Adapting Evaluations of Alternative Payment Models to a Changing Environment

Thomas W. Grannemann and Randall S. Brown

Objective. To identify the most robust methods for evaluating alternative payment models (APMs) in the emerging health care delivery system environment.

Study Design (Approach). We assess the impact of widespread testing of alternative payment models on the ability to find credible comparison groups. We consider the applicability of factorial research designs for assessing the effects of these models.

Principal Findings. The widespread adoption of alternative payment models could effectively eliminate the possibility of comparing APM results with a “pure” control or comparison group unaffected by other interventions. In this new environment, factorial experiments have distinct advantages over the single-model experimental or quasi-experimental designs that have been the mainstay of recent tests of Medicare payment and delivery models.

Conclusions. The best prospects for producing definitive evidence of the effects of payment incentives for APMs include fractional factorial experiments that systematically vary requirements and payment provisions within a payment model.

Key Words. Medicare, alternative payment models, factorial design

In 2019, under the provisions of the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA), Medicare will add 5 percent to the payments of providers who participate in a diverse array of alternative payment models (APMs) such as medical homes, accountable care organizations (ACOs), and bundled payment programs. This seemingly small change may prove to be a tipping point not only for the adoption of APMs but also for the methods we use to evaluate them. With much resting on the success of these models, there will be a more pressing need for robust information on provider responses to the new payment methods and to the incentives that determine cost and quality. However, the widespread and expanding use of APMs among public and private payers will make it more difficult to measure the causal effects of new

payment and delivery models through conventional experimental or nonexperimental methods.

In this article, we discuss how MACRA will significantly reduce the ability of the Centers for Medicare & Medicaid Services (CMS) to use current research methods to measure the impacts of APMs on the cost and quality of care. In particular, we describe why single-factor experimental or quasi-experimental designs, which have been central to evaluations at the Center for Medicare and Medicaid Innovation (Innovation Center), may be unable to provide definitive evidence of model effects when the use of APMs is widespread. We suggest fractional factorial randomized experiments as the best means of producing the evidence needed to guide the continued development of new payment and delivery models and to help calibrate payment parameters to achieve policy objectives. We also describe how perceived barriers to the use of such designs can be overcome. We conclude that, at this critical moment, when APMs are expanding, it is important that CMS and other policy makers consider which research and testing methods will best advance evidence-based policy for the Medicare and Medicaid programs in the new environment.

MACRA CREATES COMPLICATIONS AND OPPORTUNITIES

Under MACRA, eligible professionals (including physicians, nurse practitioners, and therapists) will be paid Medicare fee-for-service (FFS) rates that depend on three conditions: whether their services are furnished through a risk-bearing entity participating in an APM, whether they provide services as a medical home, and whether they meet electronic health record and quality requirements under the new Merit-Based Incentive Payment System. The Secretary of the U.S. Department of Health and Human Services (HHS) has set a goal for half of Medicare payments to be tied to APMs by 2018 (Burwell 2015). The CMS Office of the Actuary projects that APMs will account for 60 percent of payments for physician services in 2019 (Spitalnic 2015) and 100 percent by 2038. The Congressional Budget Office (CBO) estimates that enrollment in Medicare group plans will account for 36 percent of program

Address correspondence to Thomas W. Grannemann, Ph.D., Mathematica Policy Research, 955 Massachusetts Ave., Suite 801, Cambridge, MA 02139; e-mail: tgrannemann@mathematica-mpr.com. Randall S. Brown, Ph.D., is with Mathematica Policy Research, Princeton, NJ.

enrollment in 2019, growing to 40 percent by 2026 (CBO 2016). Even if slower than these projections, the rise to predominance of APMs of diverse forms, coupled with a smaller and less-representative traditional FFS component of Medicare, will make it difficult to find comparison groups unaffected by (the potentially biased) selection of patients and/or providers into these payment systems and models.

Selection Biases in Quasi-Experimental Designs Will Increase

As we enter the era of MACRA APM incentives, we can expect providers to sort themselves into the APM shared savings and risk arrangements that are the most financially advantageous to them. This self-selection and sorting may prevent quasi-experimental evaluations from producing unbiased empirical estimates of the effects of APMs or of incentives related to certain forms or payment parameters of these models.

