By Marina A. Milad, Roslyn C. Murray, Amol S. Navathe, and Andrew M. Ryan

DOI: 10.1377/hlthaff.2021.01020
HEALTH AFFAIRS 41,
NO. 4 (2022): 540–548

REVIEW ARTICLE

Value-Based Payment Models In The Commercial Insurance Sector: A Systematic Review

Marina A. Milad, University of Michigan, Ann Arbor, Michigan.

Roslyn C. Murray, University of Michigan.

Amol S. Navathe, Corporal Michael J. Cresencz Veterans Affairs Medical Center and University of Pennsylvania, Philadelphia, Pennsylvania.

Andrew M. Ryan (amryan@umich.edu), University of Michigan.

ABSTRACT Value-based payment models are a prominent strategy in health reform. Although Medicare payment models have been extensively evaluated, much less is known about value-based payment models in the commercial insurance sector. We performed the first systematic review of the quality, spending, and utilization effects of commercial models, extracting results from fifty-nine studies. Forty-one of these studies evaluated outcomes. More studies had positive results for quality outcomes (81 percent of studies) than for spending (56 percent) and utilization (58 percent). Less rigorous studies were more likely to find positive results. Given the mixed nature of the findings, commercial insurers should identify ways to strengthen value-based payment programs or leverage other strategies to improve health care value.

During the past two decades, value-based payment models, which link reimbursement to quality or spending targets, have diffused widely in the United States. Examples of these models include pay-for-performance, bundled or episode-based payment, shared savings or shared risk, and full or partial population-based payment.

Commercial insurers' deployment of value-based payment models has evolved over time. Population-based payment, also known as capitation, was credited with controlling commercial health insurance spending in the 1990s. However, its popularity waned as patients and physicians turned against its cost-cutting strategies (for example, gatekeeping, prior authorization, and narrow networks). In the early 2000s pay-for-performance programs gained prominence as payers sought to encourage higher quality in light of high-profile reports of quality shortcomings. Since 2010 value-based payment models have again shifted to require providers to take on more financial risk (exhibit 1).

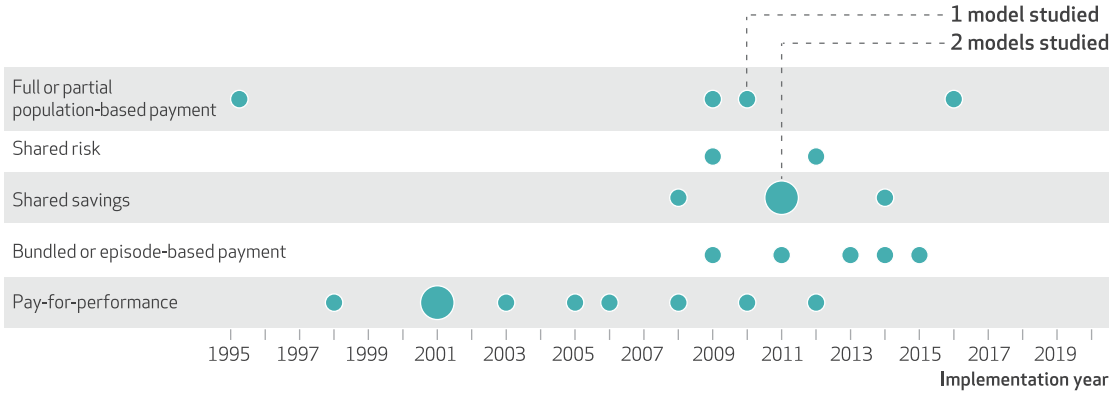
Provisions of the Affordable Care Act and the Medicare Access and CHIP Reauthorization Act of 2015 spurred and accelerated the implementation and evaluation of value-based payment models in Medicare.[1] Our study built on an extensive literature that has shown mixed and modest effects of value-based payment models on quality, spending, and utilization in public programs.[2,3]

Between 2010 and 2019 the average annual rate of spending growth per enrollee was much higher in commercial insurance (3.5 percent) than Medicare (2.0 percent).[4] Commercial prices are much higher and have grown faster than those in traditional Medicare, leading to rising premiums and out-of-pocket spending.[5] Substantial variation in the quality of care for privately insured populations has also been observed.[6] Therefore, understanding the potential for value-based payment models to improve quality and rein in commercial spending is critical.

We performed the first systematic review of commercial value-based payment models. The objectives were to summarize quality, spending, and utilization effects and explore associations between program characteristics and their outcomes.

EXHIBIT 1

Implementation of commercial value-based payment models evaluated in published studies, 2000–20



SOURCE Authors' systematic review of 59 studies published in the peer-reviewed literature between 2000 and 2020. **NOTES** The years on the *x* axis are those in which the models were implemented, which vary from the years that studies were published. The size of each bubble represents the number of value-based payment models evaluated in the studies included in the review. Definitions of payment models are in online appendix exhibit A11 (see note 7 in text).

