

EDITOR'S CHOICE

History Bias, Study Design, and the Unfulfilled Promise of Pay-for-Performance Policies in Health Care

Huseyin Naci, PhD, MHS; Stephen B. Soumerai, ScD

Suggested citation for this article: Naci H, Soumerai SB. History Bias, Study Design, and the Unfulfilled Promise of Pay-for-Performance Policies in Health Care. *Prev Chronic Dis* 2016; 13:160133. DOI: <http://dx.doi.org/10.5888/pcd13.160133>.

Editor's Note: In June 2015, *Preventing Chronic Disease* published "How Do You Know Which Health Care Effectiveness Research You Can Trust? A Guide to Study Design for the Perplexed," which used simple graphs and easy-to-understand text — in 5 case studies — to illustrate how powerful biases, combined with weak study designs that cannot control for those biases, yielded untrustworthy findings on influenza vaccination policy, health information technology, drug safety, prevention of childhood obesity, and hospital safety ("mortality reduction") programs. The target audiences for that article were policy makers, journalists, public health and medical trainees, and the general public; the primary goal was to explain how weak or strong study designs fail or succeed in controlling for biases. In the Editor's Note, we promised to add to those examples of common biases and research designs to show why people should be cautious about accepting research results — results that may have profound and long-lasting effects on health policy or clinical practice, some of which could be detrimental to health.

In this sixth case study, we revisit one of the most common and dangerous threats to research validity: history bias (ie, researchers' failure to consider relevant events or changes that precede an intervention or co-occur while it is in progress). Studies that fail to control for history can mislead policy makers and clinicians. The pay-for-performance policy used to illustrate history bias in this article is sensitive to this powerful bias, because medical practice is always changing as a result of factors unrelated to policy. Without investigating changes in a study's hoped-for outcome over time both before and after the policy or intervention being studied is implemented, investigators will probably attribute those changes to effects of the policy they are studying, causing billions of dollars of waste implementing such policies worldwide.

Introduction

The ongoing flip-flopping of research findings about the effects of medical or health policies weakens the credibility of health science among the general public, clinicians, members of Congress, and the National Institutes of Health (1–3). Even worse, poorly designed studies, combined with widespread reporting on those studies by the news media, can distort the decisions of policy makers, leading them to fund ineffective, costly, or even harmful policies. Several reports in top medical journals in 2015 (4–6) pronounced that economic incentives in Pioneer Accountable Care Organizations saved medical costs, but the reports did not control for major biases created by unfairly comparing selected high-performing organizations with less-experienced control organizations (7). The result? The US Centers for Medicare & Medicaid Services cited the findings as a reason for expanding the program nationwide.

Building on an earlier article in *Preventing Chronic Disease* (8), this article focuses on a widely accepted but questionably effective (9) health policy that compensates physicians for meeting certain quality-of-care standards, such as measuring or treating high blood pressure. Policy makers often believe that such financial incentives motivate physicians to improve their performance to maintain or increase their incomes, thereby improving patient outcomes (10). Health care systems in the United States, Canada, Germany, Israel, New Zealand, Taiwan, and the United Kingdom have committed billions of dollars to this approach in the hope that such incentives will improve the quality of health care (11). Although this monetary approach sounds good theoretically, international scientific reviews overwhelmingly find little evidence to support it (12). Giving physicians small incremental payments to do things they already do routinely (eg, measuring blood pressure) may be counterproductive and even insulting, may divert their attention from more critical concerns, and does not increase quality of care (13). Some studies even find that such compensation encourages unethical behavior by incentivizing doctors to "cherry-pick" healthy, active, wealthy patients over "costly" sick patients who are less likely to reach the performance targets. Nevertheless,



The opinions expressed by authors contributing to this journal do not necessarily reflect the opinions of the U.S. Department of Health and Human Services, the Public Health Service, the Centers for Disease Control and Prevention, or the authors' affiliated institutions.

this financial-incentive policy is entrenched in many components of the Patient Protection and Affordable Care Act (colloquially known as Obamacare), including Accountable Care Organizations, patient-centered medical homes, and health information technology (14).