Because APM-participating providers must assume some of the risk of higher costs to be eligible for enhanced payments through MACRA, we can expect that providers who opt to participate in an APM, especially the early participants, will be those most confident of achieving success and may already be ahead of the curve in embracing the transition to new forms of care delivery. Shortell et al. (2014) found physician organizations participating in ACOs to be “relatively large, well-networked practices with strong care management toolkits.” CBO acknowledged such selection effects in its scoring of the Strategic Growth Rate repeal and MACRA, and it even offers an indicator of the magnitude of selection effects; CBO projected a \$5.5 billion increase in Medicare costs over the next 10 years, in anticipation of the effects of providers selectively choosing between APMs and a value-based payment program (CBO 2014).

In quasi-experimental designs, the treatment and comparison groups must be comparable on all important factors, observed and unobserved, that are correlated with the outcomes of interest. In recent years, researchers have devoted considerable effort to developing analytic methods that minimize or compensate for differences between treatment and comparison groups. Techniques such as differences-in-differences comparisons of the treatment group to a matched comparison group, instrumental variables, and regression discontinuity methods all require rather strong assumptions about the structural relationships between the determining factors. The ability of such methods to produce valid results will become more dubious, however, if there are multiple, diverse, and complex uncontrolled confounding factors affecting the

comparison group, including those introduced by the simultaneous testing of many large models in separate demonstrations. As such demonstrations expand in size and scope, the best matched comparison cases may often be involved in one or more APMs, making it difficult to interpret results. Factorial experimental designs control such confounding factors by bringing all of the intervention options under a single design umbrella.

Simple Random Assignment to Treatment or Control Status Will Not Resolve This Problem

In the past, randomly assigning program applicants to the treatment or the control group in a test of a single intervention would eliminate the selection bias. Because the intervention could be offered to the treatment group only, and because demonstrations were relatively rare in any given geographic area, the control group provided an unbiased estimate of what would have happened to the treatment group members, on average, absent the intervention. But in the new world of widespread opportunities to engage in APMs, a provider assigned to the “control” group could—and likely would—just enter a different demonstration program or multiple programs. The researcher is then left with an ambiguous counterfactual—the treatment–control difference in outcomes represents the difference between providers exposed to the intervention being tested and those involved in other similar initiatives.

Opportunities

Under section 1115A of the Social Security Act, the Secretary of HHS has authority to test models and define the conditions for participating in such tests, including Innovation Center tests of APMs that will qualify for MACRA payment incentives. Given the benefits that will accrue to participating providers, the Secretary might reasonably require those who participate in the testing of models to comply with research protocols, such as being randomly assigned to an appropriate type of APM, as necessary to definitively evaluate APM effects.

Rather than cling to the no longer attainable concept of a pure, untainted control or comparison group against which to compare the treatment group, we propose the purposeful design and testing of alternative interventions or versions of a given model against each other. In this approach, as we discuss below, planned variations in payment model features and parameters are deliberately and systematically combined into a given number of

configurations, and each provider—participating voluntarily or as a requirement of participating in Medicare—is randomly assigned to operate under a particular configuration of such features.

LEARNING STRATEGIES FOR THE MACRA ERA

Evidence-based policy in the coming years requires the support of a research agenda that both directly addresses the most relevant policy questions and sets forth the best methods for answering these questions definitively. Even when early Innovation Center evaluations of models are complete, however, many policy questions related to APMs will remain. How strong do the incentives in shared savings and risk arrangements need to be? How much and what kind of quality reporting is worth the cost and administrative burden? How can payments be structured to support physician–patient decision making that considers cost-effectiveness over the profitability of services? Answering these questions may require testing more nuanced payment methods (multipart or blended, for example) that better align payment with value and balance incentives for the underuse and overuse of services. CMS should consider using experimental methods more often and more creatively, testing systematic variations in interventions with factorial designs, and establishing an advisory council or panel of research design experts (perhaps including provider and policy maker representation) to advise the agency about how major program models can best be designed, tested, and evaluated.

