye exam (7.4 percentage points), and diabetes care medical attention for nephropathy (3.2 percentage points) measures. Many other measures demonstrated trends toward differential improvement, but those effects were not significant. Two measures, immunization status for children and for adolescents, had point estimates that indicated worse performance in the intervention group (declines of 8.8 percentage points and 1.2 percentage points, respectively).

These results indicate that an additional 445 women were estimated to have been screened for breast cancer, an additional 500 patients with diabetes received an evidence-based eye exam, and an additional 214 patients with diabetes appropriately received screening for nephropathy (that is, achieved the medical attention for nephropathy measure).

SECONDARY OUTCOMES Secondary analyses showed that neither primary care spending nor HbA1c levels changed (appendix exhibits 3 and 4).[19] An additional analysis indicated that there were no differences in the effects of peer comparisons on quality of care for patients enrolled in Medicare Advantage or commercial plans. Primary care providers with lower baseline quality improved 0.5 percentage points ($p = 0.003$) more than those with higher baseline quality did (appendix exhibits 5–9).[19]

Sensitivity analyses that used complete cases and exchangeable covariance structures had similar results. Comparisons among the three original randomization groups (individual feedback only, individual and peer comparisons feedback, and both types of feedback plus the shared patient–primary care provider incentive for diabetes control) demonstrated similar improvements for the peer comparisons intervention (appendix exhibits 10–16).[19]

Discussion

This randomized trial demonstrated that adding peer comparisons feedback to individual feedback increased quality scores by 3.1 percentage points among physicians in Hawaii. The study intervention was designed based on the behavioral science principle of social comparisons—specifically, relative social ranking—and the results demonstrate the effectiveness of peer comparisons in motivating clinicians to use higher-value practice patterns. There are six key implications.

CLINICALLY MEANINGFUL IMPROVEMENT First, peer comparisons can improve quality performance in clinically meaningful ways. As mentioned above, in this study population an additional 445 women were estimated to have been screened for breast cancer, an additional 500 patients with diabetes received an evidence-based eye exam, and an additional 214 patients with diabetes appropriately received screening for nephropathy. Furthermore, given the reasonably high baseline performance of 79 percent of the eligible measures achieved, designing and implementing an intervention that led to a 3.1-percentage-point difference in a composite quality score was challenging. Large-scale payment changes—such as those in Medicare's Comprehensive Primary Care initiative, which provided monthly care management payments of $8–$40 per beneficiary—did not achieve improvements in process or outcome quality metrics, including measures in this study (for example, eye exams and nephropathy screening among patients with diabetes).[28] The Alternative Quality Contract of Blue Cross Blue Shield of Massachusetts was associated with similar increases of 5.1 percentage points and 2.3 percentage points, respectively, for eye exams and nephropathy screening among people with diabetes, but these increases were not significant.[22] Other health plan and health system programs that directly targeted financial incentives at physicians have generated mixed results and generally have not consistently led to improvements in the range of 3 percentage points.[1–5,29]

CONTEXT OF PAYMENT CHANGES Second, the peer comparisons intervention effect could have been influenced by its implementation in the setting of a payment change, similar to the way in which new payment programs being instituted by Medicare and other payers are using peer comparisons. This intervention was tested in the

EXHIBIT 1

Characteristics of patients covered by the Hawaii Medical Services Association in the study sample, by primary care provider (PCP) trial group, 2016

