Can performance measurement make health care systems more sustainable? Or at least more efficient?

by

Joseph White

In light of the many discussions advocating the use of pay-for-performance and performance budgeting, this paper argues that discouraging experience with both approaches should temper expectations that performance measurement can be a reform that will make health care systems more "sustainable" or even more efficient. The link between sustainability and efficiency is tenuous, and attention to performance is not new. Measurement's accuracy tends to be overstated and its costs understated or ignored. Nor does it easily lead to changed behaviour. Yet some measurement is useful for managing any complex activity. In particular, there are situations in which measures are more accurate and the proper responses to shortfalls are generally agreed. Policy-makers should look for those conditions and encourage the more limited, targeted improvement that measures then can make possible.

JEL codes: H51, I11, I18

Key words: Performance, measurement, budgeting, sustainability

Joseph White is Luxenberg Family Professor of Public Policy at Case Western Reserve University. His research has focused especially on U.S. federal budgeting and on policies to control health care costs in the U.S. and other advanced industrial democracies.

Introduction

One goal of the Joint Network on the Fiscal Sustainability of Health Systems is for health policy experts and budget policy experts to learn from each other's experiences. This article addresses a wish common to both: that performance measurement will improve value for money, so make it easier to pay for what people need. In health policy the dominant hope is that sick persons or their insurer agents will "pay for performance" (P4P) rather than for mere activity or "being there" (Busse 2016: 1123). In the budget world, the parallel idea is "performance budgeting" (PB), in which programmes will receive more money if they produce more value, or the portfolio of programmes will be adjusted to maximise the return for total spending. P4P allegedly would improve on "fee-for-service" and other payment mechanisms. PB advocates claim it improves on "paying for inputs" rather than outputs.

These ideas are so attractive that they persist even though the extensive literatures on both PB and P4P show frequent failure and disappointment. On the budgeting side, Allen Schick (2013: 2) described PB initiatives as "often tried, but rarely successful." OECD's own work has shown that performance budgeting has tended to be "performance-informed budgeting" or "presentational performance budgeting" rather than allocation based on performance. Nevertheless, budgeting for "performance and results" is a core focus of the work of OECD's network of Senior Budget Officials.[1] Reviews of P4P show weak or mixed results, and weaker results with stronger studies (Busse 2016: Markowitz and Ryan 2016; Mathes et al. 2019; Mendelson et al. 2017; Ogundeji et al. 2016; Sullivan and Soumerai 2018). Performance measurement has rarely been linked to budget allocation for health care programs (Beazley et al. 2019).

Yet policy-makers and analysts still hope health budgets can become more "performance-oriented" (OECD Joint Network 2019). This interest in part reflects the fact that at some level, as R.G. Bevan and Christopher Hood wrote, "government by targets and measured performance indicators is a form of indirect control necessary for the governance of any complex system" (2006: 518). Measurement is not only a way to limit spending; it can also be used to build support for programmes and motivate employees. "If you have tested more people for lead poisoning this year than last, or if you have rid more premises of garbage or increased the number of prison doctors by fifty percent," Gordon Chase advised managers, "then say so" (Chase and Reveal 1983: 160).[2] Numbers are a fundamental aspect of policy-making discourse, whether accurate or not (Stone 2012: Chapter 8).

The goal of this article, therefore, is to explain why experience in both budgeting and paying for healthcare should lead to much lower expectations from performance measurement, and identify the conditions for modest benefits.

Sustainability, efficiency, and focusing on performance

We first should distinguish "sustainability" from "efficiency." Efficiency is an analyst's term. An activity could be popular and growing, but analysts could consider it a waste of money. Sustainability refers to whether a given level of activity can attract the inputs needed to produce it, without some terrible side effect in budgeting, that operates at two levels. System-wide is whether total spending can attract the taxes needed to pay for it without dangerous levels of borrowing. For programmes, it is the political system's willingness to fund the programme without restrictions that are widely viewed as harmful – e.g. waiting lists for care, or allowing an epidemic.

The budgetary sustainability of health care systems is a political, not an economic, question. Economic growth should not be related to how much is devoted to medical services as opposed to food consumption or housing or recreation. Medical expenses are jobs and income too.[3] When the U.S. spends much more than other countries, that is inefficient because it gets little value for the extra money, but comparatively higher spending does not appear to create comparatively worse economic performance.

The political challenge of health care finance is how to collect the money. Any hopes that performance measurement will make it easier to collect the revenue seems naïve. As Allen Schick (2013: 25) argues, "a government that candidly reports on its performance is likely to face more opprobrium for shortfalls in results than applause for its favourable accomplishments. There is little basis for expecting improved performance to itself elevate trust or confidence levels." In health care specifically there is little reason to believe that voters assess overall performance in the ways that analysts do. For example, analysts tend to care about population health statistics, while the voters and consumers care far more about personal access to what they believe to be rescue.[4] Public expectations may rise more quickly than services improve, "so that the public thinks that they're getting less when in fact they're getting more" (Martin Roland in Galvin 2006, w417).

The argument that "value for money" is too low carries much more weight in the attitudes of policy-makers and other elites: it involves their beliefs. If anything, making this argument seems likely to reduce ability to raise revenues. In the United States, opponents of redistribution to pay for health care frequently argue that, since care is weakly related to health outcomes, expanding access for the uninsured is not so important.[5] Higher-income voters seem more likely to believe they have a moral obligation to help lower-income voters if they are not told they are getting low value for their morality.

In short, better performance is weakly related to sustainability because efficiency and sustainability are different, the sustainability issue is political rather than economic, and focusing on alleged low-value-for-money can directly reduce political sustainability. This is not to say that efficiency should not be pursued, but that sustainability is not the reason to do that.