Shifting the Focus to Medicare Payment

In the early years of the CMS Innovation Center, top priorities included fostering organizational change and discovering new ways to deliver care (Centers for Medicare and Medicaid Services 2014, 2016). This mindset put the primary locus of learning about delivery models within the participating provider organizations. Fee-for-service Medicare has long paid providers according to a uniform structure, and providers compete for patients mainly on the basis of quality, reputation, and convenience—not price. The new models introduce a diversity of payment methods, rates, and incentives. They are intended to directly reward quality and reduce the incentives for volume through APMs such as bundled payment, ACOs, or medical homes. But there is uncertainty about the extent to which providers are responsive to incentives and whether FFS providers can truly alter patterns of service use without the

controls on network and service use that are available to managed care plans. The needed evidence can come from evaluation designs that explicitly test various combinations of incentives in ways that will reveal the most desirable mix of payment methods and parameters. This mixed or blended approach to payment seeks to balance the incentives of per-unit payment, which may encourage excessive utilization, and episode-based, bundled, or capitated payment, which may encourage care stinting or cost shifting when not checked by adequate competition.

As they learn within their organizations, providers may augment the knowledge base that improves the delivery of care, but specifying the payment part of “payment and delivery models” is not something that should be left to providers. The locus of learning, at least for designing, testing, and evaluating payment model provisions, must therefore lie with payers. CMS could design payment models that qualify for MACRA incentives while giving participating providers the flexibility to experiment with delivery methods and adapt as they learn. This approach would also allow CMS to take responsibility for constructing experiments in payment methods and rates that would support its own learning strategy for refining the payment aspects of its models. The question is, Which forms of APMs produce the best outcomes in different circumstances?

Sorting Out the Effects of Multiple APMs

Factorial designs are a natural fit for testing APMs under MACRA and for sorting out their effects. This is because the defining characteristic of most APMs, and the research challenge that comes with it, is that they include a mix of payment elements, care coordination fees, shared savings, episode-based payment, and incentives linked to quality. So each APM is characterized by a specific combination of payment factors, exactly as is required for the *factors*, *levels*, and *combinations* of factorial design experiments (see, e.g., Raktoe, Hedayat, and Federer 1981). Shadish, Cook, and Campbell (2002) state that factorial designs provide a systematic basis for comparing the many possible combinations of features in an intervention. Factorial designs are also known for their efficiency in distilling the effects of an intervention that is delivered at different intensities or when there are interactions or synergies between program elements (Cox 1958; Mee 2009; Collins et al. 2014). Donald Berwick (2008) includes factorial experiments among the methods that have “more power to inform about mechanisms and contexts” than do (basic) randomized controlled trials.

In the context of MACRA, APMs can be thought of as a diverse but related set of approaches to changing provider incentives and the delivery of care. But to date, each model, which represents a specific combination of these elements, has been tested as a one-off proposition, making it difficult to sort out just what features within each model are producing desired or undesired effects. If we think of the payment elements as distinct factors that can be tested systematically and experimentally, factorial designs enable us to sort out the effects of each element and, more important, to identify which combinations of payment elements work best together. At the same time, by eliminating the need for sequential studies to determine the effect of each component of an intervention, factorial designs can produce and promote more rapid feedback and earlier findings on what works.

Using Factorial Designs to Test Systematic Variations of Models

The next generation of the Innovation Center's tests of APMs could better inform public policy if the evaluation designs include multiple treatment groups that are structured to test program variants systematically against each other rather than against the status quo, in much the same way as comparative effectiveness research seeks to assess multiple treatment alternatives, while clinical RCTs often compare a single new treatment against the usual standard of care. For some models in which the level of the intervention can be varied, such as the level of care management fees paid to providers as incentives, researchers can design the experiment to compare the effects of different payment levels. For other models, a crossed or modified factorial design might be appropriate, for example, when the goal is to test the individual and joint effects of care management fees and outcome-based incentives. Factorial models are used routinely in marketing and industrial studies, and researchers have started to use them in evaluations of health care delivery options (Box, Hunter, and Hunter 2005; Zurovac et al. 2013).

Factorial (and orthogonal fractional factorial) designs are especially promising for assessing APMs under MACRA. The entities or health care professionals being studied—such as physician practices, hospitals, or care coordinators in a health plan—could be randomly assigned to a strategically chosen mix of payment parameters and intervention features. Even if there are too few intervention participants to test all possible combinations of intervention components against each other, researchers can use creative designs such as fractional factorial designs or efficient orthogonal designs to estimate the relative effects of each component—and some interactions of components.

