

Withholds to Slow Medicare Spending

A Better Deal Than Cuts

Jonathan S. Skinner, PhD

James N. Weinstein, DO, MS

Elliott S. Fisher, MD, MPH

CONSENSUS HAS EMERGED ON THE NEED TO SLOW Medicare spending growth, but there is no agreement on how best to do so. Although many approaches have been suggested, Congress is almost certain to consider across-the-board cuts in reimbursement rates. It is not hard to understand why: cuts are easy to implement and appear to deliver clear savings. But cuts in fees are a poor long-term solution to the problem of increasing health care costs. They make it more likely that physicians will decide not to accept new Medicare patients; further penalize already efficient systems; cause some to increase utilization to make up for revenue losses; and—most importantly—do little to encourage the collaborative efforts needed to improve health, better coordinate care, reduce regional duplication, and help beneficiaries avoid unnecessary care.

In this article, a “withhold” approach is proposed to slow Medicare spending. The objective is to achieve just as much savings while allowing motivated hospitals, physicians, and others who deliver Medicare services a way to recover the cuts and thus maintain, or even increase, their incomes. The idea is to return the withheld money—with interest—if cost growth in the health care region, or within the hospital or physician network, is less than expected. In the absence of such savings, Medicare keeps the money. Either way, Medicare saves.

Most readers are familiar with taxes being withheld from wages or even from government payments. Our withhold proposal is different, however, with several antecedents in health policy. During the 1980s, Germany instituted a payment system linking higher physician reimbursement rates to lower overall health care cost growth.¹ Withholds are also a key contributor to Grand Junction’s success in creating a shared commitment among physicians to improving care while restraining cost growth.² The key questions are, what is the “reference group” (ie, the relevant universe of patients used to determine how much of the withhold is returned), how much savings can be achieved, and who gets the money back?

The simplest reference group would be geographic areas such as hospital referral regions or states. However, the disadvantage of such pools is that the refund for an individual physician or hospital will depend on the behavior of others in the region with whom they have no existing relationship. Thus, Medicare should also allow accountable care organizations (ACOs) or other integrated health systems responsible for defined populations of Medicare beneficiaries to carve out their own pool, making it specific to their own patients rather than the region as a whole.

How much revenue could this proposal generate? Based on projected Medicare spending from the Congressional Budget Office (CBO) and our own calculations, a 6% long-term withhold plan could save more than \$400 billion over 10 years. Increasing the withhold to 15% over an initial 3-year transition period could generate more than \$1 trillion over a 10-year period.

How the Withhold Works for ACOs

Who gets the rebate under our proposed withholding system? To see how this would work for ACOs, consider 3 health care systems, A, B, and C. Each has projected 2015 Medicare payments for attributed patients of \$100 million. Medicare would withhold 6% of every bill submitted but would keep track of annual overall reimbursements for patients enrolled in (or attributed to) each of the systems A, B, and C. Suppose system A had managed to reduce its Medicare expenditures for patients under its care from its projected spending of \$100 million to just \$90 million. System A’s total reimbursements from Medicare would be \$90 million less the \$5.4 million (6% of \$90 million) held in escrow, for a total of \$84.6 million. Because system A saved so much money, it would receive the entire \$5.4 million back (plus interest), assuming essential quality targets are achieved, so the system is made whole and receives the original \$90 million—no price cuts. Medicare saves \$10 million relative to projected costs, a 10% saving.

Author Affiliations: Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, New Hampshire (Drs Skinner, Weinstein, and Fisher); Department of Economics, Dartmouth College, Hanover, New Hampshire (Dr Skinner); Dartmouth Hitchcock Medical Center, Lebanon (Dr Weinstein); and Department of Medicine, Dartmouth Medical School, Hanover (Dr Fisher).

Corresponding Author: Elliott S. Fisher, MD, MPH, Dartmouth Institute for Health Policy and Clinical Practice, 35 Centerra Dr, Lebanon, NH 03766 (elliott.s.fisher@dartmouth.edu).

System B is not as successful in cutting costs. This system bills \$98 million but only receives that amount less the 6% withhold (\$5.88 million), or \$92.12 million. Medicare returns \$1.88 million of the original \$5.88 million withheld, so system B ends up receiving \$94 million, or 6% less than the original projected spending. Once again, Medicare achieves its targeted savings.