Advocates for performance budgeting normally claim that traditional budgeting focuses on inputs rather than results. The OECD notes that performance budgeting "represents a profound change in the character of the budget process, from a traditionally closed domain of budget specialists, focused on the numbers, to a more accessible, transparent and multi-disciplinary exercise." (OECD 2019) Yet the core process of traditional agency budgeting has been agencies justifying requests for inputs *by claiming that will lead to better outputs.* Changes to entitlement programmes often are justified by policy analysis of their likely effects – as in the many savings in the U.S. Affordable Care Act that were based on reports by MedPAC and other analysts (White 2018). Allen Schick argues that the many achievements of governments "before the first formal PB systems were launched" show that governments did focus on results before. Moreover, "there are often less differences between input-based and performance-based budgets than appears on paper." For instance, "it is difficult, sometimes impossible, to determine whether the amounts budgeted by governments are optimal when it lacks market prices or data on the cost of inputs" (Schick 2013: 5, 6). In practice, governments cannot be managed without decisions about what inputs to fund; agency personnel, equipment, and activities are inputs. Budgeting that purports to de-emphasise inputs only shifts the locus of control of inputs, from central authorities to agency managers.

Similarly, claims that new measurement create a focus on performance deserve scepticism. At a minimum, *reputation* for performance is rewarded in any system in which patients have some choice of provider and pay per service. Those with better reputations get more business. The core argument for measurement then should not be that FFS does not reward performance, but that whoever chooses (patients, or physicians through their referrals) tends to misjudge performance. While this is quite possible, we should remember that criteria such as providers' credentials, their facilities, or what people you know report about their experiences are logically relevant. Thus accreditation agencies may make those factors part of how they determine approvals, while patient satisfaction surveys in essence are attempts to quantify word-of-mouth reputation.

Also similarly to budgeting, the distinction between funding "activity" and funding "performance" is hard to maintain. In practice, a great deal of P4P is actually paying for activity. The most frequently (and legitimately) cited example of P4P having arguably positive effects, the NHS Quality and Outcomes Framework, paid GPs for documenting provision of specific services.[6] OECD's own assessment of P4P (OECD 2010: 107-108) emphasised increasing activity:

> *"results from P4P schemes suggest what common sense would tell you: quality of care increases when you pay for it. Paying for preventative and public health services appears to be particularly effective and can increase coverage of cancer screening, vaccination rates, etc. Often, primary care physicians neglect preventative services such as screening for cancer, measuring blood pressure and treating it, counselling patients to stop smoking or to improve their diet. The most successful P4P programmes pay additionally for providing these services or reaching some target."*[7]

In theory, P4P would pay physicians and hospitals based on measured improvements in patients' conditions, however achieved. Most P4P schemes however, do not operate this way for many reasons, beginning with the measurement difficulties we consider next.

Measurement

Whether applied to budgeting or payment for health care, performance measurement is part of a "recurring trope of public service reform" as described by Christopher Hood and Ruth Dixon (2015: 44-45): "the argument that policies to improve public management and service delivery ought to be based on proper evidence of 'what works" and that 'proper evidence' is often taken to mean well-developed performance numbers." They highlight one difficulty that has broader implications: governments continually alter the measures they use, "changing administrative record-keeping in ways that make it impossible to make before-and-after comparisons" (45). Anyone familiar with measurement in either the NHS or American Medicare could surely cite numerous examples. For example, one of the U.S. Hospital Inpatient Quality Reporting Program process-of-care measures had eight specification changes between 2006 and 2015 (Parast et al. 2015). The very frequency of change suggests concerns that any given set of measures is inadequate. As Bevan and Hood (2006: 520) argue, the case for measurement begins with core assumptions that, "measurement problems are unimportant, that the part on which performance is measured can adequately represent performance on the whole… that distribution of performance does not matter... [and] that this method of governance is not vulnerable to gaming by agents." All these assumptions are difficult to meet.

Measures are often statistically invalid or based on unreliable data. Measures that focus on adverse events often involve fairly rare events, so that results are not stable. This is made worse if, as in private sector P4P in the United States, the data includes only a portion of a medical provider's practice (McDonald et al. 2009). If providers collect the data, it may be gamed. Indeed, the policy might even encourage manipulation in the providers' self-interest. On the other hand, data collected by outsiders may (often for good reasons) be distrusted by the providers. For example, administrative data used for billing may be mined to evaluate performance, but is often incomplete or flawed. These problems are illustrated in U.S. Medicare's Merit-based Incentive Payment System (MIPS). MedPAC (2018a: 446) recommended abandoning MIPS because, with "quality of care and payment adjustments for quality" that are "based on measures that clinicians themselves choose to report," the system "will be inequitable because clinicians will be evaluated and compared on dissimilar measures" – while, "in addition, many clinicians will not be evaluated at all because, as individuals, they will not have a sufficient number of cases for statistically reliable scores." Letting physicians select scores for their own evaluation may seem like asking for trouble. Yet it can also seem a necessary response to the representativeness problem.

If measures are too narrow, there is a danger of "hitting the target and missing the point" (Bevan and Hood: 521). They provide a particularly scary example: that, "the waiting time for new ophthalmology outpatient appointments at a major acute hospital had been achieved by cancellation and delay of follow-up appointments… as a consequence, 25 patients lost their vision over two years." The most scandalous failures in the NHS would not have affected hospitals' star ratings, because the types of mortality involved were not measured (532-33).

In order to measure performance by an individual, organisation or programme, metrics should address not the final condition of the patient or population but the change that can be associated with the measured activity. That, however, requires measuring the original state of the system, not just the outcome. In health care this is the "risk adjustment" problem: one hospital's cardiac patients may have worse outcomes than another's because the first hospital has a deserved reputation for quality and so attracts sicker patients. Similar problems exist for many government programmes, such as crime control and education.

Adjusting for underlying conditions, however, expands the measurement challenge. In health care, the risk is defined by diagnosis and testing, that are usually performed by the same providers who are to be held responsible for the degree of cure. Ensuring accurate diagnosis is a measurement frontier that has received little attention (Berenson and Singh, 2018). The potential for gaming is obvious: for example, if a provider is paid based on a patient's risk profile, there is a strong incentive to diagnose as much illness as possible.