Variable	Overall (*N* = 74,778)	Group 1 (*n* = 28,249)	Groups 2 and 3 together (*n* = 46,529)	Group 2 only (*n* = 28,050)	Group 3 only (*n* = 18,479)
Age, years					
Less than 18	18%	25%	14%	21%	4%
18–34	16	16	17	16	18
35–49	21	19	23	21	25
50–64	25	23	26	24	29
65 or more	19	17	20	18	25
Sex					
Female	53%	54%	52%	53%	50%
Male	48	47	48	47	50
Health plan type					
Commercial	79%	76%	81%	82%	80%
Medicare	6	6	6	6	8
Medicaid	14	17	13	13	12
Urban status					
Rural	12%	13%	12%	16%	4.6%
Urban	88	88	88	84	95
Island					
Hawaii	1%	0%	1%	1%	0%
Maui	11	12	11	15	4
Oahu	88	88	89	84	95
Had a visit in 2015 with:					
PCP	81%	81%	80%	80%	81%
Mid-level provider	4	3	6	6	5
Mean PCP visits per member (no.)	3	3	3	3	3
Most common comorbidities for adult members					
Hypertension	41%	40%	41%	38%	44%
Diabetes	18	18	17	15	21
Obesity	17	14	18	17	19
Mean combined ERG score	1.8	1.8	1.8	1.7	2.0
Patients with diabetes					
Adults	9%	8%	10%	8%	12%
With HbA1c checked					
No	11%	13%	9.8%	11%	9.0%
Yes	89	87	90	90	91
Mean HbA1c level	7.5	7.4	7.5	7.5	7.5
Patient median household income by ZIP code level	$78,720	$77,445	$79,499	$79,514	$79,477
Patients in ZIP codes by median share with high school education[a]					
Below median	42%	45%	41%	39%	43%
Above median	58	56	60	61	57
Patients in ZIP codes by median share with college education[b]					
Below median	34%	36%	33%	33%	33%
Above median	66	64	67	68	67

SOURCE Authors' analysis of data from the randomized controlled trial. **NOTES** Providers in group 1 (the control group) received feedback on their performance as individuals only. Providers in group 2 also received feedback on their performance that compared them to their peers. Providers in group 3 received both types of feedback and shared with their patients with diabetes an incentive for those patients' glycemic control. ERG is Episode Risk Group. HbA1c is hemoglobin A1c. [a]ZIP codes where the share of residents with a high school diploma or higher is above the ZIP code median of 91.95 percent. [b]ZIP codes where the share of residents with a bachelor's degree or higher is above the ZIP code median of 27.95 percent.

EXHIBIT 2

Characteristics of primary care providers (PCPs) in the Hawaii Medical Services Association in the study sample, by PCP trial group, 2016

Variable	Overall (*N* = 88)	Group 1 (*n* = 29)	Groups 2 and 3 together (*n* = 59)	Group 2 only (*n* = 32)	Group 3 only (*n* = 27)
Sex					
Female	33%	31%	34%	41%	26%
Male	67	69	66	60	74
Specialty					
Family medicine	33%	21%	39%	25%	56%
General practice	8	7	9	16	0
Internal medicine	41	45	39	34	44
Pediatrics	16	24	12	22	0
APRN or physician assistant	2	3	2	3	0
Island					
Maui	15%	10%	17%	25%	7%
Oahu	85	90	83	75	93
Practice site location					
Rural	15%	10%	17%	25%	7%
Urban	85	90	83	75	93
Residency program location					
Hawaii	44%	45%	44%	56%	30%
Other US state	56	55	56	44	70
Medical school location					
Hawaii	52%	48%	54%	56%	52%
International	16	17	15	13	19
Other US state	32	35	31	31	30
Mean age, years	53	52	54	54	53
Mean panel size	850	974	789	877	684
Mean share of patients in 2015 by type of insurance					
Commercial	80%	77%	81%	5%	9%
Medicare	6	6	7	82	80
Medicaid	14	17	12	13	11
Physicians in organization (mean)	46	45	47	50	44
2015 ERG score (mean)					
Pediatric patients	0.62	0.66	0.60	0.66	0.50
Adult patients	1.9	1.7	1.8	1.6	2.1
Combined	1.8	1.7	1.8	1.7	2.0
Composite 2015 measure score (mean)	0.79	0.82	0.77	0.80	0.75