In this article, our aim is to help the public and policy makers understand how a pervasive bias can undermine the results of poorly designed studies of pay-for performance programs published in even the world's leading medical journals. We also point to observational study designs and systematic reviews of the total body of evidence to find more trustworthy conclusions on the efficacy of pay-for-performance (12). Although randomization is frequently not feasible for evaluating such public policies (15), we also present an example of a randomized controlled trial that supports the conclusions drawn from strong observational study designs.

The Threat: History Bias

The most pervasive threat to the credibility of studies of pay-for-performance (and many other health interventions) is history bias. History biases are simple to understand: they are events unrelated to the policy under study that occur before or during the implementation of that policy and that may have a greater effect on the policy's hoped-for outcome than the policy itself. These events call into question the conclusions of studies evaluating the policy. They can be any event, such as a concurrent improvement in physician practice that successfully identifies and treats patients with high blood pressure, or widespread news media coverage of a new drug or new national guidelines supporting a life-saving treatment (for example, β blockers to prevent acute myocardial infarction [16]). The American Heart Association's physician educational campaign Get With the Guidelines led to a gradual improvement in hypertension management (17). If a study of a pay-for-performance program targeting hypertension management is launched after or in the middle of such a campaign and does not account for that campaign's effect on any improvements, the pay-for-performance program may take credit for the success, even though the blood pressure improvements really resulted from the unmeasured historical changes in physician practices (18).

Weak pre-post designs that did not control for history bias

Many studies have evaluated the United Kingdom's pay-for-performance program — a national policy that provides financial incentives to physicians to improve patient care. The largest pay-for-performance program of its kind, this national policy, introduced in April 2004, offered family physicians up to an additional 25% of salary for meeting certain performance standards.

Figure 1 shows data from a study that did not protect against history bias when it evaluated the United Kingdom's pay-for-performance program (19,20). The objective of this study was to determine whether the program led to improvements in a set of quality indicators — measurable elements of practice that indicate quality of care (in our example, target total cholesterol levels). This study had data only for the same month in which the policy was implemented and 2 points afterwards.

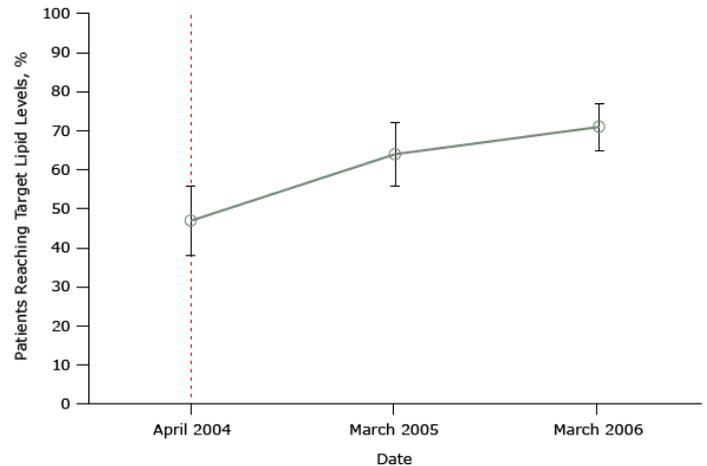


Figure 1. Mean percentage of patients achieving a selected quality indicator — a target total cholesterol level of ≤ 5 mmol/L — in a sample of family practices that participated in a study evaluating the effect of the United Kingdom's pay-for-performance policy. Dashed line indicates when the pay-for-performance policy was implemented (April 2004). Figure is based on data extracted from Table 2 of Tahrani AA, McCarthy M, Godson J, Taylor S, Slater H, Capps N, et al. Diabetes care and the new GMS contract: the evidence for a whole county. *Br J Gen Pract* 2007;57(539):483–5 (19).

Figure 2 shows data from another study that evaluated the United Kingdom's pay-for-performance policy. This study also did not account for secular trends, yet it was published in a major medical journal and is highly cited (21). The authors' assessment included only 2 points in time before and only 2 points after program implementation.