OECD's "User's Guide" to PB identifies a further difficulty: identifying whether poor performance means a programme is operated or designed poorly so should be cut, or faces a tough task that requires more resources. Therefore, (2008: 5), "in most cases the finance ministry does not use performance results to financially reward or punish agencies… poor performance may not be the agency's fault; poor performance caused by underfunding would hardly be improved by a further cut in funds."[8]

Bevan and Hood identify the risk of "gaming by agents," and there are many dramatic examples of documentation misleading because it is dishonest. In defence policy, these range from cover-ups of friendly fire incidents to false results of tests on equipment.[9] But the broader problem is that documenting is not the same as performing, involves extra work, and is independently shaped by payment incentives. In the NHS QOF, measured performance (and so payments for it) far exceeded expectations, partially because estimates

of baseline levels were too low but also because administrative staff was increased so as to improve reporting (Roland and Campbell 2014). When payment was reduced recorded performance also declined, but whether practice changed as much is less clear (Minchin et al. 2018).

Measurement becomes more difficult as the object of measurement becomes more complex, and so is especially challenging for health care or across a government.

The variety of tasks and of conditions that could affect performance for a single hospital, never mind a wider health care system, far exceeds what is faced in even technologically sophisticated production organisations like an aircraft carrier or nuclear power plant. Doctors and units in hospitals do many, many different things. Because practices and their patients differ so much, it is very hard to define measurements that are appropriate for a wide range of clinicians. It is very hard to compare performance of a police force and a primary school, but it is hardly easier to apply the same measures to a paediatrician and a cardiac surgeon. The result is that measures which are useful for managing performance within a unit (both in budgeting and health care) can rarely be used to compare performance across units. In the words of an OECD review (Shaw 2015: 5), "the line ministries themselves" are best able to identify measures that fit their programmes. "However, from the budgeting perspective, information generated in this highly decentralised way does not provide central, comparable data on performance, thus prohibiting strategic comparisons necessary in budget allocation processes."

This paper does not suggest that measures for a particular activity, such as ambulance response times, are not useful for assessing that activity – though even such statistics can be gamed (Bevan and Hamblin 2009). However, the immense variety of activities in even a moderate-sized government, or health care system, elicits a blizzard of measures that can overwhelm both those who seek to manage a system (e.g. Central Budget Authorities or managers of sickness funds) and the operators who do the work. In large and complex systems, accuracy and representativeness are inconsistent with manageability.

Although measurement is challenging enough, there are other reasons why improving performance through measurement is difficult.

PB and P4P in practice

As a budgeting tool, performance measurement could help save money in a more "sustainable" way (that is, without political blowback) by either guiding cuts in times of fiscal stress, discouraging incremental increases, shaping the limited incremental increases so they are most likely to increase value-for-money, or helping agencies manage more efficiently and therefore request less to do their jobs.

Yet cuts in response to crisis normally are "intended to be implementable in short order" and therefore, "not focused on efficiencies" (Shaw 2015: 24). Schick (2013: 9) summarised experience across countries, saying that the downturn from the Great Recession had not "swayed governments to emphasize performance issues in budget negotiations…negotiators have other things on their minds when they are pressured by time and fiscal constraints to hammer out a budget agreement." Similarly, a report on U.S. states which referred to performance budgeting as a "noble idea" noted that when "in crisis mode" states "resorted to mainly across-the-board cuts, furloughs, layoffs, and in some cases tax increases to attempt to achieve balance… both effective and ineffective programs are treated equally." (IBM 2011).

This is not to say that evaluation never affects decisions, but the process of changing minds about programmes does not fit well with the budget cycle. Major changes – "shift points" or "punctuations" in a previous budgetary "equilibrium" – are generally driven from outside the budget process (White 1994). In Schick's (2013: 11) words, "the budget is not the main driver of change but the means of accounting for changes made by other means."

Studies of budgeting for health care in particular show that experts' conclusions have at best moderate effects on allocations.[10] Experts, as in Canada, may hope that population health spending will reduce reliance on the acute care system, but that, "is more commonly portrayed as an ideal objective than as a realistic one." (Abelson et al. 2017: 9). Indeed, after the financial crisis, "public health was an easy target for budget cuts and curative services were more successful in holding on to (and increasing) financial resources" (Rechel 2019: 26). Leading researchers report that analytic priority-setting frameworks rarely lead to "successful disinvestment," and "almost none" make "claims for improved efficiency and equity" (Angell et al. 2016).

Although requirements to present performance information appear to have spread to more countries over time (Keller 2018), the scope and complexity challenges of PB have been particularly evident in countries that have tried harder to do it. "Countries with the most experience with performance budgeting" therefore "have steadily reduced the number of programmes and indicators over time" (OECD 2019: 39). "Performance budgeting momentum has in many cases, slowed under the weight of its own expectations," another OECD overview concluded, so that, "the number of performance indicators is being consolidated to provide more meaningful data metrics and reduce onerous reporting burdens" (Shaw 2015: 8). "Information overload," Allen Schick notes, “is a chronic problem in the time-compressed, deadline-driven world of budgeting,” and "exacerbated when PB adds new data, classifications and analyses to the old." PB "almost always increases the costs of generating and processing budget information," while within the agencies, "PB becomes discredited when spending units which produce much of the information perceive that their efforts have been in vain" (Schick 2013: 11, 12, 11).

Similarly, PB in American states has had weak effects at best for reasons mentioned above, such as: "difficulty in gaining clear agreement among stakeholders on the primary purpose of programs and activities," the limits of what is measurable, the weak links between agency performance and program outcomes, "budget decisions being made on the basis of priorities of elected leaders," and "incentives for agencies to choose easily achievable targets, or cheat in the use of measures" (Kamensky 2014). In all countries, it is most likely to shape budget allocations at the margins under favourable economic conditions, because "expansive budgets have sufficient space to accommodate both allocations based on evidence and allocations based on politically-expedient responses to voter preferences and group demands" (Schick 2013: 13).