Study Data And Methods

LITERATURE REVIEW Our search strategy sought to identify studies discussing commercial value-based payment models in the US. We used the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) Protocols and the Population, Intervention, Comparator, and Outcome framework and searched the Ovid MEDLINE and Scopus databases (appendix exhibit A1) for peer-reviewed, English-language articles published between January 1, 2000, and July 28, 2020.[7] We incorporated sixteen additional papers that were identified through prior systematic reviews of value-based payment models across all payers.[8] We did not include studies that evaluated consumer-focused benefit design (for example, narrow networks or high-deductible health plans), which encourage consumers to seek fewer or lower-cost services. We excluded two studies that did not allow us to extract program details specific to commercial payers.

DATA EXTRACTION We organized model design (appendix exhibits A4 and A5), individual study descriptions (appendix exhibit A6), outcomes (appendix exhibit A7), and study strength (appendix exhibit A8) data elements into evidence tables.[7] One author (Marina Milad) extracted data elements from each study, and another (Roslyn Murray) independently verified extracted elements for accuracy. The team discussed and resolved inconsistencies. Appendix exhibits A4 and A5 include articles that only discussed program design and implementation or articles from which results were not reported for commercial enrollees alone.[7]

Program effects were categorized as positive (all outcomes improved), mixed positive (more than half of the outcomes improved), no effects or mixed effects (on average, the outcomes did not improve), mixed negative (more than half of the outcomes did not improve), and negative (no outcomes improved). We used a p value less than or equal to 0.05 for statistical significance. We defined program success as improving quality, reducing spending, and improving the appropriateness of utilization as defined by individual studies.

We evaluated the methodological rigor of the study and assigned a study strength rating based on the empirical approach and whether the study demonstrated that the assumptions required for causal interpretation were supported. We assigned medium or high strength ratings to studies that enabled credible causal inferences (for example, difference-in-differences analyses). Studies received a high rating only if authors provided clear evidence to support the assumptions required for causal inference (for example, parallel trends for difference-in-differences). We assigned low strength ratings to studies that used a pre-post or post-only comparison, as these approaches are more subject to bias. Our focus on rigor and potential bias is similar to that of other evidence appraisal frameworks.[9] We provide detailed descriptions of results only for studies characterized as medium or high strength.

STATISTICAL ANALYSIS We performed a regression analysis to evaluate the association between model effectiveness (positive, not positive) and study strength (low, medium, high), as well as outcome evaluated (quality, spending, utilization). We grouped positive and mixed positive effectiveness results together as positive (equal

to 1) and grouped no effect or mixed effects and mixed negative results as not positive (equal to 0). Because of the small cell sizes for the negative category, we included these results in the category of mixed negative effects. The model was estimated at the study outcome level ($n = 74$) and included dummy variables representing study strength and outcome evaluated. Standard errors were clustered at the study level.

LIMITATIONS Our study had several limitations. First, our systematic review might not represent the full experience of commercial value-based payment models. We identified fifty-nine studies published over the course of a twenty-year period, and only forty-one of them evaluated outcomes. This limited number of peer-reviewed publications may have resulted from several factors, including hesitation to publish unsuccessful findings and a preference to communicate results in industry and media outlets instead of peer-reviewed literature. Second, only a few of the published studies accounted for savings payouts in their spending results; therefore, the literature overstates the impact of value-based models on net savings for commercial insurers.[10–13] Third, the variability in program design and lack of consistent terminology made it challenging to categorize programs, thus limiting the ability to draw conclusions by model. Relatedly, we were not able to quantitatively evaluate associations among payment model, governance factors, and outcomes because of insufficient power from limited observations. Similarly, our regression analysis may have been underpowered, and findings from this analysis should be considered suggestive.

Study Results

We identified 1,255 unique studies in the systematic review, completed full-text evaluations of 277 studies, and included 59 studies, 41 of which evaluated outcomes. The outcomes included in these evaluations were quality, spending, and utilization (exhibit 2). The fifty-nine studies that we included in our analysis are listed as citations 2–60 in appendix exhibit A12.[7] These studies are further characterized in appendix exhibit A3, where the forty-one studies that present quality, spending, or utilization outcomes are identified as well.[7] Exhibit 3 summarizes results of the thirty studies with high and medium study strength by outcome and value-based payment model. Many studies evaluated more than one outcome.