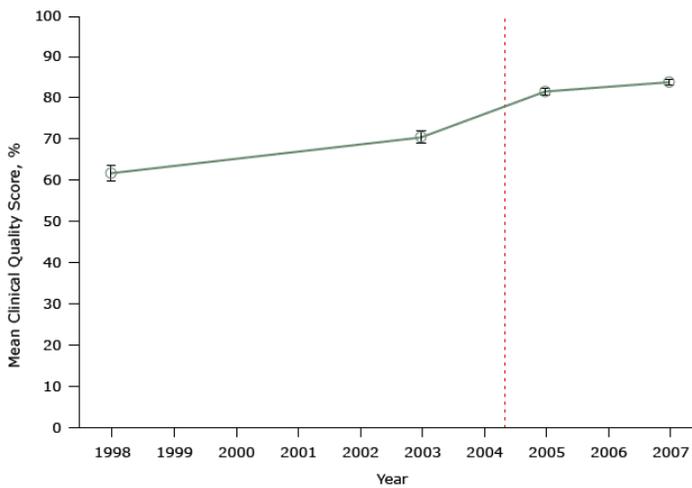


Figure 2. Mean clinical quality scores for diabetes at 42 practices participating in a study evaluating the effect of the United Kingdom’s pay-for-performance policy. The scale for scores ranges from 0% (no quality indicator was met for any patient) to 100% (all quality indicators were met for all patients). Dashed line indicates when the pay-for-performance policy was implemented (April 2004). Figure is based on data extracted from Table 1 in Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of pay for performance on the quality of primary care in England. *N Engl J Med* 2009;361(4):368–78 (21).

The key problem with both of these studies, which purport to show the positive effects of the national pay-for-performance policy, is the use of only 2 data points during a long period before program implementation and 2 data points afterwards. From so few data, it is impossible to know what was happening to affect diabetes scores unrelated to the pay-for-performance policy before that policy was implemented. So we do not know if any small changes after policy implementation resulted from the pay-for-performance program or from some other changes in physicians’ practice. If anything, it appears that improvements before implementation (from 1998 to 2003) — to the extent these are detectable by examining only 2 data points — may have lessened or flattened, not increased after implementation of pay-for-performance. These studies are examples of a simple pre–post design. Only one or 2 observations (points) before and after pay-for-performance cannot control for secular trends (history) before program implementation. The pre-existing trajectory of good quality of care both before and after the intervention is unknown, and it is impossible to know whether the policy had any effect on this trend.

Strong interrupted time-series design that controls for history bias

Figure 3 illustrates a result of one of the most convincingly negative studies showing that the United Kingdom’s pay-for-performance had no detectable effects on quality of care for patients with hypertension. Using a strong interrupted time-series design and 7 years of monthly data (84 time points) for 400,000 patients before and after the program’s implementation, Serumaga et al showed that the pay-for-performance program started in the middle of a slight rise in the percentage of patients who began blood pressure treatment (22).