As noted above, the numerous reviews of P4P have given little reason for optimism that it will fix health care. There are many potential avenues by which performance measurement could encourage providers to improve. Their managers might be subject to direct threats, as in the "targets and terror" approach in the NHS (Bevan and Hood 2006). Customers might switch providers in response to published (e.g. "star") ratings of performance – though coming up with accurate ratings that are aggregated enough to be usable by consumers is very difficult and they might not even respond to safety scandals (Bevan and Hamblin 2009; Hibbard 2008; Lavery et al. 2012). Medical professionals might learn from

measurement that they are not as good as they thought, and so realise they could do better – out of professional values or concern for reputation.

These benefits, however, require that the measures be accurate and believed and that is not easy, especially because there is often dispute about defining good practice.[11] The United States, again, may offer a worst-case scenario because measures and rankings come from multiple sources, which tend to disagree. In the U.S., four respected national hospital rating systems generated quite different results: "eighty-three hospitals were rated by all four rating systems, with no hospital rated as a high performer by all four. Only three hospitals were rated as high performers by three of the four systems" (Austin et al. 2015: 427.). Different reputable data sources for hospital surgery care yield very different lists of outliers (Lawson et al. 2015). The quality classifications in one of the most highly-publicised examples, joint replacements in California, involve quite different lists from different insurers.[12] These differences suggest that measurement is hard, and so one should not assume the single source in some other system is accurate.

In a few cases measurement may have encouraged major improvements. Perhaps the case that was most influential in promoting further measurement efforts involved cardiac artery bypass graft (CABG) surgery in New York state: hospitals appear to have responded to low ratings by improving their own processes (Chassin 2002). Yet even this success story must be tempered. Later reports suggest that physicians who are ranked on their performance have begun to avoid more difficult patients, leading even to greater mortality for patients who need percutaneous revascularisation in states with reporting processes than in nearby non-reporting states (Waldo et al. 2015; also see Rosenbaum 2015).

In fact, P4P may involve more serious risks than occur with performance budgeting. Performance budgeting routines can seem wasteful and pointless to the people who operate programmes, but generally do not interfere with most of the agency operators' work. In the case of health care, however, collecting information to measure performance can be much more intrusive, because the information is normally recorded by those caregivers.

Measurement burdens interact with the failings of electronic medical records (EMRs). In the U.S., one report by leading researchers (McGlynn et al. 2014: 2150) concluded that,

> *"Physicians, hospitals, and health plans view measurement as burdensome, expensive, and indifferent to the complexity of care delivery. Patients and their care-givers believe that performance reporting misses what matters most to them and fails to deliver the information they need to make good decisions. In an attempt to overcome these troubles, measure developers are creating ever more measures, and payers are requiring their use in more settings and tying larger financial rewards or penalties to performance. We believe that doing more of the same is misguided…"*

"Measurement fatigue" (Cassel et al. 2145) should be especially severe in the U.S., where providers are subjected to measures by many different payers. The result of excessive measurement is now being described as an epidemic of "physician burnout" (Noseworthy et al. 2017; Jha et al. 2018), as physicians are even reported, in some studies, to spend more time on documentation than patient care.[13] Documentation burdens may be less severe in other countries, but still, as with the NHS QOF, can require substantial resources.[14]

P4P can also create negative, rather than positive, incentives for quality. Donald Berwick (1995) once described "pay for performance" as "toxic to true organizational performance." He explained that the core idea, merit pay, has long been known to lead to making "the supervisor the customer," suppression of possibly harmful information, can inhibit

co-operation, costs a great deal to administer, because full accuracy is impossible will seem unfair, and can reduce intrinsic motivation for quality by putting price tags on everything.[15] Woolhandler, Ariely and Himmelstein (2012) criticise the underlying theory of incentives. "The quality improvement literature," they note, "has pinpointed many causes of quality breaches in medical care… But 'not trying' is rarely cited". Yet P4P implicitly blames lack of motivation for poor quality care." Monetary rewards not only may miss the point but may also backfire. "Tangible rewards – particularly monetary ones – undermine motivation for tasks that are intrinsically interesting or rewarding." P4P creates the equivalent of detailed contracts, making expectations for care far more specific than in the past. Transaction cost economics however, shows that making more complete contracts raises administrative and legal costs, while lists of rules "implicitly permit everything else." Therefore, while "injecting different monetary incentives into health care can certainly change it," that is "not necessarily in the ways that policy makers would plan, much less hope for."

To summarise, experience with both performance budgeting and P4P give little reason to expect performance measurement to make health care systems more than marginally more efficient, never mind more sustainable. Yet we should return to the earlier point that measurement is used in many situations, and can have some positive effect. In what health policy situations, and how?

Conditions for modest success

As a general rule, it is much easier (though not easy given gaming) to assess activity than outcomes. P4P therefore will be more successful if the goal is to increase activity and if activity can be increased at acceptable cost.

Measures to encourage specific activities must involve manageable and accurate measurement, and not ask providers of care to attend to more than a few items at a time. So they must be limited in number. They also need to ask people to do things they know how to do. If people are punished for failure they cannot avoid, they will subvert and deceive – and probably should. Worse, this will change their underlying attitudes and worsen performance on other tasks (Bevan and Hood 2006).