Using a *fractional factorial plan*, researchers could identify the combinations of features that must be tested to obtain unbiased estimates of the first-order effects of each model feature or component (see, e.g., Dey and Mukerjee 1999). Such an experimental strategy would enable policy makers to learn which elements to include in an intervention *and* to evaluate the relative improvements in savings, quality of care, and other outcomes that are expected to result from investing more heavily in a given program component. These designs also ensure that the key payment provisions and program requirements to be tested are systematically varied across participants; when this does not occur, evaluators are left with the challenging task of linking ex post facto differences in outcomes to any observed differences in the implementation of such provisions. Table 1 shows the advantages and disadvantages of factorial designs compared with traditional and current designs.

Using Factorial Designs to Accelerate APM Development

The factorial design approach can accelerate learning about APMs because it provides for simultaneous testing of multiple versions of a model. For example, four dimensions of the payment intervention could be varied: (1) the level of the FFS payment for services rendered (such as visits), (2) the level of the per-case or per-episode payment for assessment and care management, (3) the payment for achieving quality benchmarks, and (4) the reward for savings achieved. Although the early CMS tests of APMs included only one or two variants of these parameters (with self-selection of provider participants into risk strata), with expanded participation in APMs under MACRA, multiple combinations of factors could be tested simultaneously (perhaps crossing a higher versus a lower levels of incentives for each factor); participants would be randomly assigned to one of up to 2^4 (i.e., 16) possible combinations of two incentive levels on each of the four dimensions.

This example is illustrated in Table 2. In this Karnaugh-map-type diagram, cells that share a border (including wraparounds top to bottom and left-most to right-most) differ in only one factor and so may make logical groups for comparing options and interpreting results. CMS could recruit, say, 240 qualified organizations to participate (the number would be scaled to reflect research objectives and statistical power requirements), randomly assigning each to one of the 16 possible combinations (15 per cell). (Many more variations could be tested concurrently with efficient orthogonal designs without filling all the possible cells.) Unbiased estimates of the overall effect of each of the four payment dimensions would be obtained by comparing outcomes for

Table 1: Alternative Approaches to Testing APMs in the MACRA Era

<i>Design (Model)</i>	<i>Advantages</i>	<i>Disadvantages</i>
<i>Pre-MACRA Era</i>		
Quasi-experimental design Most CMMI models: Pioneer ACOs, Comprehensive Primary Care (CPC), Bundled Payment for Care Improvement (BCPI)	<ul style="list-style-type: none"> • Providers can choose to participate in model and version • Can be implemented on a voluntary basis 	<ul style="list-style-type: none"> • Difficult to implement strong incentives in voluntary program • Can be difficult to ensure truly comparable comparison group
Traditional randomized designs (Some Medicare benefit demos of an earlier era; a few CMMI models: Million Hearts, Home Health Value-Based Purchasing, Medicare Care Choices)	<ul style="list-style-type: none"> • Can provide causal evidence of effect of a defined intervention 	<ul style="list-style-type: none"> • Requires mandatory assignment or participants to agree to randomization
Factorial experimental design (Proposed Medicare Part B Drug Model)	<ul style="list-style-type: none"> • Can test multiple interventions and compare results 	<ul style="list-style-type: none"> • Requires mandatory assignment or participants to agree to randomization
<i>Additional MACRA Era Factors</i>		
Quasi-experimental design	<ul style="list-style-type: none"> • Builds on recent CMMI experience 	<ul style="list-style-type: none"> • Inability to define an uncontaminated control group
Traditional randomized design	<ul style="list-style-type: none"> • Experimental design supports causal inferences 	<ul style="list-style-type: none"> • Inability to define an uncontaminated control group
Fractional factorial experimental design	<ul style="list-style-type: none"> • Experimental design supports causal inferences • Ability to consistently compare multiple interventions 	<ul style="list-style-type: none"> • Requires comprehensive up-front design • Requires mandatory participation or favorable incentives in all cells

the 120 participants assigned to the high option on that dimension (e.g., higher per-case payment) to outcomes for the 120 participants assigned to the low option on that dimension. For example, 15 organizations would be randomly assigned to receive the usual FFS rate for services, a higher per-case care