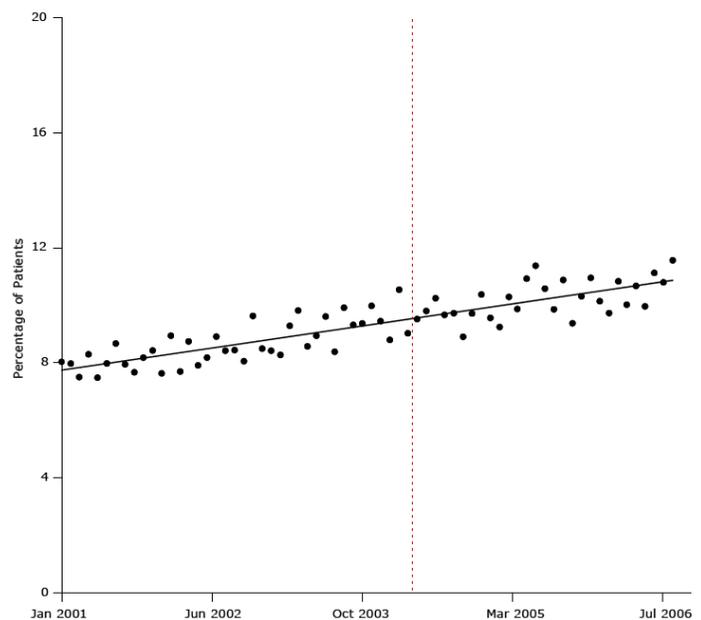


Figure 3. Percentage of study patients who began antihypertensive drug treatment from January 2001 through July 2006. Dashed line indicates when the United Kingdom’s pay-for-performance policy was implemented (April 2004). Figure is based on data extracted from bottom panel, Figure 3, in Serumaga B, Ross-Degnan D, Avery AJ, Elliott RA, Majumdar SR, Zhang F, et al. Effect of pay for performance on the management and outcomes of hypertension in the United Kingdom: interrupted time series study. *BMJ* 2011;342:d108 (22).

Every figure in the article shows flat or slightly improving treatment over many years and no effect of the \$2 billion program that links family physician’s income to measures of health care quality. The existence of a long prepolicy trend, established by data for January 2001 through April 2004, to control for history bias (eg, pre-existing physician improvements in quality) enabled a valid assessment of the effect of the policy on changes in the level or

The opinions expressed by authors contributing to this journal do not necessarily reflect the opinions of the U.S. Department of Health and Human Services, the Public Health Service, the Centers for Disease Control and Prevention, or the authors’ affiliated institutions.

trend in health outcomes over many years. The stronger study (interrupted time-series design) showed that the United Kingdom's pay-for-performance program had no effects, whereas the 2 weak studies (pre-post design) contributed only to false or exaggerated hopes.

Figure 4 shows a remarkably similar negative result of paying hospitals for their measured performance. Hospital pay-for-performance programs and physician pay-for-performance programs are developed under similar assumptions: linking pay-for-performance with certain hospital outcomes are expected to motivate hospital leaders to meet targets to maintain or increase their incomes, thereby improving patient outcomes.

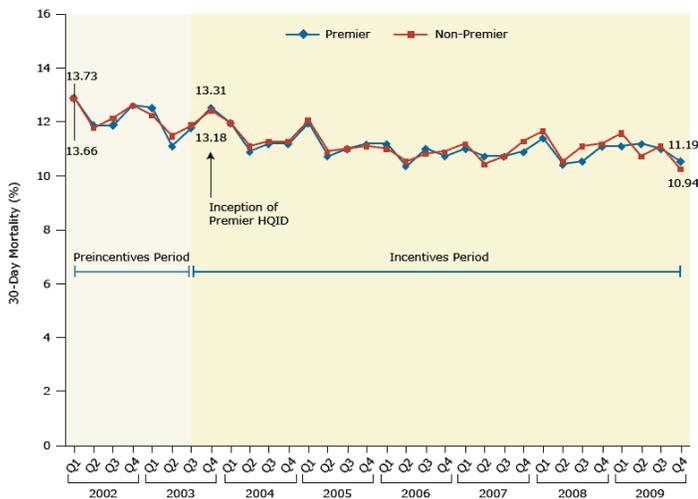


Figure 4. Mortality at 30 days among all hospitals examined before (from first quarter 2002) and after (through fourth quarter 2009) implementation of a pay-for-performance intervention (Premier Hospital Quality Incentives Demonstration [HQID]), which targeted 4 conditions beginning in late 2003: acute myocardial infarction, congestive heart failure, and pneumonia, and patients who underwent coronary artery bypass grafting. Changes at hospitals participating in the pay-for-performance intervention (Premier) were similar to changes at hospitals not participating (non-Premier) for all 4 conditions. Figure is reproduced from Jha AK, Joynt KE, Orav EJ, Epstein AM. The long-term effect of premier pay for performance on patient outcomes. *N Engl J Med* 2012;366(17):1606–15 with permission from the *New England Journal of Medicine* (23).