We can see these principles in some examples of positive change, and in some more mixed examples. The U.S. Hospital Inpatient Quality Reporting Program targeted process-of-care measures, mostly for treatment of cardiac patients. Between 2003 and 2015, most improved so dramatically that they "topped out." They were relatively clear, simple, cheap, and could be put on a check-list (Kahn et al. 2015).[16]

Perhaps the most broadly-adopted reform which could be interpreted as P4P is the international move to "activity-based payment" – often payment by DRGs – for hospitals, rather than fixed budgets. A full discussion of DRGs is beyond the scope of this paper (one is Busse et al. 2010); we should as a start remember that DRGs can be used as a management tool in budget-making instead of as a direct payment mechanism. Activity measures are useful for traditional budgeting by giving a way to compare inputs to outputs. Budget control in a DRG payment system requires some limit on payment for activities, such as overall fee reductions if volume exceeds targets (a volume-related fee schedule) or a system in which services beyond some total receive lower fees. It is clear however, that policy-makers have sought, with some success, to increase hospital productivity by paying for this version of "performance." Broader effects on population health are best described as uncertain.

An especially complex example of hospital payment for activity, a form of case-mix funding, was implemented in Victoria, Australia beginning in 1993, and associated with some savings (Duckett 1995). A baseline activity level was calculated and hospitals paid more, but only about half of the average payment, for increasing services. This was meant to be attractive because the marginal payment exceeded marginal cost. The system was designed to separate out budgets for non-care functions such as capital, research, and education. Savings were not due solely to this approach: Duckett reports that labour law changes also made it easier for hospitals to reduce staff, while the government funded buy-outs. Later observations suggested also that the buffering of non-care functions had not worked as intended: in essence, hospitals found it easier to make budget savings by raiding research and other funds, rather than increasing efficiency.[17] Nevertheless paying for incremental activity probably increased activity, and in a context in which hospitals had to increase activity to compete for scarce funds, should have improved efficiency. No effort was made to relate the funding to health outcomes.

Another major "value for money" initiative implemented in many countries has been to create bundled payments for most of the care for a given condition, with payments adjusted according to quality measures (OECD 2010, Chapter 5; Srivastava et al. 2016, Chapter 3). The goal has been to reduce incentives for excess treatments while encouraging relatively low-cost interventions that are believed to increase value. This approach has been implemented most widely for patients with Type 2 diabetes (Type 1 patients are more difficult to manage so normally excluded). Diabetes is one disease for which there are fairly widely-accepted process of care measures, intermediate outcome measures, and definitions of who qualifies as a patient. Some of the measures therefore may be included in rating the performance of primary care physicians (see, e.g., Mousques and Daniel 2015).

Dutch policymakers encouraged formation of Chronic Care Groups, consortiums of providers that would be paid bundles for treating diabetes, cardiac risk, and/or COPD. The diabetes initiative attracted far more participation because physicians agreed there were measures that were good practice but not everyone was doing. Implementation for cardiac care was limited because of diagnosis disagreement: the doctors wanted to describe the at-risk group far more broadly than the insurers could accept. Doctors were less willing to participate for COPD because it involved many fewer patients, so greater administrative costs per patient, and there was less agreement about treatment methods. Care groups for Type 2 diabetes appear to have adhered more fully to recommended processes of care. Outcomes are less clear because, while death rates were lower in the "treatment" group, there are strong indications that sicker patients were disproportionately excluded. There appears to be no sense that there were meaningful savings.[18]

Treatment of dialysis patients in the United States should be another opportunity to improve performance through measurement. Dialysis is a routine and frequent treatment with clear standards of quality. Compared to other countries, U.S. outcomes for dialysis patients have been poor, and this can be explained in part by practice patterns. For example, the U.S. has tended to use dialysis catheters more, and fistulas less, than other countries; treatment time on average is shorter in the U.S. than in Japan and Europe, and U.S. patients miss more treatments. U.S. patients are less likely to receive food or nutritional supplements during their treatment, and staff in U.S. facilities tend to have less training (Foley and Hakim 2009). Under these circumstances, improving quality with measurement (if accurate) should be not only "low-hanging fruit" but a moral imperative.

Medicare therefore in 2012 implemented the End Stage Renal Disease Quality Incentive Plan (ESRD QIP) which reduced payments by up to 2% for providers "that do not achieve

or make progress toward specified quality measures" (MedPAC 2018b: 3). The ESRD QIP has been associated with improvements on measures such as use of fistulas rather than catheters, and proper targeting of anaemia, especially a reduction in overtreatment (Weiner and Watnick 2017; Saunders et al. 2017). Because it uses penalties rather than rewards, and improved dialysis is unlikely to increase costs for other services, it should have increased efficiency.

Yet even this strong case for pay-for-performance reveals difficulties. There is disagreement between CMS and its expert advisor, the National Quality Forum, about which measures to use. The set of measures has been "fluid, with frequent addition of new" ones (Weiner and Watnick 2017). Some of the improvements may have been based on promotion of strong research evidence (e.g. regarding fistulas) rather than incentives. Some of the data is questionable, especially the patient satisfaction scores (Brady et al. 1365). As in many other cases (e.g. Kahn et al. 2015), the sanctions are more likely to apply to providers with more disadvantaged patients, whose outcomes can be shaped by their social disadvantages or personal behaviour. Reducing providers' resources in response is not likely to improve performance (Saunders et al. 2017). The ESRD QIP seems like a necessary response to blatant quality failures, but we should hope that such examples are rare.

Conclusion

Both budget-makers and people who attempt to manage health care systems should realise that evaluating and assessing of performance is part of their job. Yet hopes that performance measurement initiatives will make health care systems more sustainable are not likely to be met.

Both budgeting and health care payment may now involve too much *measurement* and too little *observation*. The ethos of measurement is so strong that U.S. health system CEOs worried about burnout from excessive record-gathering recommended that physicians answer surveys about their degree of burnout (Noseworthy et al. 2017). Not everything can be measured; measurement limits attention to what can be identified in advance for counting, but the costs of measurement may arise in activities not identified beforehand. Observation means open-ended attention to what is happening on the shop floor, on the ward, or in the prison yard. Organisations need the equivalent of the prison warden who stands at the entrance to the lunchroom greeting the inmates and watching how they interact, and the informal communication that shares information in unstructured directions. Yet observation and communication can be threatened by documentation burdens (Michel 2017).