QUALITY OF CARE

▶ **PAY-FOR-PERFORMANCE:** Overall, quality of care improved or remained stable in pay-for-performance programs. Of the thirteen studies evaluating outcomes of pay-for-performance models, two demonstrated no or mixed effects on quality, six demonstrated mixed positive effects, three demonstrated positive effects, and two did not evaluate quality effects (exhibit 2, appendix exhibit A9).[7] The seven studies with a

EXHIBIT 2

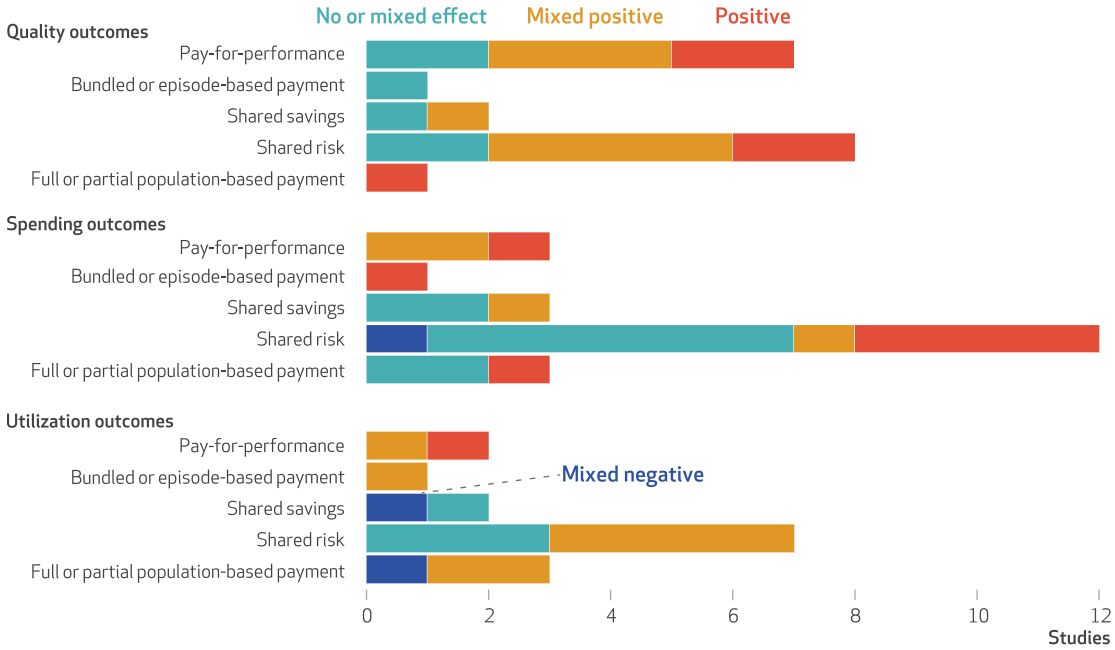
Summary of sample for literature review of commercial value-based payment models, 2000–20

		Primary payment method[a]					
	Total	Pay-for-performance	Bundled or episode-based payment	Shared savings	Shared risk	Partial population-based payment	Full population-based payment
Total no. of studies	59	22	9	5	16	1	6
No. of studies that evaluated outcomes	41	13	5	4	14	1	4
No. of models with evaluated outcomes[b]	24	9	5	4	2	1	3
Outcomes evaluated[c]							
Quality	25	11	3	2	8	0	1
Spending	28	4	4	3	12	1	4
Utilization	20	5	2	3	7	0	3
Study strength rating[d]							
Low	11	6	3	1	0	0	1
Medium	12	6	1	2	1	1	1
High	18	1	1	1	13	0	2

SOURCE Authors' systematic review of 59 studies published in the peer-reviewed literature between 2000 and 2020. [a]Definitions of payment models are in appendix exhibit A11 (see note 7 in text). [b]The number of models that evaluated outcomes refers to the total number of unique commercial value-based payment models included in our literature review. Several studies evaluated the same model. [c]Many studies evaluated more than one outcome; therefore, the total number of outcomes is greater than the number of studies included. [d]Study strength was determined by assessing the empirical approach and whether the study demonstrated that the assumptions required for causal interpretation were supported. A low rating was assigned to studies that used a pre-post or post-only comparison. A medium or high rating was assigned to studies that enabled credible causal inferences (for example, difference-in-differences analyses). Studies received a high rating only if authors provided clear evidence to support the assumptions required for causal inference (for example, parallel trends for difference-in-differences).

EXHIBIT 3

Effects of commercial value-based payment models on quality, spending, and utilization, as reported in studies with medium and high strength, 2000–20



SOURCE Authors' systematic review of studies published in the peer-reviewed literature between 2000 and 2020. **NOTES** $n = 30$ studies with medium or high strength. *Quality* is defined as the result of a process (such as breast cancer screening rates), patient experience (such as wait time), or patient health measure (HbA1c level for patients with diabetes). *Spending* is defined as the amount of money the payer spends on services per case or per procedure, or the payer's total annual or monthly spending. *Utilization* is defined as the quantity of patient services used, such as the quantity of hospital admissions or emergency department visits. Program effects are categorized as positive (all outcomes improved), mixed positive (more than half of the outcomes improved), no effects or mixed effects (among the measures evaluated, measures did not improve, on average), and mixed negative (more than half of the outcomes did not improve). Effects categorized as positive demonstrated improved outcomes (higher quality, lower spending, and more appropriate utilization), and those categorized as negative demonstrated worsened outcomes (lower quality, higher spending, and less appropriate utilization). Because of the limited number of observations, our regression analysis may be underpowered, and findings from this analysis should be considered suggestive. Definitions of payment models are in appendix exhibit A11 (see note 7 in text).