A study by Jha et al (23) used an interrupted-time series design with a comparison series to investigate differences in patient mor-

tality rates between hospitals participating in a pay-for-performance program and hospitals not participating. The study compared data for patients with one of 4 conditions: acute myocardial infarction, congestive heart failure, and pneumonia, and patients who underwent coronary artery bypass grafting. The 7-year trends in 30-day mortality for the pay-for-performance and non-pay-for-performance hospitals almost completely overlapped, leaving little doubt that the program had no detectable effect on long-term mortality (Figure 4).

It is worth noting, however, that hospitals in the pay-for-performance program opted into that program and could have already had better outcomes than non-participating hospitals before the pay-for-performance program began (ie, they were anticipating the financial rewards). But the equivalent trends and pre-program improvements in mortality for both study groups makes it less likely that this bias would have changed the conclusion.

Interestingly, a more recent study of pay-for-performance effects on 30-day in-hospital mortality rates among patients with pneumonia, heart failure, or acute myocardial infarction in one region of the United Kingdom was also compromised by already occurring declines in mortality rates and a lack of clear differences in mortality rates between the study and comparison groups (24). These short-term declines in mortality rates were not maintained in the long term (25).

Strongest designs for controlling for history bias: randomized controlled trials

The strongest design for evaluating policies is a randomized controlled trial (RCT). In such study designs, random allocation of participants into intervention and control groups increases the likelihood that the only difference between the group receiving the pay-for-performance intervention and the control group (the one not participating in pay-for-performance) is the intervention itself. In a recent RCT, physicians randomized to a pay-for-performance intervention were eligible to receive up to \$1,024 per patient who met target cholesterol levels, whereas physicians in the control groups received no economic incentives for achieving better outcomes (26).

Studies with strong, trustworthy designs, such as this RCT, suggest that paying physicians according to their measured performance on quality metrics (eg, reduction in low-density lipoprotein levels) does not improve outcomes (Figure 5). Physician payments did not produce any meaningful changes in quality of care compared with an equivalent group receiving no incentives.

The opinions expressed by authors contributing to this journal do not necessarily reflect the opinions of the U.S. Department of Health and Human Services, the Public Health Service, the Centers for Disease Control and Prevention, or the authors' affiliated institutions.

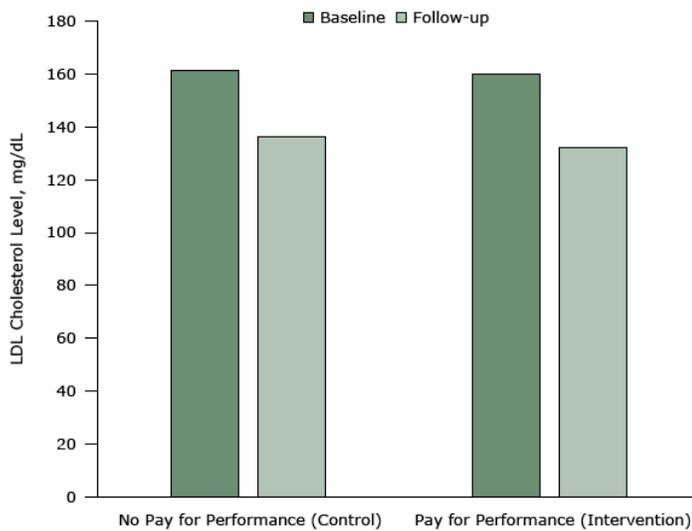


Figure 5. Mean low-density lipoprotein (LDL) cholesterol levels at baseline and 12-month follow-up in an intervention (pay-for-performance) group (in which incentives were provided to physicians) and a control group (no pay-for-performance). The intervention was conducted from 2011 to 2014 in 3 primary care practices in the northeastern United States. Patients in the control group achieved a mean reduction of 25.1 mg/dL in LDL cholesterol levels from a baseline of 161.5 mg/dL. Patients in the pay-for-performance group achieved a mean reduction of 27.9 mg/dL from a baseline of 159.9 mg/dL. The difference between the 2 groups was neither statistically significant nor clinically meaningful. Figure is based on data extracted from Asch DA, Troxel AB, Stewart WF, Sequist TD, Jones JB, Hirsch AG, et al. Effect of financial incentives to physicians, patients, or both on lipid levels: a randomized clinical trial. *JAMA* 2015;314(18):1926–35 (26).