These lessons are clear in the literatures on both performance budgeting and paying-for-performance. The similarities should alert both policy communities that the general approach itself is not as promising as many would hope.

Much of what has been argued in this paper is implied in many statements about "best practices" for PB or P4P. The danger is that ideal circumstances will be confused with best practices. Analysts should remember that the reason "best practices" tend to be rare is that they can be very difficult to create.

Notes

1 For information on that activity see www.oecd.org/gov/budgeting/seniorbudgetofficialsnetworkonperformanceandresults.htm

2 As is shown by the forwards in the book, Chase's work is in many ways the core statement of what the founders of Harvard's John F. Kennedy School of Government wanted its students to learn.

3 For more discussion see White (2014), 75-81.

4 For one good example, focusing on consumers' interest in quality information, see Hibbard 2008.

5 See, for example, Sullivan 1992, and the discussion in White 2010. This rhetorical promotion of public health is not accompanied by proposals to spend more on it.

6 There are many descriptions of the QOF; one good source is Doran et al. 2006, including the supplemental appendix available online at www.nejm.org/doi/suppl/10.1056/NEJMsa055505/suppl_file/nejm_doran_375sa1.pdf

7 Note that success here is defined as improving quality, not efficiency.

8 This is a fundamental reason why budget analysis must focus on inputs as well as outputs.

9 The U.S. experience may be most thoroughly documented, as in Wheeler ed. 2011; the most horrifying story involves the M-16 used in the Vietnam War, but there are many others.

10 I do not mean to suggest that the "experts" are necessarily right but that is another topic.

11 Debates about guidelines are too extensive to cover here – but, as one example, readers might look at the conflicts in many countries over when women should receive mammograms, e.g. Biller-Adorno and Juni 2014.

12 As Robinson and MacPherson (2012) report, differences between Anthem's and Blue Shield's lists of quality hospitals resulted in part from insurers' and hospitals' bargaining strategies. I found even less overlap in 2014, and somewhat more than in 2014 in 2016. In an interview, one Anthem official explained that they did not "have real detailed data about the procedures" so made rough assumptions such as that high volume indicated higher quality, looked at infection rates "and some internal performance goals," generated a list and then asked physicians if it looked OK. This method likely did no harm, and certainly the higher-priced hospitals had little or no evidence to support their prices – but it does not fit the idea that patients were steered to "centers of excellence" based on measurement.

13 We should remember that performance measurement is not the only goal for EMRs, which are also promoted as improving co-ordination – though there is reason for concern about that also. Reports of the time involved may be a bit misleading because they do not adjust for time that would have been spent on paper records, but findings that half or more of time is spent feeding the record are common. See Arndt et al. 2017, Sinsky et al. 2016, Young et al. 2018.

14 In my own interviews I have heard numerous complaints about documentation burdens in other countries.

15 See also Berwick and Bisognano 2019 for more recent worries about excess measurement.

16 See Kahn et al. 2015. The examples were giving a cardiac patient aspirin on arrival at the hospital; prescribing aspirin at discharge; giving an ACE inhibitor soon after arrival; prescribing a beta blocker at discharge; assessment of heart function and, for pneumonia

patients, assessment of oxygen levels. Some may wonder why incentives were necessary, but nevertheless these were positive effects.

17 Personal communication with Dr. Duckett, 8 July 2019.

18 My account here is based on a series of interviews in the Netherlands in 2018. The same three conditions were made subject to chronic disease management programs in Denmark. My respondents in Denmark saw less variation in implementation across diseases, but more variation across locations and more doubt about how seriously the initiatives were implemented. They noted substantial up-front costs, so while there may have been modest quality improvements, it is not clear value for money increased. None asserted it had.

References

Abelson, J., S. Allin, M. Grignon, D. Pasic, and M.Walli-Attaei (2017), "Uncomfortable trade-offs: Canadian policy makers' perspectives on setting objectives for their health systems", *Health Policy* 121: 9-16.

Angell, B., J. Pares and G. Mooney, "Implementing priority setting frameworks: Insights from leading researchers", *Health Policy* 120: 1389-1394.

Arndt, B. G., J. W. Beasley, M. D. Watkinson, J. L. Tempte, W. J. Tuan, C. A. Sinsky and V. J. Gilchrist (2017), "Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations", Annals of Family Medicine 15(5):419-426.

Austin, J. M., A. K. Jha, P. S. Romano, S. J. Singer, T. J. Vogus, R. M. Wachter and P. J. Pronovost (2015), "National Hospital Rating Systems Share Few Common Scores And May Generate Confusion Instead of Clarity" *Health Affairs* 34(3): 423-430.

James, C., I. Beazley, C. Penn, L. Philips and S. Dougherty (forthcoming), "Decentralisation in the health sector and responsibilities across levels of government ", *OECD Journal on Budgeting - Special Issue on Health.*

Berenson, R. and H. Singh (2018), "Payment Innovations To Improve Diagnostic Accuracy and Reduce Diagnostic Error", *Health Affairs* 37(11): 1828-1835.

Berwick, D. M. (1995), "The Toxicity of Pay for Performance", *Quality Management in Health Care* 4(1): 27-33.

Berwick, D. M. and M. Bisognano (2019), "Keynote Three: I Have Changed My Mind." Video from International Forum on Quality and Safety in Healthcare, at https://www.youtube.com/watch?v=mPMJznD_Yis beginning at 3'13"

Bevan, G. and R. Hamblin (2009), "Hitting and missing targets by ambulance services for emergency calls: effects of different systems of performance measurement within the UK", *Journal of the Royal Statistical Society* 172(1): 161-190.

Bevan, G. and C. Hood (2006), "What's Measured is What Matters: Targets and Gaming in the English Public Health Care System", *Public Administration* 84(3): 517-538.

Biller-Andorno, N. and P. Juni (2014), "Abolishing mammography screening programs? A view from the Swiss Medical Board", *New England Journal of Medicine* 370(21): 1965-1967.