high or medium study strength found small improvements in quality measures (appendix exhibit A7).[7] One study highlighted mixed findings among five pay-for-performance programs, with stronger improvement in HbA1c screening (4.0 percentage points; $p = 0.02$), diabetes eye exams (ranging from 4.0 [$p < 0.01$] to 7.0 [$p = 0.04$] percentage points), and well-child visits (5.0 percentage points; $p = 0.02$) but lower improvement in low-density lipoprotein testing (−3.0 percentage points; $p < 0.01$), chlamydia screening (−11.0 percentage points; $p < 0.01$), diabetes urine testing (−7.0 percentage points; $p < 0.01$), and well-child visits for adolescents (−5.0 percentage points; $p < 0.01$) for the treatment group compared with the control group.[14] The Quality Incentive Program in California was associated with higher cancer screening rates (ranging from 3.5 percent [$p < 0.01$] to 3.6 percent [$p = 0.02$])[15,16] but lower chlamydia screening rates (−5.3 percent; $p < 0.01$),[16] as well as no significant changes for several other measures compared with the control.[15,16] A Massachusetts-based model improved smoking status documentation among enrollees compared with the control group (adjusted odds ratio: 1.3; $p < 0.01$).[17] The Quality Blue Primary Care program in Louisiana improved HbA1c screening (3.92 percent; $p < 0.01$) more for the treatment group compared with the control group, whereas other screening test rates were not statistically different.[18] Finally, the Physician Group Incentive Program in Michigan demonstrated mixed positive effects, including improvements in overall quality (1.6 percent; $p < 0.01$)[19] and for several individual quality measures[20] (appendix exhibit A7).[7]

▶ **BUNDLED OR EPISODE-BASED PAYMENT:** Quality of care in bundled or episode-based payment programs also either improved or remained stable. Of the five studies evaluating outcomes of these models, one demonstrated no effects on quality, one demonstrated mixed positive effects, one demonstrated positive effects,

and two did not evaluate quality effects (exhibit 2, appendix exhibit A9).[7] However, the studies that found positive or mixed positive outcomes had low study strength. The only medium-strength study, which evaluated UnitedHealthcare's bundle for breast, lung, and colon cancer, did not find differences between the treatment and control groups on quality[21] (appendix exhibit A7).[7]

▶ **SHARED SAVINGS OR SHARED RISK:** Evaluations of shared savings or shared risk models found that these programs either improved or had no effect on quality. Of the eighteen studies evaluating outcomes of shared savings or shared risk models, three demonstrated no or mixed effects on quality, five demonstrated mixed positive effects, two demonstrated positive effects, and eight did not evaluate quality effects (exhibit 2, appendix exhibit A9).[7] Seventeen studies had medium or high study strength. Maryland's Multi-payer Patient-Centered Medical Home pilot demonstrated mixed positive quality effects, with improved cervical cancer screening (relative risk: 1.08 in year 2; $p < 0.05$), an increase in adolescent well-care visits (3 percentage points in year 1 and 5 percentage points in year 3; both $p < 0.05$), and reduced use of postpartum care (RR: 0.37 in year 3; $p < 0.01$) compared with the baseline between treatment and control groups. However, there were no statistically significant differences in other measures.[22] Between 2009 and 2016 the Massachusetts Alternative Quality Contract achieved positive and mixed positive quality effects across three process measure domains of adult preventive care, pediatric care, and chronic care management compared with national and New England performance.[10,11,13,23,24] However, there were no quality improvements related to substance use disorders[25,26] (appendix exhibit A7).[7]

▶ **FULL OR PARTIAL POPULATION-BASED PAYMENT:** Of the five studies evaluating outcomes of population-based payment models, only Hawaii's Population-based Payments for Primary Care program, classified as a high-quality study, evaluated quality. The program increased a composite quality score composed of thirteen measures[27] (2.3 percentage points; $p < 0.01$; appendix exhibit A7).[7]

SPENDING

▶ **PAY-FOR PERFORMANCE:** Of the thirteen studies evaluating outcomes of pay-for-performance models, three demonstrated mixed positive effects on spending, one demonstrated positive effects, and nine did not evaluate spending effects (exhibit 2, appendix exhibit A9).[7] Among the three studies with medium study strength, one found positive effects and two found mixed positive effects on spending. No studies evaluating spending were classified as high study strength. The Physician Group Incentive Program reduced total adult medical (−1.1 percent; $p < 0.01$) and pediatric (−5.1 percent; $p < 0.01$)[20] spending between 2009 and 2011. Another study found that the program led to a lower spending trajectory compared with the control but that the program did not achieve differences in medical-surgical spending.[19] Further, between 2010 and 2013 the program was associated with a reduction in drug spending for the intervention group (odds ratio: 0.82; $p < 0.01$), but these effects were reversed, conditional on any pharmaceutical use (3.9 percent; $p < 0.01$).[19] Effects of the Quality Blue Primary Care program varied by spending category: Total (RR: 0.92; $p < 0.001$), medical (RR: 0.87; $p < 0.01$), and specialty (RR: 0.95; $p < 0.01$) spending increased at a lower rate in the treatment group, whereas spending increased at a higher rate for ambulatory emergency department (RR: 1.081; $p = 0.02$) and emergency department (RR: 1.10; $p < 0.01$) visits. Effects for additional categories of spending were not statistically significant[18] (appendix exhibit A7).[7]