Rigorous systematic reviews of the entire body of pay-for-performance studies: the most trustworthy evidence

It is important to reiterate that no study is perfect, and no single study can determine the truth, whatever that may be. The accumulation of knowledge over time is the best way to assess a health care treatment or policy. However, given that most studies do not control for history or other biases, it is essential to single out the most rigorous systematic reviews — literature syntheses that eliminate weakly designed studies (the simple pre–post study designs illustrated in Figure 1 and Figure 2 or studies that simply correlate pay-for-performance with quality of care at one point in time) and summarize the remaining evidence.

One international systematic review (Box) (12) found that not only was there little evidence to support pay-for-performance’s effects on quality of medical care, some studies found that it sometimes had the unintended consequence of discouraging doctors from treating the sickest patients.

Box. Conclusion of International Study by Houle et al in *Annals of Internal Medicine*, 2012: Does Performance-Based Remuneration for Individual Health Care Practitioners Affect Patient Care?: A Systematic Review (12)

Although uncontrolled before–after studies suggested that P4P [pay for performance] improves adherence to quality-of-care indicators for chronic illnesses . . . higher-quality studies with contemporaneous control groups or analyses that considered secular trends failed to confirm these benefits. Most important, 4 large interrupted time series analyses conducted in the United Kingdom to evaluate the effect of their primary care P4P scheme introduced in 2004 found that quality scores for incentivized indicators were increasing for patients . . . before P4P began; there was no convincing evidence that the quality of care increased at a faster rate in the 3 years after P4P implementation than before. [T]he current evidence for P4P targeting individual practitioners is insufficient to recommend wholesale adoption in health care systems at this time.

In addition to the international systematic review, other recent well-conducted systematic reviews supported the conclusion that questions the efficacy of pay-for-performance and advised against its widespread implementation — which has occurred despite the negative evidence (27). For example, Dutch researchers conducted an “umbrella review” — a review of all systematic reviews on pay-for-performance policies — to consider the totality of the evidence (28). They found that most systematic reviews unequivocally concluded that evidence showing effectiveness for pay-for-performance policies is weak, mixed, and inconclusive; many studies failed to find a meaningful effect attributable to the policy. As we have illustrated in this article, studies with weak designs that do not control for biases found more positive results than those with strong designs (29).

Closing Comments

Despite its unfulfilled promise and discouraging evidence, this costly and ineffective approach to improving health care is a widespread component of current national and international health care policies. It is entrenched in many policies created by the Affordable Care Act (14). Part of the problem is the explosion in statistical techniques that attempt to “adjust for” or “correct” unquestionably dissimilar study and comparison groups rather than graphing the actual data over time so policy makers (who appreci-

The opinions expressed by authors contributing to this journal do not necessarily reflect the opinions of the U.S. Department of Health and Human Services, the Public Health Service, the Centers for Disease Control and Prevention, or the authors’ affiliated institutions.

ate simple graphical displays) can actually look at the size of effects. Exaggeration of the effects of government programs through such “black box” statistics by the news media further widens the divide between reality and perception of policy effects (30). Weakly designed studies may also facilitate the proliferation of policies that encourage physicians to achieve “target rates” for health care procedures even when these procedures may be harmful for some patients. Oftentimes, what’s measured is what matters, and quality may deteriorate in areas that are not incentivized (31). Pay-for-performance programs may even result in collateral damage (13), diverting resources from under-resourced facilities such as safety net hospitals that provide care for vulnerable and high-need populations (32).

If we wish to encourage efficiency in medicine when our government and private health care programs are consuming almost one-fifth (17%) of the gross national product, it may be time to insist on strong experimental and quasi-experimental research designs (such as RCTs, interrupted time-series designs, and systematic reviews) in pilot tests of expensive policies. Investments of private and taxpayer funds should be based on solid evidence of safety and efficacy. The alternative, the present system, relies on weak and uncontrolled research designs, misleads policy makers and the public, and will ultimately lead to perverse effects, such as unsustainable costs, unhappy clinicians, and policies that may damage rather than improve the quality of medical care (30).