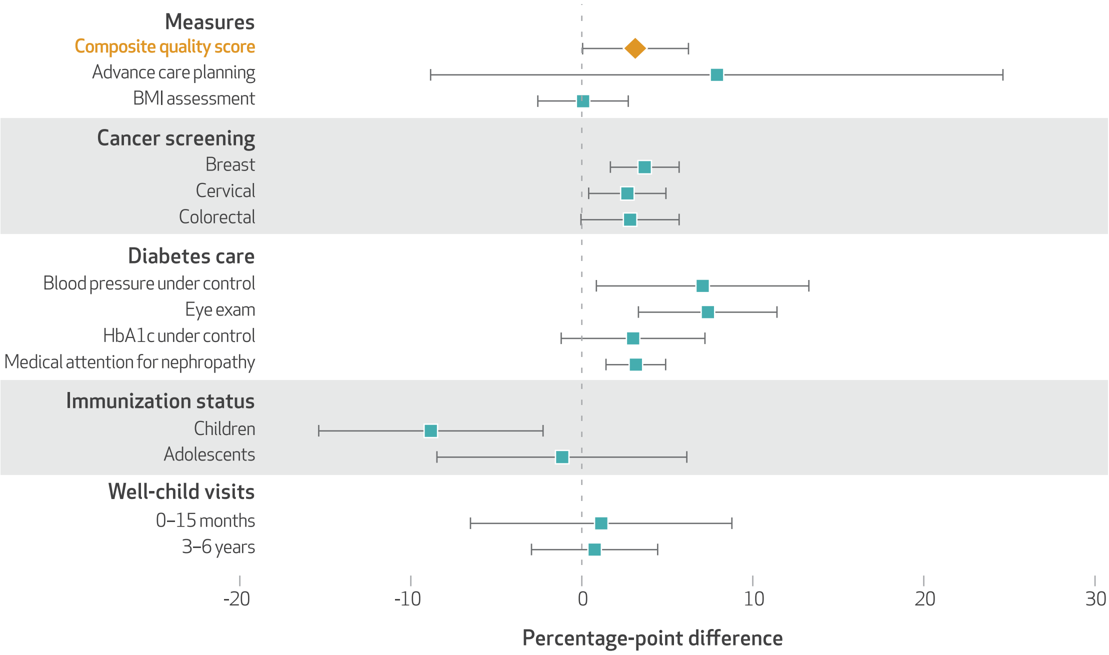
SOURCE Authors' analysis of data from the randomized controlled trial. **NOTES** Groups 1, 2, and 3 are explained in the notes to exhibit 1. ERG is Episode Risk Group. APRN is advanced practice registered nurse.

context of a shift from fee-for-service to the capitated Population-Based Payments for Primary Care system. It is possible that the intervention would not have generated similar results under fee-for-service. The fact that system payments were not visit based might have allowed primary care providers and practices more time to focus on achieving quality measures. Additionally, other features of the system—some of which were designed using insights from behavioral economics (for example, improvements on quality measures, not just the attainment of high thresholds, were rewarded)—might have enhanced the effect of the intervention. However, in view of past experiments with changing payments, financial incentives can be insufficient to drive changes in clinician practice. This could be because the financial incentives have not been large enough, but it could also be because financial incentives might displace intrinsic motivation. In contrast, providing peer comparisons might be more effective at activating clinicians' intrinsic motivation to provide high-quality care as well as to adhere to professional norms.

PEER VERSUS INDIVIDUAL COMPARISONS Third, this study highlights the fact that providing peer comparisons outperforms providing individual feedback alone and suggests that design features leveraging social comparisons could be important adjuncts that could make other inter-

EXHIBIT 3

Percentage-point differences in scores on quality measures of primary care providers who received individual feedback only (group 1), compared to those of providers who also received feedback comparing them to their peers (group 2), 2016



SOURCE Authors' analysis of data from the randomized controlled trial. **NOTES** The error bars show 95% confidence intervals. When we used the Holm-Bonferroni correction to adjust for multiple testing, the *p* values for the percentage-point differences were as follows: composite quality score (not applicable); advance care planning (>0.99); body mass index (BMI) assessment (>0.99); screening for breast cancer (<0.001), cervical cancer (0.20), and colorectal cancer (0.38); diabetes care—blood pressure under control (less than 140/90 mmHg) (0.21), eye exam (<0.001), hemoglobin A1c under control (no more than 9.0) (0.97), and medical attention for nephropathy (<0.001); immunization status for children (0.08) and adolescents (>0.99); and well-child visits at ages 0–15 months (>0.99) and at ages 3–6 years (>0.99).

ventions more effective. It is important to note the exact nature of the study interventions. For example, the peer comparisons intervention included graphic displays of quality measures and approximately twice-monthly email messages that were "pushed" to primary care providers to prompt them to pay attention to their performance. The individual feedback intervention did not include email pushes, and though it was designed together with the peer comparisons display, its display might not have been as easy for primary care providers to interpret.