▶ **BUNDLED OR EPISODE-BASED PAYMENT:** Of the five studies evaluating outcomes of bundled or episode-based payment models, one demonstrated no effect on spending, one demonstrated mixed positive effects, two demonstrated positive effects, and one did not evaluate spending effects (exhibit 2, appendix exhibit A9).[7] However, only one study was characterized as medium study strength. UnitedHealthcare's cancer care model was estimated to have saved more than $33 million for treating 810 patients with breast, colon, or lung cancer[21] (appendix exhibit A7).[7]

▶ **SHARED SAVINGS OR SHARED RISK:** Of the eighteen studies evaluating outcomes of shared savings or shared risk models, one demonstrated mixed negative effects on spending, eight demonstrated no or mixed effects, two demonstrated mixed positive effects, four demonstrated positive effects, and three did not evaluate spending effects (exhibit 2, appendix exhibit A9).[7] Seventeen of the studies had medium or high study strength. One study found mixed negative spending trends. Compared with traditional Medicare, Aetna's accountable care organization resulted in greater total (ranging from $205 to $538; $p < 0.05$), inpatient ($291; $p < 0.05$), non–evaluation and management outpatient ($195; $p < 0.05$), evaluation and management (ranging from $20 to $48; $p < 0.05$), and emergency department ($32; $p < 0.05$) spending.[28] The Maryland multipayer pilot led to lower increases in outpatient payments relative to baseline in the first program year (−$146; $p = 0.03$) but was not

Value-based payment models may be enhanced by benefit design structures and delivery system reforms.

associated with significant changes for other years or inpatient spending categories.[22] The Alternative Quality Contract decreased average spending (ranging from 2.3 percent to 11.9 percent), with greater relative savings in more mature cohorts.[10,11,13,29] This spending trend was also achieved for enrollees with behavioral health risk (–$238).[30] The Alternative Quality Contract was one of a few models that evaluated net savings relative to incentive payments and development costs. Provider incentive payments exceeded savings from reduced utilization between 2009 and 2011. However, after 2012 the Alternative Quality Contract generated net savings.[13,31] An evaluation of the CareFirst Total Care and Cost Improvement program also accounted for incentive payments in spending results but did not find net reductions[12] (appendix exhibit A7).[7]

▶ **FULL OR PARTIAL POPULATION-BASED PAYMENT:** Of the five studies evaluating outcomes of population-based payment models, three demonstrated no or mixed effects on spending, and two demonstrated positive effects (exhibit 2, appendix exhibit A9).[7] Of the four studies with medium or high strength, one found spending reductions, and three found no or mixed effects. A New York–based patient-centered medical home pilot did not reduce spending in the first two years.[32] The Population-based Payments for Primary Care model showed no total spending reductions and mixed results across spending categories and population subgroups.[27,33] The California Public Employees' Retirement System accountable care organization demonstrated 10 percent lower per member spending in the first intervention year compared with the control, but no statistical analysis was reported[34] (appendix exhibit A7).[7]

UTILIZATION

▶ **PAY-FOR-PERFORMANCE:** Only five of the thirteen pay-for-performance studies evaluating outcomes examined utilization (exhibit 2, appendix exhibit A9).[7] Two of these five studies were medium- or high-strength studies, and they found mixed positive and positive effects. The Quality Blue Primary Care program increased primary care and decreased specialty visits while decreasing inpatient admissions.[18] Participants in the Physician Group Incentive Program had lower odds of thirty- and ninety-day readmissions and emergency department visits than nonparticipants[19] (appendix exhibit A7).[7]

▶ **BUNDLED OR EPISODE-BASED PAYMENT:** Two of the five studies evaluating outcomes of bundled or episode-based payment models examined utilization, but only one of the two was a medium- or high-strength study (exhibit 2, appendix exhibit A9).[7] The authors found mixed positive effects on utilization. The Arkansas Health Care Payment Improvement Initiative increased the probability of undergoing colonoscopies, a clinically underused service (17.2 percent; $p < 0.01$), but had no statistically significant effect on the other outcomes evaluated[35] (appendix exhibit A7).[7]