Notes

Dr Soumerai is co-chair of the Evaluative Sciences and Statistics Concentration of Harvard University’s PhD Program in Health Policy, Boston, Massachusetts. This project was supported by a Thomas O. Pyle Fellowship and a Developmental Research Design grant (Dr Soumerai) from the Department of Population Medicine, Harvard Medical School, and Harvard Pilgrim Health Care Institute, Boston; and a grant from the Commonwealth Fund (no. 20120504). Dr Soumerai received grant support (no. 5U58DP002719) from the Centers for Disease Control and Prevention’s Natural Experiments for Translation in Diabetes (NEXT-D). Dr Naci received no financial support for developing this article. We are indebted to Dr Sumit Majumdar for his outstanding contributions to an earlier version of the article. We are grateful to Caitlin Lupton for editorial assistance and graphic design. The Commonwealth Fund is a national, private foundation in New York City that supports independent research on health care issues and makes grants to improve health care practice and policy. The views presented here are those of the authors and not necessarily those of The Commonwealth Fund, its directors, officers, or staff.

Author Information

Corresponding Author: Stephen B. Soumerai, ScD, Professor of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Landmark Center, 401 Park Dr, Boston, MA 02215. Telephone: 617-509-9942. Email: ssoumerai@hms.harvard.edu.

Author Affiliation: Huseyin Naci, Department of Social Policy, London School of Economics and Political Science, London, United Kingdom.

References

1. Majumdar SR, Soumerai SB. The unhealthy state of health policy research. *Health Aff (Millwood)* 2009;28(5):w900–8.
2. Prasad V, Vandross A, Toomey C, Cheung M, Rho J, Quinn S, et al. A decade of reversal: an analysis of 146 contradicted medical practices. *Mayo Clin Proc* 2013;88(8):790–8.
3. National Institutes of Health. Enhancing reproducibility through rigor and transparency; 2015. <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-103.html>. Accessed January 20, 2016.
4. Soumerai SB, Koppel R. Accountable care organizations: did they reduce medical costs in one year? *The Incidental Economist*; 2015 May 5 [blog]. <http://theincidentaleconomist.com/wordpress/accountable-care-organizations-did-they-reduce-medical-costs-in-one-year/>. Accessed April 14, 2016.
5. Nyweide DJ, Lee W, Cuerdon TT, Pham HH, Cox M, Rajkumar R, et al. Association of Pioneer Accountable Care Organizations vs traditional Medicare fee for service with spending, utilization, and patient experience. *JAMA* 2015; 313(21):2152–61.
6. McWilliams JM, Chernew ME, Landon BE, Schwartz AL. Performance differences in year 1 of pioneer accountable care organizations. *N Engl J Med* 2015;372(20):1927–36.
7. Koppel R, Soumerai SB. Not-so-useful findings. Flawed research on health care costs begets flawed health care policy. *US News World Report* 2015 May 18; News, opinion & analysis: policy dose [blog]. <http://www.usnews.com/opinion/blogs/policy-dose/2015/05/18/when-health-care-cost-studies-get-the-facts-wrong>.
8. Soumerai SB, Starr D, Majumdar SR. How do you know which health care effectiveness research you can trust? A guide to study design for the perplexed. *Prev Chronic Dis* 2015; 12:150187.