We chose to display performance for all peer primary care providers in the same practice to increase the local nature of the comparison, thereby heightening its salience to the providers. We also delivered email pushes twice monthly to provide frequent prompts, but not so frequent as to generate a dismissive response. The multimodal aspects of the peer comparisons and their effectiveness underscore both the importance of testing behaviorally designed interventions with different features and their potential utility to policy makers and payers trying to stimulate practice improvements.

EFFECTIVENESS OF PEER COMPARISONS

Fourth, this study highlights evidence about the effectiveness of peer comparisons feedback and the relative social ranking hypothesis when trying to influence broader practice changes, compared with targeting specific choices about prescribing medications. Previous studies have shown the effectiveness of peer comparisons interventions in reducing antibiotic prescriptions by 16 percentage points among primary care practitioners treating adult patients[6] and 6.7 percentage points among pediatricians.[7] These studies used text or graphical displays to show comparisons on guideline-based prescribing for antibiotics in upper respiratory infections. In the context of opioid prescribing, peer comparisons among clinicians with self-perceptions of low opioid prescribing achieved reductions of two prescriptions per hundred patients.[16] None of these interventions attempted to address multiple changes in practice, nor were any of them accompanied by a large shift in payment. Thus, an unanswered question has been whether peer comparisons can influence broader practice patterns that may be relevant to more patients, clinicians, and policy makers. One other study used comparative feedback to improve diabetes care.[30] However, our study has provided evidence for using peer comparisons feedback to improve a composite of quality metrics that collectively are closer to more traditional assessment of the quality of clinician practice. Two pediatric measures demonstrated negative effects, though this could be explained by high baseline rates of achievement.

EFFECT OF PROVIDERS' BASELINE QUALITY Fifth, the fact that patients of primary care providers with lower baseline quality experienced significantly larger improvements in the quality of care is notable. This may be due in part to greater opportunity to improve, but in the context of other pay-for-performance studies that have failed to show such an effect, it may also reinforce the ability of peer comparisons to activate intrinsic motivation and desires to adhere to professional norms. Regardless of the mechanism, it is promising to see that the peer comparisons intervention had an outsize effect among patients with providers of low baseline quality. However, given the post hoc nature of these analyses, these results should be interpreted with caution.

QUALITY IMPROVEMENT Sixth, this study underscores the effectiveness of peer comparisons as a way to improve health care quality, particularly in the setting of payment changes. This provides important evidence to support Medicare's decisions to provide comparative feedback as part of recently implemented primary care and specialty payment reform programs.

Conclusion

A peer comparisons intervention that displayed quality information in a dashboard designed using insights from behavioral economics and implemented in the setting of a broad payment system change improved quality scores by 3.1 percentage points, relative to individual feedback alone. This highlights the ability of peer comparisons to improve health care quality and supports recent Medicare payment program designs that have made sharing comparative feedback a key component of their approach. ■