Brady, B. M., B. Zhao, J. Niu, W. C. Winkelmayer, A. Milstein, G. M. Chertow and K. F. Erickson (2018), "Patient-Reported Experiences of Dialysis Care Within a National Pay-for-Performance System", *JAMA Internal Medicine* 178(10): 1358-1367.

Busse, R., A. Geissler, W. Quentin and M. Wiley eds. (2011), *Diagnosis-Related Groups in Europe*, WHO and Open University Press.

Busse, R. (2016), "Pay-for-performance: Time to act but also to provide further evidence", *Health Policy* 120: 1123-1124.

Cassell, C. K., P. H. Conway, S.F. Delbanco, A. K. Jha, R. S. Saunders and T. H. Lee (2014), "Getting More Performance from Performance Measurement", *New England Journal of Medicine* 371(23): 2145 – 2147.

Chase, G. and E. C. Reveal (1983), *How to Manage in the Public Sector*, Boston: McGraw-Hill.

Chassin, Mark R. 2002. "Achieving and Sustaining Improved Quality: Lessons From New York State and Cardiac Surgery", *Health Affairs* 21(4): 40-51.

Doran, T., C. Fullwood, H. Gravelle, D. Reeves, E. Kontopantelis, U. Hiroeh and M. Roland (2006), "Pay-for-Performance Programs in Family Practices in the United Kingdom", *New England Journal of Medicine* 355(4): 375-384.

Duckett, S. J. (1995), "Hospital payment arrangements to encourage efficiency: the case of Victoria, Australia", *Health Policy* 34: 113-134.

Foley, R. N. and R. M. Hakim (2009), "Why Is the Mortality of Dialysis Patients in the United States Much Higher than the Rest of the World?, *Journal of the American Society of Nephrology* 20: 1432-1435.

Galvin, R. (2006), "Pay-For-Performance: Too Much Of A Good Thing? A Conversation With Martin Roland", *Health Affairs* 25: w412-w419.

Hibbard, J. (2008), "Editorial: What Can We Say About the Impact of Public Reporting? Inconsistent Execution Yields Variable Results", *Annals of Internal Medicine* 148(2): 160-161.

Hood, C. and R. Dixon (2015), *A Government that Worked Better and Cost Less? Evaluating Three Decades of Reform and Change in UK Central Government*, Oxford: Oxford University Press.

IBM Center for Business and Government (2011), "Is Performance Budgeting Hopeless?", Blog, July 13, www.businessofgovernment.org/blog/performance-budgeting-hopeless

Jha, A. K., A. R. Iliff, A. A. Chaoui, S. Defossez, M. C. Bombaugh and Y. R. Miller (2018), "A Crisis in Health Care: A Call to Action on Physician Burnout", Massachusetts Medical Society report, at www.massmed.org/news-and-publications/mms-news-releases/physician-burnout-report-2018/

Kahn, C. N. III, T. Ault, L. Potetz, T. Walke, J. Hart Chambers, and S. Burch (2015), "Assessing Medicare's Hospital Pay-for-Performance Programs And Whether They Are Achieving Their Goals", *Health Affairs* 34(8): 1281-1288.

Kamensky, J. (2014), "Performance Budgeting: Lessons from the States", IBM Center for Business and Government blog, 26 April, http://businessofgovernment.org/blog/performance-budgeting-lessons-states

Keller, A. (2018), "2018 OECD Annual Performance Budgeting Survey: Key Findings and Trends", Presented at the 14th Annual Meeting of the OECD SBO Performance and Results Network," Paris, 26-27 November, 2018, www.slideshare.net/OECD-GOV/international-trends-in-performance-budgeting-anne-keller-oecd

Laverty, A. A., P. C. Smith, U. J. Pape, A. Mears, R. M. Wachter and C. Millett (2012), "High-Profile Investigations Into Hospital Safety Problems in England Did Not Prompt Patients To Switch Providers", *Health Affairs* 31(3): 593-601.

Lawson, E. H., D. S. Zingmond, B. L. Hall, R. Louie, R. H. Brook, and C.Y. Ko (2015), "Comparison Between Clinical Registry and Medicare Claims Data on the Classification of Hospital Quality of Surgical Care", *Annals of Surgery* 261(2): 290-296.

Markowitz, A. A. and A. M. Ryan (2017), Pay for Performance: Disappointing Results or Masked Heterogeneity?, *Medical Care Research and Review* 74(1): 3-78.

Mathes, T., D. Pieper, J. Morche, S. Polus, T. Jaschinski, M. Eikermann (2019), "Pay for performance for hospitals", *Cochrane Systematic Review* (05 July), www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD011156.pub2/full

McDonald, R., J. White and T. R. Marmor (2009), "Paying for Performance in Primary Medical Care: Learning from 'Success' and 'Failure' in England and California", *Journal of Health Politics, Policy and Law* 34(5): 747-776.

McGlynn, E. A., E. C. Schneider and E. A. Kerr (2014), "Reimagining Quality Measurement", *New England Journal of Medicine* 371(23): 2150-2153.

MedPAC (Medicare Payment Advisory Commission) (2018a), *Report to the Congress: Medicare Payment Policy*, March, MedPAC.

MedPAC (Medicare Payment Advisory Commission) (2018b), "Outpatient Dialysis Services Payment System", *paymentbasics* October, www.medpac.gov/docs/default-source/payment-basics/medpac_payment_basics_16_dialysis_final.pdf

Mendelson A., K. Kondo, C. Damberg, A. Low, M. Motuapuaka, M. Freeman, M. O'Neil, R. Relevo and D. Kansagara (2017), "The Effects of Pay-for-Performance Programs on Health, Health Care Use, and Processes of Care: A Systematic Review", Annals of Internal Medicine 166(5): 341-353.