System C is completely unsuccessful in cutting total per-beneficiary costs and bills \$102 million. It gets none of the 6% withhold back. This is therefore equivalent to a 6% cut in reimbursement rates. Medicare ends up paying \$95.88 million, above the \$94 million target amount, which is what would have occurred under a 6% price cut. Thus, the withhold approach is no worse for system C and is better in terms of revenue gains for both systems A and B.

How the Withhold Works at the Regional Level

As noted earlier in this article, Medicare would withhold 6% from every bill it pays, whether to physicians, hospitals, home health care agencies, or medical equipment suppliers. Each physician, hospital, or other supplier in turn would receive a rebate based on whether cost growth is contained within its own region. The regionally targeted withhold would encourage multi-stakeholder collaboratives to address excess spending, including not only inadequate quality of care, but also poor population health and duplicative services.³ Ostrom received the 2009 Nobel Prize in Economics for her work on how local self-governed initiatives have successfully addressed complex social challenges such as declining fisheries or fresh water supplies.⁴ There are some hints, at least, that similar processes might be able to help in health care,⁵ and a regional withhold would create strong incentives for such initiatives to emerge more broadly in health care.

This withhold proposal has advantages over simply cutting reimbursement rates. At worst, if no regions or physician groups achieve savings, the withhold acts just like a cut in reimbursement rates, with identical revenue effects. However, the withhold proposal is better than a cut in the reimbursement rate because it gives physicians and others a way out: by improving care for their patients (through better population-based prevention, care coordination, and chronic disease management) and by scaling back on costly or duplicative services.

This proposal differs in 3 ways from the current sustainable growth rate (SGR) formula. First, the rates are set at a regional or ACO level, thus creating stronger incentives to

reduce inefficiencies than at the national level. Second, the withhold would apply to every bill paid by Medicare. This would allow Congress to replace draconian SGR cuts—imposed only on physician payments—with smaller withholding rates spread across all Medicare spending (and all who deliver Medicare reimbursed services). Third, a withhold ensures that Medicare already has the money and does not need to impose unpredictable cuts after the fact. Indeed, all withhold rates can and should be announced well in advance.

Further modifications deserve consideration. For example, different spending growth targets—and thus withholds—could be set for regions where there is more or less potential for achieving savings, such as high- or low-cost regions. Another possibility would be to set more generous growth targets (ie, lower withhold rates) for clinicians, practice groups, hospitals, or integrated systems that join Medicare-sponsored ACO-like reforms, such as the Medicare Shared Savings Programs. This would provide stronger incentives for physicians and others to move to more integrated delivery models. The key, however, is to ensure that the reform affects all Medicare spending, allowing Medicare to save considerable amounts of revenue with relatively modest rates of withholding. Also, a defined set of essential quality measures could ensure that physician groups and hospitals do not skimp on quality to gain cost savings.

This proposal does not correct a variety of other long-standing problems, such as Medicare's continued tendency to pay too much for some specialist and facility-based services and too little for others, including primary care. The withhold approach, however, can help keep the US Treasury solvent, provide meaningful incentives to reduce overutilization, and improve the coordination of care. And it's a much better deal for physicians.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Henke K-D, Murray MA, Ade C. Global budgeting in Germany: lessons for the United States. *Health Aff (Millwood)*. 1994;13(4):7-21.
2. Bodenheimer T, West D. Low-cost lessons from Grand Junction, Colorado. *N Engl J Med*. 2010;363(15):1391-1393.
3. Lucas FL, Siewers A, Goodman DC, Wang D, Wennberg DE. New cardiac surgery programs established from 1993 to 2004 led to little increased access, substantial duplication of services. *Health Aff (Millwood)*. 2011;30(8):1569-1574.
4. Ostrom E. Beyond markets and states: polycentric governance of complex economic systems. *Am Econ Rev*. 2010;100:641-672.
5. Gawande A, Berwick D, Fisher E, McClellan M. 10 Steps to better health care. *New York Times*. August 12, 2009:A27.