Table 2: 2 × 2 × 2 × 2 Factorial Design for Hypothetical APM Model

		Level of FFS Payments					
		High		Low			
Level of per-case or episode payment	High	1	2	3	4	Low	Reward for total Medicare savings achieved
		5	6	7	8	High	
	Low	9	10	11	12	Low	
		13	14	15	16		
		Low	High		Low		
		Payment for achieving quality measures					

Note: Shades identify rows or columns associated with the different levels of factors. [Color figure can be viewed at wileyonlinelibrary.com]

management fee, and a smaller reward for achieving quality and savings performance targets. Fifteen other organizations would be randomly assigned to receive a higher FFS rate for services, a lower per-case payment for care management, a smaller reward for achieving quality targets, and a greater share of any cost savings generated. Many other variants are also possible, such as tying the receipt of shared savings to the achievement of quality targets.

This approach would promote a quicker and more expansive understanding of the effects of alternative APM features and their interactions. A four-factor model with 2⁴ cells would provide estimates of four main effects (one for each of the factors) and a rich set of interaction effects. Many types of comparisons would be supported. We could compare, for example, the cells with the highest performance incentives (cells 6, 7, 10, 11) with the cells with the lowest performance incentives (cells 1, 4, 13, 16). Or we could compare the set of cells 2, 3, 14, and 15 with cells 6, 7, 10, and 11, to learn whether strong savings incentives diminish the effect of quality incentives.

Factorial designs need not be limited to two levels for each factor; there could be multiple levels of the FFS and episode-based dimensions of payment. Alternatively, these two factors could be combined into a single multiple-level mixed-payment-methods factor representing a sliding scale (with a higher FFS component at one end and a higher episode-based component at the other) to find the optimal mix that balances the incentives at the margin for over-providing care in pure FFS with incentives for under-providing care in episode-based payment. This approach could inform decisions about payment parameters and greatly shorten the time required to identify the combinations of model features that maximize the value achieved in improved

quality and cost outcomes. If implemented on a larger scale, the approach could also help to identify what combinations of payment features and incentives work best in particular types of markets, and for particular types of practices and patients.

Factorial experiments could be used to test a wide range of APMs, including ACOs, bundled payment, and medical home interventions. The findings would show the associated tradeoffs in cost and quality, and how payment parameters can be manipulated to achieve desired objectives, such as better patient outcomes, potentially avoiding years of payment policy guesswork and adverse incentives from poorly calibrated payment. In each type of intervention, it would be possible to systematically experiment with various payment parameters and other requirements.

Making Factorial Experiments Practical

The CMS Innovation Center has eschewed experimental methods for most of its major model tests, stating concerns that randomization (1) may be inappropriate for evaluating complex, multicomponent interventions, (2) could be infeasible if it leads to lower participation of providers, and (3) raises legal and ethical issues if patients are randomized to different levels of care (Howell, Conway, and Rajkumar 2015). Our view is that none of these are compelling reasons to choose nonexperimental designs over factorial experiments for APMs.

First, factorial designs are ideally suited to break down estimates of the overall effect of a complex intervention into the effects of each of its component factors and their interactions. Some may ask whether it would be too costly to test many variations of a model, given the need for sufficient samples in every factorial cell; or they may perceive such tests as being too administratively complex or requiring additional time to implement. However, the power of factorial design to detect effects depends on the overall sample size and the combinations model features that are being compared, not on the number in any one cell. Efficiency comes from drawing on results across many cells, so for the same sample size or cost many more variants of APMs could be tested with enough power to detect meaningful impacts (Dziak, Inbal, and Collins 2012; Pennsylvania State University, no date). Although the size and cost of factorial designs will depend on the need for precision, the time and effort required for administration and evaluation is no greater than for other strong designs or current Innovation Center models. In fact, evaluation tasks are greatly

simplified, compared with quasi-experimental designs, as there is no need to search for credible comparison groups or to adjust for or rule out correlation of the intervention with uncontrolled causal factors.