Previous versions of this article were presented at the Annual Meeting of the American Society of Health Economists, June 24, 2019; the Annual Meeting of the Society of General Internal Medicine, May 9, 2019; and the AcademyHealth Annual Research Meeting, June 24, 2019, all in Washington, D.C. Amol Navathe reports receiving grants from the Hawaii Medical Service Association (HMSA) for this work. He also reports receiving grants from the Anthem Public Policy Institute, Cigna, and Oscar Health; personal fees from Navvis and Company, Navigant Inc., Lynx Medical, Indegene Inc., Sutherland Global Services, Agathos Inc., and the National University Health System of Singapore; receiving personal fees and equity from NavaHealth; receiving speaking fees from the Cleveland Clinic; serving as a board member of Integrated Services Inc., a subsidiary of the HMSA, without compensation; and receiving an honorarium from Elsevier Press, none of which were related to this article. Kevin Volpp reports receiving grants and personal fees from CVS Health; personal fees and other fees from VAL Health; and grants from Humana, Merck, Weight Watchers, the HMSA, and Oscar Health Insurance, none of which were related to this article. Since 2016 Ezekiel Emanuel reports receiving speaker's fees from Tanner Healthcare System, Mid-Atlantic Permanente Group, American College of Radiology, Marcus Evans, Loyola University, Oncology Society of New Jersey, Good Shepherd Community Care, Remedy Partners, Medzel, Kaiser Permanente Virtual Medicine, Wallace H. Coulter Foundation, Lake Nona Institute, Allocation, Partners Chicago, Pepperdine University, Huron, American Case Management Association, Philadelphia Chamber of Commerce, Blue Cross and Blue Shield of Minnesota, UnitedHealth Group, Futures Without Violence, Children's Hospital of Philadelphia, Washington State Hospital Association, Association of Academic Health Centers, Blue Cross Blue Shield of Massachusetts, American Academy of Ophthalmology, Lumeris, Roivant Sciences Inc., Medical Specialties Distributors LLC, Vizient University Healthcare System, Center for Neuro-Degenerative Research, Colorado State University, Genentech Oncology Inc., Council of Insurance Agents and Brokers, Grifols Foundation, America's Health Insurance Plans, Montefiore Physician Leadership Academy, Greenwall Foundation, Medical Home Network, Healthcare Financial Management Association, Ecumenical Center of University of Texas Health San Antonio, American Association of Optometry, Associação Nacional de Hospitais Privados, National Alliance of Healthcare Purchaser Coalitions, Optum Labs, Massachusetts Association of Health Plans, District of Columbia Hospital Association, and Washington University. Since January 2017 he has been a venture partner with Oak HC/FT, a firm that invests in health services but not pharmaceuticals or devices. Emanuel is on the board of two start-ups, Village MD and Oncology Analytics. He has investments in Gilead, Allergan, Amgen, Baxter, and UnitedHealth Group. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the US government.

NOTES

1 Rosenthal MB. Beyond pay for performance—emerging models of provider-payment reform. N Engl J Med. 2008;359(12):1197–200.

2 Rosenthal MB, Frank RG, Li Z, Epstein AM. Early experience with pay-for-performance: from concept to practice. JAMA. 2005;294(14):1788–93.

3 Van Herck P, De Smedt D, Annemans L, Remmen R, Rosenthal MB, Sermeus W. Systematic review: effects, design choices, and context of pay-for-performance in health care. BMC Health Serv Res. 2010;10(1):247.

4 Eijkenaar F, Emmert M, Scheppach M, Schöffski O. Effects of pay for performance in health care: a systematic review of systematic reviews. Health Policy. 2013;110(2–3):115–30.

5 Jha AK, Joynt KE, Orav EJ, Epstein AM. The long-term effect of Premier pay for performance on patient outcomes. N Engl J Med. 2012;366(17):1606–15.

6 Meeker D, Linder JA, Fox CR, Friedberg MW, Persell SD, Goldstein NJ, et al. Effect of behavioral interventions on inappropriate antibiotic prescribing among primary care practices: a randomized clinical trial. JAMA. 2016;315(6):562–70.

7 Gerber JS, Prasad PA, Fiks AG, Localio AR, Grundmeier RW, Bell LM, et al. Effect of an outpatient antimicrobial stewardship intervention on broad-spectrum antibiotic prescribing by primary care pediatricians: a randomized trial. JAMA. 2013;309(22):2345–52.

8 Hallsworth M, Chadborn T, Sallis A, Sanders M, Berry D, Greaves F, et al. Provision of social norm feedback to high prescribers of antibiotics in general practice: a pragmatic national randomised controlled trial. Lancet. 2016;387(10029):1743–52.