▶ **SHARED SAVINGS OR SHARED RISK:** Ten of the eighteen studies evaluating outcomes of shared savings or shared risk models evaluated utilization, nine of which had medium or high strength ratings (exhibit 2, appendix exhibit A9).[7] These nine studies found varied effects on utilization. Maryland's Multi-payer Patient Centered Medical Home pilot demonstrated mixed negative effects.[22] Four studies evaluating the Alternative Quality Contract achieved mixed positive effects.[13,30,36,37] Notably, the eight-year program evaluation found that 71 percent of the reduction in spending was attributed to lower utilization.[13] Four studies found no effects[12,24,26,38] (appendix exhibit A7).[7]

▶ **FULL OR PARTIAL POPULATION-BASED PAYMENT:** Three of the five studies that evaluated outcomes of full or partial population-based payment studies assessed utilization, and all were medium- or high-strength studies (exhibit 2, appendix exhibit A9).[7] These studies found mixed positive and mixed negative effects on utilization. The Population-based Payments for Primary Care model had mixed negative results, with a decrease in primary care (–3.9 percentage points; $p < 0.01$)[27] and nuclear medicine utilization (–18.1 percent; $p < 0.01$)[33] but an increase in drug utilization (15.6 percentage points; $p < 0.01$).[27] The California Public Employees' Retirement System accountable care organization reduced thirty-day readmission rates (–1.1 percentage points in year 1 and –0.2 percentage points in year 2) and increased average length-of-stay (5.9 percent in year 2)[34] (appendix exhibit A7).[7]

IMPLEMENTATION CHARACTERISTICS Twenty-four models studied had evaluated outcomes

(exhibit 2 and appendix exhibit A3).[7] Of these models, evaluations of eighteen models discussed the importance of program implementation factors along with technical assistance. Across all payment models, factors reported as important contributors to success included involving interdisciplinary stakeholders in program design and governance; providing technical assistance for establishing a data infrastructure; disseminating web-based, real-time performance reports on outcomes; and providing opportunities for shared learning.

ACCOUNTING FOR STUDY STRENGTH When controlling for study strength, we found a greater number of studies with positive results for quality outcomes (81 percent of studies) compared with spending (56 percent) and utilization (58 percent). In addition, studies with low strength were more likely to have positive results (78 percent of studies) compared to those with medium (67 percent) and high (51 percent) study strength (appendix exhibit A10).[7] We were not able to analyze the association between individual models and outcomes or study strength because of the small number of observations. Nevertheless, we observed that pay-for-performance models achieved fewer positive effects when we accounted for study strength.

Discussion

This first systematic review of commercial value-based payment models produced three main findings. First, between 2000 and 2020 only forty-one peer-reviewed studies evaluated outcomes of commercial value-based payment programs. Most published studies were evaluations of pay-for-performance models; few were of models with downside risk. Second, our review found that value-based payment models tended to improve quality outcomes, but there was less evidence of spending reductions and improvements in the appropriateness of utilization. Finally, studies with methodologically strong designs were less likely than those using less rigorous methods to find positive results.

Our study built on research demonstrating the mixed and modest effects of value-based payment models on outcomes in the public sector.[3,4] Compared with the literature on public programs, we found that commercial value-based payment models may have more positive effects on quality but similarly mixed results on spending and utilization. Thus, we found substantial gaps in the evidence supporting the effectiveness of value-based payment models in the commercial health insurance sector.

Commercial payers need to identify ways to strengthen value-based payment programs or turn to other strategies to improve health care value.

Policy Implications

Our findings suggest five key implications for the future of commercial value-based payment. First, evidence from this review suggests that success in improving quality, reducing spending, and improving appropriate utilization after a shift to greater risk-sharing is far from guaranteed. More empirical work is needed to determine whether new payment models can reduce spending and improve utilization without negatively affecting patients.

Second, value-based payment models may be enhanced by benefit design structures and delivery system reforms. For example, accountable care organizations, which use risk-based arrangements, can adopt narrow or tiered networks to drive patients to seek services from high-value providers as a lever for controlling spending. Yet the degree to which payers are aligning demand-side benefit designs with supply-side payments, and the effectiveness of this alignment, is unclear from the literature between 2000 and 2020.

Third, financial incentives should be deployed alongside support for providers. Numerous studies emphasized that successful implementation of value-based payment models requires technical assistance for the development of data infrastructure, collaborative involvement from providers and diverse stakeholder groups in program design and governance, and regular performance reporting and opportunities for shared learning. Relatedly, it may be more challenging for commercial payers to implement and providers to participate in value-based payment models in markets with competing models and overlapping program designs and reporting needs.

Fourth, with the rise of value-based payment models among commercial payers and the versatility of models, alignment of commercial and public-sector value-based payment models can facilitate the adoption and sustainability of these models, particularly among providers in markets with numerous payers. Using consistent performance measures across value-based payment models, for example, can support providers serving a broad payer mix in focusing on population health.