Second, on the issue of participation, factorial designs allow all or nearly all participants to engage in the intervention and so avoid difficulties of unengaged control or comparison groups. Such designs must nonetheless seek to encourage participation that is fully representative. Factorial experiments with voluntary participation must be structured to appeal to potential participants regardless of the model requirements or payment parameters to which they might be randomly assigned. Although all model forms need not be equally attractive to all providers, every cell to which a participant could be assigned (which might include a participation incentive payment) must be perceived as no worse than the participant's status quo alternative to prevent attrition after random assignment. These conditions can constrain the design of models to be tested on a voluntary basis, but they allow all participants to benefit from participation, if in different ways, because none are assigned to a pure control group.

There are tradeoffs in structuring the financial incentives for APM tests. Factorial experiments may produce internal validity with unbiased estimates of effects on participants but could impair external validity. This can arise if the design leads to a smaller and more self-selected set of participants than a design that does not leave demonstration applicants uncertain about the model features and payment parameters to which they will be assigned. Selecting the best designs for testing APMs in the coming years will require realistically assessing the benefits and risks of alternative designs and payment incentives.

Another way to ensure representative participation is to make participation mandatory for providers who deliver particular Medicare services in a specific geographic area. CMS has pursued several mandatory demonstrations for Medicare, including the Comprehensive Care for Joint Replacement Model, the Home Health Value-Based Purchasing Model, and the Medicare Part B Drug Payment Model. These have been proposed through the rule-making process, which gives providers and the public an opportunity to voice any concerns in advance. While mandatory demonstrations can be controversial if they entail unequal payment methods for providers, they can be an important tool for gaining unbiased evidence of the effects of distinctly different payment models. Mandatory factorial design demonstrations could be employed to efficiently and expeditiously test many variations, each

representing a change in payment that is substantively meaningful but small enough to be politically acceptable.

Third, on ethical and legal requirements, APM interventions do not entail different levels of benefits or prescribe forms of care for patients. Care delivery differences associated with APMs are no greater, and typically less, than those in many randomized design demonstrations successfully conducted by HHS/CMS in the past, such as the RAND Health Insurance Experiment (Newhouse et al. 1993), the National Long-Term Care Channeling Demonstration (Carcagno and Kemper 1988), the Medicare Case Management Demonstration (Schore, Brown, and Cheh 1999), the National Home Health Prospective Payment Demonstration (Cheh 2001), the Cash and Counseling Demonstration (Brown and Dale 2007), Demonstration of Informatics for Diabetes Education and Telemedicine (Moreno et al. 2008), the Medicare Care Coordination Demonstration (Brown et al. 2008; Peikes et al. 2010), the demonstrations of disease management for severely and chronically ill beneficiaries (Chen et al. 2008) and for dual eligible beneficiaries (Esposito et al. 2008), and the Medicare Health Support Pilot (McCall and Cromwell 2011).

Our view is that for APMs the benefits of factorial experimental designs, including stronger internal validity and the ability to test more variations and their interactions, will far outweigh the risks to external validity, which can be kept in check with creative design and effective implementation. The benefits of randomization for establishing causality are substantial. Within complex and interactive health care delivery systems, many factors that determine use, cost, and quality outcomes are highly correlated. Randomization provides the only fully reliable means to establish causal effects and definitively rule out the many confounding factors that may affect outcomes. Factorial methods that test a wide range of financial parameters at once may be the fastest way to build up evidence for the best model forms and parameters. Ideally, the parameters tested would span the range of policy-relevant forms of the model, but a narrower range of parameters could be tested if it appears that wide variation in payment would impede participation. While providers may object to short-term inequities associated with randomization, the benefits of accelerated learning in Medicare may more quickly lead to payments that better support care delivery. Policy makers in Congress and the executive branch should consider how much difference in payments among providers can be tolerated in the short run to produce evidence for a more effective program in the long run. Creative designs can help to reduce the disparities in the financial impact of being assigned to a particular demonstration cell.

Obtaining the strongest causal evidence of a program's effects often means investing a good deal of effort upfront in overcoming political objections and practical operational challenges with solutions that support a strong experimental design (Gueron 2002). CMS demonstrations in the past often randomized patients to test new benefits, but its only prior experience randomizing providers was in the home health demonstrations. Politically it can be more difficult to randomize doctors and hospitals than beneficiaries and home health agencies, whose interests may be less vigorously advocated in Washington. APM tests need not require provider participants to quit innovating in care delivery even if they are randomly assigned to defined payment models that differently support such activities. In fact, Medicare can best learn about what drives care delivery innovation by testing variations in model requirements and incentives.