9 CMS.gov. Comprehensive Primary Care Plus [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; [last updated 2019 Dec 11; cited 2019 Dec 30]. Available from: https://innovation.cms.gov/initiatives/Comprehensive-Primary-Care-Plus

10 CMS.gov. Oncology Care Model [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; [last updated 2019 Dec 4; cited 2019 Dec 30]. Available from: https://innovation.cms.gov/initiatives/Oncology-Care/

11 Navathe AS, Emanuel EJ. Physician peer comparisons as a nonfinancial strategy to improve the value of care. JAMA. 2016;316(17):1759–60.

12 Liao JM, Fleisher LA, Navathe AS. Increasing the value of social comparisons of physician performance using norms. JAMA. 2016;316(11):1151–2.

13 Emanuel EJ, Ubel PA, Kessler JB, Meyer G, Muller RW, Navathe AS, et al. Using behavioral economics to design physician incentives that deliver high-value care. Ann Intern Med. 2016;164(2):114–9.

14 Song H, Tucker AL, Murrell KL. The diseconomies of queue pooling: an empirical investigation of emergency department length of stay. Manage Sci. 2015;61(12):3032–53.

15 Sacarny A, Barnett ML, Le J, Tetkoski F, Yokum D, Agrawal S. Effect of peer comparison letters for high-volume primary care prescribers of quetiapine in older and disabled adults: a randomized clinical trial. JAMA Psychiatry. 2018;75(10):1003–11.

16 Michael SS, Babu KM, Androski C Jr, Reznek MA. Effect of a data-driven intervention on opioid prescribing intensity among emergency department providers: a randomized controlled trial. Acad Emerg Med. 2018;25(5):482–93.

17 Milani RV, Wilt JK, Entwisle J, Hand J, Cazabon P, Bohan JG. Reducing inappropriate outpatient antibiotic prescribing: normative comparison using unblinded provider reports. BMJ Open Qual. 2019;8(1):e000351.

18 Volpp KG, Navathe A, Lee EO, Mugishii M, Troxel AB, Caldarella K, et al. Redesigning provider payment: opportunities and challenges from the Hawaii experience. Healthc (Amst). 2018;6(3):168–74.

19 To access the appendix, click on the Details tab of the article online.

20 Navathe AS, Emanuel EJ, Bond A, Linn K, Caldarella K, Troxel A, et al. Association between the implementation of a population-based primary care payment system and achievement on quality measures in Hawaii. JAMA. 2019;322(1):57–68.

21 Optum. Symmetry® Episode Risk Groups® (ERG®): predict future health care utilization [Internet]. Eden Prairie (MN): Optum; c 2018 [cited 2019 Dec 30]. Available from: https://www.optum.com/content/dam/optum3/optum/en/resources/sell-sheet/symmetry-episode-risk-groups-erg-sell-sheet.pdf

22 Song Z, Safran DG, Landon BE, He Y, Ellis RP, Mechanic RE, et al. Health care spending and quality in year 1 of the Alternative Quality Contract. N Engl J Med. 2011;365(10):909–18.

23 Angrist JD, Pischke J-S. Mostly harmless econometrics: an empiricist's companion. Princeton (NJ): Princeton University Press; 2009.

24 McWilliams JM, Hatfield LA, Chernew ME, Landon BE, Schwartz AL. Early performance of accountable care organizations in Medicare. N Engl J Med. 2016;374(24):2357–66.

25 Joffe MM, Ten Have TR, Feldman HI, Kimmel SE. Model selection, confounder control, and marginal structural models: review and new applications. Am Stat. 2004;58(4):272–9.

26 Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65–70.

27 Rubin DB. Multiple imputation for nonresponse in surveys. New York (NY): John Wiley and Sons; 2004.

28 Dale SB, Ghosh A, Peikes DN, Day TJ, Yoon FB, Taylor EF, et al. Two-year costs and quality in the Comprehensive Primary Care initiative. N Engl J Med. 2016;374(24):2345–56.

29 Asch DA, Troxel AB, Stewart WF, Sequist TD, Jones JB, Hirsch AG, et al. Effect of financial incentives to physicians, patients, or both on lipid levels: a randomized clinical trial. JAMA. 2015;314(18):1926–35.

30 Kiefe CI, Allison JJ, Williams OD, Person SD, Weaver MT, Weissman NW. Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. JAMA. 2001;285(22):2871–9.