9. Glasziou PP, Buchan H, Del Mar C, Doust J, Harris M, Knight R, et al. When financial incentives do more good than harm: a checklist. *BMJ* 2012;345:e5047.
10. Roland M, Dudley RA. How financial and reputational incentives can be used to improve medical care. *Health Serv Res* 2015;50(Suppl 2):2090–115.
11. Eijkenaar F. Pay for performance in health care: an international overview of initiatives. *Med Care Res Rev* 2012;69(3):251–76.
12. Houle SK, McAlister FA, Jackevicius CA, Chuck AW, Tsuyuki RT. Does performance-based remuneration for individual health care practitioners affect patient care?: a systematic review. *Ann Intern Med* 2012;157(12):889–99.
13. Woolhandler S, Himmelstein DU. Collateral damage: pay-for-performance initiatives and safety-net hospitals. *Ann Intern Med* 2015;163(6):473–4.
14. Burwell SM. Setting value-based payment goals — HHS efforts to improve U.S. health care. *N Engl J Med* 2015;372(10):897–9.
15. Chaix-Couturier C, Durand-Zaleski I, Jolly D, Durieux P. Effects of financial incentives on medical practice: results from a systematic review of the literature and methodological issues. *Int J Qual Health Care* 2000;12(2):133–42.
16. Lee TH. Eulogy for a quality measure. *N Engl J Med* 2007;357(12):1175–7.
17. LaBresh KA, Ellrodt AG, Gliklich R, Liljestrand J, Peto R. Get with the guidelines for cardiovascular secondary prevention: pilot results. *Arch Intern Med* 2004;164(2):203–9.
18. Shadish WR, Cook TD, Cambell DT. *Experimental and quasi-experimental designs for generalized casual inference*. Belmont (CA): Wadsworth Cengage Learning; 2002.
19. Tahrani AA, McCarthy M, Godson J, Taylor S, Slater H, Capps N, et al. Diabetes care and the new GMS contract: the evidence for a whole county. *Br J Gen Pract* 2007;57(539):483–5.
20. Tahrani AA, McCarthy M, Godson J, Taylor S, Slater H, Capps N, et al. Impact of practice size on delivery of diabetes care before and after the Quality and Outcomes Framework implementation. *Br J Gen Pract* 2008;58(553):576–9.
21. Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of pay for performance on the quality of primary care in England. *N Engl J Med* 2009;361(4):368–78.
22. Serumaga B, Ross-Degnan D, Avery AJ, Elliott RA, Majumdar SR, Zhang F, et al. Effect of pay for performance on the management and outcomes of hypertension in the United Kingdom: interrupted time series study. *BMJ* 2011;342:d108.
23. Jha AK, Joynt KE, Orav EJ, Epstein AM. The long-term effect of premier pay for performance on patient outcomes. *N Engl J Med* 2012;366(17):1606–15.
24. Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M. Reduced mortality with hospital pay for performance in England. *N Engl J Med* 2012;367(19):1821–8.
25. Kristensen SR, Meacock R, Turner AJ, Boaden R, McDonald R, Roland M, et al. Long-term effect of hospital pay for performance on mortality in England. *N Engl J Med* 2014;371(6):540–8.
26. Asch DA, Troxel AB, Stewart WF, Sequist TD, Jones JB, Hirsch AG, et al. Effect of financial incentives to physicians, patients, or both on lipid levels: a randomized clinical trial. *JAMA* 2015;314(18):1926–35.
27. Scott A, Sivey P, Ait Ouakrim D, Willenberg L, Naccarella L, Furler J, et al. The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database Syst Rev* 2011;(9):CD008451.
28. Eijkenaar F, Emmert M, Scheppach M, Schöffski O. Effects of pay for performance in health care: a systematic review of systematic reviews. *Health Policy* 2013;110(2-3):115–30.
29. Schatz M. Does pay-for-performance influence the quality of care? *Curr Opin Allergy Clin Immunol* 2008;8(3):213–21.
30. Soumerai SB, Starr D, Majumdar SR. How do you know which health care effectiveness research you can trust? a guide to study design for the perplexed. *Prev Chronic Dis* 2015;12:E101.
31. Bevan G. Have targets done more harm than good in the English NHS? No. *BMJ* 2009;338:a3129.
32. Gilman M, Hockenberry JM, Adams EK, Milstein AS, Wilson IB, Becker ER. The financial effect of value-based purchasing and the hospital readmissions reduction program on safety-net hospitals in 2014: a cohort study. *Ann Intern Med* 2015;163(6):427–36.