Finally, the role of hospitals in commercial value-based payment arrangements is uncertain. Although hospitals have participated in a variety of upside-only arrangements, they have been less involved in models that require them to take on risk. Hospitals' dependence on high service volume is at odds with risk-based models, which incentivize improved population health to reduce downstream service use and control spending. As hospitals are a key driver of overall health care spending, additional research is needed to increase understanding of how these entities respond to value-based models that adopt greater risk-sharing.

Conclusion

Our review of the literature on value-based payment models in the commercial insurance sector suggests that these programs have been less successful than anticipated. They have been somewhat successful at improving health care quality; however, the most methodologically rigorous studies were less likely to find evidence of quality improvement. Evidence of value-based payment models' impact on spending and utilization is less conclusive. Commercial payers need to identify ways to strengthen value-based payment programs or turn to other strategies to improve health care value. ■

Andrew Ryan acknowledges receiving funding from the Agency for Healthcare Research and Quality (Grant No. R01 HS026244). Amol Navathe reports grants from Hawaii Medical Service Association, Commonwealth Fund, Robert Wood Johnson Foundation, Donaghue Foundation, Pennsylvania Department of Health, Ochsner Health System, United Healthcare, Blue Cross Blue Shield of North Carolina, Blue Shield of California, and Humana; personal fees from Navvis Healthcare; equity from Agathos, Inc.; personal fees and equity from Navahealth; personal fees from YNHHSC/CORE, Maine Health Accountable Care Organization, Singapore Ministry of Health, Elsevier Press, Medicare Payment Advisory Commission, Cleveland Clinic, Analysis Group, VBID Health, Advocate Physician Partners, Federal Trade Commission, and Catholic Health Services Long Island; equity from Embedded Healthcare; and noncompensated board membership for Integrated Services, Inc., all outside the submitted work in the past three years. In addition to her affiliation with the University of Michigan, Marina Milad is employed by Accenture of Chicago, Illinois. The views and opinions expressed in this article are solely those of the authors and do not necessarily reflect the position of Accenture.

NOTES

1 Abrams MK, Nuzum R, Zezza MA, Ryan J, Kiszla J, Guterman S. The Affordable Care Act's payment and delivery system reforms: a progress report at five years [Internet]. New York (NY): Commonwealth Fund; 2015 May 7 [cited 2022 Feb 8]. Available from: https://www.commonwealthfund.org/publications/issue-briefs/2015/may/affordable-care-acts-payment-and-delivery-system-reforms

2 Chee TT, Ryan AM, Wasfy JH, Borden WB. Current state of value-based purchasing programs. Circulation. 2016;133(22):2197–205.

3 Agarwal R, Liao JM, Gupta A, Navathe AS. The impact of bundled payment on health care spending, utilization, and quality: a systematic review. Health Aff (Millwood). 2020;39(1):50–7.

4 Kamal R, McDermott D, Ramirez G, Cox C. How has U.S. spending on healthcare changed over time? [Internet]. Washington (DC): Peterson-KFF Health System Tracker; 2020 Dec 23 [Feb 8] Available from: https://www.healthsystemtracker.org/chart-collection/u-s-spending-healthcare-changed-time/

5 Johnson B, Kennedy K, Kurowski D, Bloschichak A, Clayton E, Biniek JF, et al. Comparing commercial and Medicare professional service prices [Internet]. Washington (DC): Health Care Cost Institute; 2020 Aug 13 Feb 8]. Available from: https://healthcostinstitute.org/hcci-research/comparing-commercial-and-medicare-professional-service-prices

6 McKellar MR, Landrum MB, Gibson TB, Landon BE, Fendrick AM, Chernew ME. Geographic variation in quality of care for commercially insured patients. Health Serv Res. 2017;52(2):849–62.

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

8 Kaufman BG, Spivack BS, Stearns SC, Song PH, O'Brien EC. Impact of Accountable Care Organizations on utilization, care, and outcomes: a systematic review. Med Care Res Rev. 2019;76(3):255–90.

9 Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

10 Song Z, Safran DG, Landon BE, Landrum MB, He Y, Mechanic RE, et al. The "Alternative Quality Contract," based on a global budget, lowered medical spending and improved quality. Health Aff (Millwood). 2012;31(8):1885–94.

11 Song Z, Rose S, Safran DG, Landon BE, Day MP, Chernew ME. Changes in health care spending and quality 4 years into global payment. N Engl J Med. 2014;371(18):1704–14.

12 Afendulis CC, Hatfield LA, Landon BE, Gruber J, Landrum MB, Mechanic RE, et al. Early impact of CareFirst's patient-centered medical home with strong financial incentives. Health Aff (Millwood). 2017;36(3):468–75.

13 Song Z, Ji Y, Safran DG, Chernew ME. Health care spending, utilization, and quality 8 years into global payment. N Engl J Med. 2019;381(3):252–63.