A strong commitment to making experimental methods work will be needed, as Medicare evolves, to rigorously test APM payment incentives. Fractional factorial experiments combined with insightful qualitative evaluations to explain success and shortcomings in care delivery innovations may offer the best chance of supporting evidence-based continuous improvements, not only for providers but for the Medicare program itself. HHS has a compelling history of conducting randomized design demonstrations in which concerns and objections were overcome by well-considered designs and engaging participants in the experimental process. To advance payment methodology effectively and efficiently in the long run, CMS will need to test models with incentives strong enough to make a difference and evaluation designs capable of establishing causal relationships and accurately measuring those differences.

CONCLUSION: DESIGNS FOR THE TIMES

The expanding use of alternative payment models makes this a pivotal time to use broader and more effective learning methods to support evidence-based policy making in Medicare and Medicaid. Given the extensive participation and self-selection of providers into many APMs, the use of nonexperimental methods to produce unbiased estimates of their effects will soon become infeasible in many cases. Without solid evidence showing which combinations of APM payment and delivery elements yield the best results, Medicare could find itself on a long, meandering path of continual revisions to payment policy, chasing what are ultimately elusive solutions

to the problem of delivering affordable quality care. CMS and private payers need a surer, faster path toward definitive evidence of the most effective forms of APMs. This evidence can best come from well-designed factorial experiments because of the particular capability of this method to efficiently and rigorously test a wide range of payment features and parameters against one another in a systematic way.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The authors thank HSR reviewers, Eugene Rich and other Mathematica colleagues, and current and former government officials for helpful comments on earlier versions of this paper. Our employer, Mathematica Policy Research, allowed us time to prepare the paper as a professional activity.

Disclosures: None.

Disclaimer: None.

REFERENCES

- Berwick, D. M. 2008. "The Science of Improvement." *Journal of the American Medical Association* 299 (10): 1182–4.
- Box, G. E. P., J. S. Hunter, and W. G. Hunter. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery*. Hoboken, NJ: John Wiley.
- Brown, R. S., and S. B. Dale. 2007. "The Research Design and Methodological Issues for the Cash and Counseling Evaluation." *Health Services Research* 42 (1 pt. 2): 414–5.
- Brown, R., D. Peikes, A. Chen, and J. Schore. 2008. "15-Site Randomized Trial of Coordinated Care in Medicare FFS." *Health Care Financing Review* 30 (1): 5–25.
- Burwell, S. 2015. "Setting Value-Based Payment Goals – HHS Efforts to Improve U.S. Health Care." *New England Journal of Medicine* 372 (10): 897–9.
- Carcagno, G. J., and P. Kemper. 1988. "An Overview of the Channeling Demonstration and its Evaluation." *Health Services Research* 23 (1): 1–22.
- Centers for Medicare and Medicaid Services. 2014. *Center for Medicare and Medicaid Innovation*. Report to Congress, December 2014.
- Centers for Medicare and Medicaid Services. 2016. "Innovation Models" [accessed on May 18, 2016]. Available at <https://innovation.cms.gov/initiatives/index.html#views=models>
- Cheh, V. 2001. *The Final Evaluation Report on the National Home Health Prospective Payment Demonstration: Agencies Reduce Visits While Preserving Quality*. Princeton, NJ: Mathematica Policy Research Inc.