Michel, Lucie (2017), "A Failure to Communicate? Doctors and Nurses in American Hospitals", *Journal of Health Politics, Policy and Law* 42(4): 709-717,

Minchin, M., M. Roland, J. Richardson, S. Rowark and B. Guthrie (2018), "Quality of Care in the United Kingdom after Removal of Financial Incentives", *New England Journal of Medicine* 379(10): 948-957.

Mousquès, J. and F. Daniel (2015), "The Impact of Multiprofessional Group Practices on the Quality of General Practice", *Questions d'économie de la Santé* , IRDES, 211.

Noseworthy, J. et al. (2017), "Physician Burnout Is A Public Health Crisis: A Message To Our Fellow Health Care CEOs, Health Affairs blog (March 28), at www.healthaffairs.org/do/10.1377/hblog20170328.059397/full/

OECD, 2008. "Performance Budgeting: A User's Guide." *OECD Observer Policy Brief*, March, www.oecd.org/gov/budgeting/Performance-Budgeting-Guide.pdf.

OECD (2010), *Value for Money in Health Spending*, OECD Health Policy Studies, OECD Publishing, Paris, https://doi.org/10.1787/9789264088818-en.

OECD (2019a), *OECD Good Practices for Performance Budgeting*, OECD Publishing, Paris, https://doi.org/10.1787/c90b0305-en.

OECD (2019b), "Synthesis Note", Report on the 7th Meeting of the Joint Network of Senior Budget and Health Officials, 14-15 February, http://www.oecd.org/gov/budgeting/SBO-health-sythesis-note-2019.pdf

Ogundeji, Y.K., J.M. Bland, T.A. Sheldon (2016), "The Effectiveness of Payment for Performance in health care: a meta-analysis and exploration of variation in outcomes", *Health Policy* 120: 1141-1150.

Parast, L., B. Doyle, C. L. Damberg, K. Shetty, D. A. Ganz, N. S. Wenger and P. G. Shekelle (2015), "Perspective: Challenges in Assessing the Process-Outcome Link in Practice. *Journal of General Internal Medicine* 30(3): 359-364.

Rechel, B. (2019), "Funding for public health in Europe in decline?" *Health Policy* 123: 21-26.

Robinson, J. C. and K. MacPherson (2012), "Payers Test Reference Pricing and Centers of Excellence To Steer Patients to Low-Price and High-Quality Providers", *Health Affairs* 31(9): 2028-2036.

Roland, M. and S. Campbell (2014), "Successes and Failures of Pay for Performance in the United Kingdom", *New England Journal of Medicine* 370(20): 1944-1949.

Rosenbaum, L. (2015), "Scoring No Goal: Further Adventures in Transparency", *New England Journal of Medicine* 373(15): 1385-1388.

Saunders, M., R. Haenal Lee and M. H. Chin (2017), "Early winners and losers in dialysis center pay-for-performance", *BMC Health Services Research* 17: 816-824.

Schick, A. (2013), "The metamorphoses of performance budgeting", *OECD Journal on Budgeting* 2013(2): 1-31.

Shaw, T. (2016), "Performance budgeting practices and procedures", *OECD Journal on Budgeting*, vol. 15/3, https://doi.org/10.1787/budget-15-5jlz6rhqdvhh.

Sinsky, C. A., L. Colligan, L. Li, M. Prgomet, S. Reynolds, L. Goeders, J. Westbrook, M. Tutty and G. Blike (2016), "Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study of 4 Specialties." *Annals of Internal Medicine* 165(11): 753-760.

Srivastava, D., M. Mueller and E. Hewlett (2016), *Better Ways to Pay for Health Care*, OECD Health Policy Studies, OECD Publishing, Paris.

Stone, D. (2012), *Policy Paradox: The Art of Political Decision Making 3rd ed.*, New York: W.W. Norton.

Sullivan, K. and S. Soumerai (2018), "Pay for performance: a dangerous health policy fad that won't die", *STAT*, at www.statnews.com/2018/01/30/pay-for-performance-doctors-hospitals/

Sullivan, L. W. (1992), "The Bush Administration's Health Care Plan", *New England Journal of Medicine* 327(11): 801-804.

Waldo, S.W., J. M. McCabe, C. O'Brien, K. F. Kennedy, K. E. Joynt and R.W. Yeh (2015), "Association Between Public Reporting of Outcomes With Procedural Management and Mortality for Patients With Acute Myocardial Infarction", *Journal of the American College of Cardiology* 65(11): 1119-1126.

Weiner, D. and S. Watnick (2017), "The ESRD Quality Incentive Program: Can We Bridge the Chasm?", *Journal of the American Society of Nephrology* 28(6): 1697-1706.

Wheeler, W. T. (2011), *The Pentagon Labyrinth: 10 Short Essays to Help You Through It*, Washington, DC: Center for Defense Information, http://pogoarchives.org/labyrinth/full-labyrinth-text-w-covers.pdf

White, J. (1994), "Almost Nothing New Under the Sun: Why the Work of Budgeting Remains Incremental. Public Budgeting & Finance 14(1): 113-134.

White, J. (2010), "My Health Policy Nightmare", *Health Matrix* 20: 423-436.

White, J. (2014), "The challenge of budgeting for healthcare programmes", *OECD Journal on Budgeting*, vol. 14/1, https://doi.org/10.1787/budget-14-5jxst2mfm923.

White, J. (2018), "Hypotheses and Hope: Policy Analysis and Cost Controls (or Not) in the Affordable Care Act", *Journal of Health Politics, Policy and Law* 43(3): 455-482.

Woolhandler, S., D. Ariely and D. Himmelstein (2012), "Will Pay for Performance Backfire? Insights From Behavioral Economics", Health Affairs blog, 11 October, www.healthaffairs.org/do/10.1377/hblog20121011.023909/full/

Young, R. A., S. K. Burge, K. A. Kumar, J. M. Wilson and D. F. Ortiz (2018), "A Time-Motion Study of Primary Care Physicians' Work in the Electronic Health Record Era", *Family Medicine* 50(2): 91-99