14 Pearson SD, Schneider EC, Kleinman KP, Coltin KL, Singer JA. The impact of pay-for-performance on health care quality in Massachusetts, 2001–2003. Health Aff

(Millwood). 2008;27(4):1167–76.
15 Rosenthal MB, Frank RG, Li Z, Epstein AM. Early experience with pay-for-performance: from concept to practice. JAMA. 2005;294(14): 1788–93.
16 Mullen KJ, Frank RG, Rosenthal MB. Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. Rand J Econ. 2010;41(1):64–91.
17 Kruse GR, Chang Y, Kelley JHK, Linder JA, Einbinder JS, Rigotti NA. Healthcare system effects of pay-for-performance for smoking status documentation. Am J Manag Care. 2013;19(7):554–61.
18 Shi Q, Yan TJ, Lee P, Murphree P, Yuan X, Shao H, et al. Evaluation of the Quality Blue Primary Care program on health outcomes. Am J Manag Care. 2017;23(12):e402–8.
19 Cross DA, Cohen GR, Harris Lemak C, Adler-Milstein J. Sustained participation in a pay-for-value program: impact on high-need patients. Am J Manag Care. 2017;23(2): e33–40.
20 Lemak CH, Nahra TA, Cohen GR, Erb ND, Paustian ML, Share D, et al. Michigan's fee-for-value physician incentive program reduces spending and improves quality in primary care. Health Aff (Millwood). 2015; 34(4):645–52.
21 Newcomer LN, Gould B, Page RD, Donelan SA, Perkins M. Changing physician incentives for affordable, quality cancer care: results of an episode payment model. J Oncol Pract. 2014;10(5):322–6.
22 Marsteller JA, Hsu Y-J, Gill C, Kiptanui Z, Fakeye OA, Engineer LD, et al. Maryland Multipayor Patient-centered Medical Home Program: a 4-year quasiexperimental evaluation of quality, utilization, patient satisfaction, and provider perceptions. Med Care. 2018;56(4):308–20.
23 Chien AT, Song Z, Chernew ME, Landon BE, McNeil BJ, Safran DG, et al. Two-year impact of the Alternative Quality Contract on pediatric health care quality and spending. Pediatrics. 2014;133(1):96–104.
24 Sharp AL, Song Z, Safran DG, Chernew ME, Fendrick AM. The effect of bundled payment on emergency department use: Alternative Quality Contract effects after year one. Acad Emerg Med. 2013;20(9): 961–4.
25 Stuart EA, Barry CL, Donohue JM, Greenfield SF, Duckworth K, Song Z, et al. Effects of accountable care and payment reform on substance use disorder treatment: evidence from the initial 3 years of the Alternative Quality Contract. Addiction. 2017; 112(1):124–33.
26 Donohue JM, Barry CL, Stuart EA, Greenfield SF, Song Z, Chernew ME, et al. Effects of global payment and accountable care on medication treatment for alcohol and opioid use disorders. J Addict Med. 2018;12(1): 11–8.
27 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.
28 Newhouse JP, Price M, Hsu J, Landon B, McWilliams JM. Delivery system performance as financial risk varies. Am J Manag Care. 2019; 25(12):e388–94.
29 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.
30 Barry CL, Stuart EA, Donohue JM, Greenfield SF, Kouri E, Duckworth K, et al. The early impact of the "Alternative Quality Contract" on mental health service use and spending in Massachusetts. Health Aff (Millwood). 2015;34(12): 2077–85.
31 Song Z. Accountable care organizations: early results and future challenges. J Clin Outcomes Manag. 2014;21(8):364–71.
32 Vats S, Ash AS, Ellis RP. Bending the cost curve? Results from a comprehensive primary care payment pilot. Med Care. 2013;51(11):964–9.
33 Dinh CT, Linn KA, Isidro U, Emanuel EJ, Volpp KG, Bond AM, et al. Changes in outpatient imaging utilization and spending under a new population-based primary care payment model. J Am Coll Radiol. 2020;17(1 Pt B):101–9.
34 Markovich P. A global budget pilot project among provider partners and Blue Shield of California led to savings in first two years. Health Aff (Millwood). 2012;31(9):1969–76.
35 Chen JL, Chernew ME, Fendrick AM, Thompson JW, Rose S. Impact of an episode-based payment initiative by commercial payers in Arkansas on procedure volume: an observational study. J Gen Intern Med. 2020; 35(2):578–85.
36 Huskamp HA, Greenfield SF, Stuart EA, Donohue JM, Duckworth K, Kouri EM, et al. Effects of global payment and accountable care on tobacco cessation service use: an observational study. J Gen Intern Med. 2016;31(10):1134–40.
37 Joyce NR, Huskamp HA, Hadland SE, Donohue JM, Greenfield SF, Stuart EA, et al. The Alternative Quality Contract: impact on service use and spending for children with ADHD. Psychiatr Serv. 2017;68(12): 1210–2.
38 Afendulis CC, Fendrick AM, Song Z, Landon BE, Safran DG, Mechanic RE, et al. The impact of global budgets on pharmaceutical spending and utilization: early experience from the Alternative Quality Contract. Inquiry. 2014;51: 0046958014558716.