- Chen, A., R. Brown, D. Esposito, J. Schore, and R. Shapiro. 2008. *Report to Congress on the Evaluation of Medicare Disease Management Programs*. Princeton, NJ: Mathematica Policy Research.
- Collins, L. M., J. J. Dziak, K. C. Kugler, and J. B. Trail. 2014. "Factorial Experiments: Efficient Tools for Evaluation of Intervention Components." *American Journal of Preventive Medicine* 47 (4): 498–504.
- Cox, D. R. 1958. *Planning of Experiments*. New York: John Wiley.
- Dey, A., and R. Mukerjee. 1999. *Fractional Factorial Plans*. New York: John Wiley.
- Dziak, J. J., N. Inbal, and L. M. Collins. 2012. "Multilevel Factorial Experiments for Developing Behavioral Interventions: Power, Sample Size, and Resource Considerations." *Psychological Methods* 17 (2): 153–75.
- Esposito, D., R. S. Brown, A. Y. Chen, J. L. Schore, and R. Shapiro. 2008. "Impacts of a Disease Management Program for Dually Eligible Beneficiaries." *Health Care Financing Review* 30 (1): 27–45.
- Gueron, J. M. 2002. "The Politics of Random Assignment: Implementing Studies and Affecting Policy." In *Evidence Matters: Randomized Trials in Education Research*, edited by F. Mosteller, and R. Boruch, pp. 15–49. Washington, DC: Brookings Institution.
- Howell, B. L., P. H. Conway, and R. Rajkumar. 2015. "Guiding Principles for Center for Medicare & Medicaid Innovation Model Evaluation." *Journal of the American Medical Association* 313 (23): 2317–8.
- McCall, N., and J. Cromwell. 2011. "Results of the Medicare Health Support Disease-Management Pilot Program." *New England Journal of Medicine* 365 (18): 1704–12.
- Mee, R. W. 2009. *A Comprehensive Guide to Factorial Two-Level Experimentation*. New York, NY: Springer.
- Moreno, L., R. Shapiro, S. Dale, L. Foster, and A. Chen. 2008. *Final Congressional Report on the Informatics for Diabetes Education and Telemedicine (IDEATel) Demonstration: Final Report on Phases I and II. Final Revised Report*. Princeton, NJ: Mathematica Policy Research.
- Newhouse, J. P., and the Insurance Experiment Group. 1993. *Free for All: Lessons From the RAND Health Insurance Experiment*. Cambridge, MA: Harvard University Press.
- Peikes, D., A. Y. Chen, J. L. Schore, and R. S. Brown. 2010. "Effects of Care Coordination on Hospitalization, Quality of Care, and Health Care Expenditures among Medicare Beneficiaries: 15 Randomized Trials." *Journal of the American Medical Association* 301 (6): 601–16.
- Pennsylvania State University. no date. "Some Common Misconceptions About Factorial Experiments" [accessed on February 27, 2017]. Available at <https://methodology.psu.edu/ra/most/femiscon>
- Raktoe, B. L., A. Hedayat, and W. T. Federer. 1981. *Factorial Designs*. New York: John Wiley.
- Schore, J. L., R. S. Brown, and V. A. Cheh. 1999. "Case Management for High-Cost Medicare Beneficiaries." *Health Care Financing Review* 20 (4): 87–101.

- Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for General Causal Inference*, pp. 263–6. Belmont, CA: Wadsworth.
- Shortell, S. M., S. R. McClellan, P. P. Ramsay, L. P. Casalino, A. M. Ryan, and K. R. Copeland. 2014. “Physician Practice Participation in Accountable Care Organizations: The Emergence of the Unicorn.” *Health Services Research* 49 (5): 1519–36.
- Spitalnic, P. 2015. “Estimated Financial Effects of the Medicare Access and CHIP Reauthorization Act of 2015 (H.R.2).” Memorandum. Baltimore, MD: Centers for Medicare & Medicaid Services, Office of the Actuary, April 9, 2015.
- U.S. Congressional Budget Office. 2014. “Cost Estimate H.R. 2810 SGR Repeal and Medicare Beneficiary Access Act of 2013” [accessed on January 24, 2014]. Available at <https://www.cbo.gov/publication/45040>.
- U.S. Congressional Budget Office. 2016. “Medicare Baseline” [accessed on March 24, 2016]. Available at <https://www.cbo.gov/sites/default/files/51302-2016-03-Medicare.pdf>
- Zurovac, J., L. Moreno, J. Crosson, R. Brown, and R. Schmitz. 2013. “Using Multifactorial Experiments for Comparative Effectiveness Research in Physician Practices with Electronic Health Records.” *eGEMS (Generating Evidence and Methods to Improve Patient Outcomes)* 1 (3): article 5.

SUPPORTING INFORMATION

Additional supporting information may be found online in the supporting information tab for this article:

Appendix SA1: Author Matrix.