The Unfulfilled Promise of Pay-for-Performance (PfP) in Health Care

Menzies Centre for Health Policy and Centre for Research Excellence in Medicines and Ageing
University of Sydney
Friday 7th November 2017

Professor Stephen Soumerai
Harvard Medical School and Harvard Pilgrim Health Care Institute
Department of Population Medicine
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
Paying doctors bonuses for better health outcomes makes sense in theory. But it doesn't work.

It can even lead doctors to shun treating the sickest patients.

Updated by Stephen B. Soumerai and Ross Koppel | Jan 25, 2017, 9:40am EST
Economic Incentives: Theory vs. Evidence

- Pay-For Performance: extra $ to MDs for meeting quality standards
  - Typical standard: measuring or treating high BP

- Early research on PfP had serious design flaws
  - Led to ineffective, costly, or even harmful policies

- Later research w/ stronger designs overturned the optimistic findings
  - Yet PfP is entrenched in Obamacare, Medicare, and seven countries’ health policies
Why does PfP fail?

- Overreliance on economic theory in the absence of empirical research

- Improved performance pre-dates PfP

- Social factors are not counted (e.g., poverty, language, spousal abuse)

- “The idea that everyone’s professionalism and everyone’s good will has to be bought with tips is bizarre.” (Uwe Reinhardt, NYT, 2015)
The Truth Wears Off

Is there something wrong with the scientific method?

by Jonah Lehrer

December 13, 2010

Many results that are rigorously proved and accepted start shrinking in later studies.
A Few Examples of Clinical/Health Policies that Contradict Rigorous Research

- Pioneer Accountable Care Organizations (US)
- US roll-out of health IT to affect safety, mortality, costs
- Charging the sick more for health insurance (“Wellness Penalties”)
- Mailed Rxing feedback (O’Connell et al. UNSW)
“You can't fix by statistical adjustment what you bungled by design.”

Hierarchy of Strong and Weak Designs: Capacity to Control for Biases

Strong Design: Often Trustworthy Effects

Intermediate Design: Sometimes Trustworthy Effects

Weak Designs: Rarely Trustworthy Effects (No Controls for Common Biases.)
## Hierarchy of Strong and Weak Designs: Capacity to Control for Biases

<table>
<thead>
<tr>
<th>Strong Design: Often Trustworthy Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple RCTs</strong></td>
</tr>
<tr>
<td>The “gold standard” of evidence, incorporating systematic review of all studies.</td>
</tr>
<tr>
<td><strong>Single RCT</strong></td>
</tr>
<tr>
<td>A single, strong randomized experiment, but sometimes not generalizable</td>
</tr>
<tr>
<td><strong>Interrupted time series with control series (CITS)</strong></td>
</tr>
<tr>
<td>Baseline trends often allow visible effects and control for biases. Two controls.</td>
</tr>
</tbody>
</table>
## Hierarchy of Strong and Weak Designs: Capacity to Control for Biases

<table>
<thead>
<tr>
<th>Intermediate design: Sometimes Trustworthy Effects</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single ITS</strong></td>
<td>Controls for trends, but no comparison.</td>
</tr>
<tr>
<td><strong>Before and after with comparison group</strong></td>
<td>Pre-post change using two single observations. Comparability of baseline unclear.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weak Designs: Rarely Trustworthy Effects (No Controls)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Uncontrolled pre-post</strong></td>
<td>Single observations before and after intervention, no baseline or control group.</td>
</tr>
<tr>
<td><strong>Cross-sectional designs</strong></td>
<td>Simple correlation, no baseline, no measure of change.</td>
</tr>
</tbody>
</table>
Different Effects That Can Be Observed in Time Series

- Intervention before and after
- Different patterns of change before and after intervention
A Pervasive Threat: History Bias

Events unrelated to the policy under study that occur before or during policy and may be responsible for the outcome.

– E.g., A national flu campaign
Weak pre-post designs that did not control for history bias in UK P-f-P Program

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. Tahrani et al., Br J Gen Pract, 2007.
Another Poor Research Design: UK PfP

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% to 100%. Campbell et al., N Engl J Med 2009.
Figure 3. Percentage of study patients who began antihypertensive drug treatment from January 2001 through July 2006. Serumaga et al., BMJ 2011.
No Effect of PfP Cervical Screen among Australian GPs

Source: Green J. Health Serv Res. 2013 Aug;48(4):1415-32
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. Jha et al., N Engl J Med 2012.
Figure 5. Mean low-density lipoprotein (LDL) cholesterol levels at baseline and 12-month follow-up in a pay-for-performance group and a control group (no pay-for-performance) 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. Asch et al., JAMA 2015.
“Although uncontrolled before-after studies suggested that P4P improves adherence to quality-of-care indicators for chronic illnesses...higher-quality studies with contemporaneous control groups or that considered secular trends failed to confirm these benefits.”

Umbrella Review: Review of all Systematic Reviews

“They found that most systematic reviews unequivocally concluded that evidence showing effectiveness for pay-for-performance policies is weak, mixed and inconclusive.”

2017 Cochrane Systematic Review of PfP in Outpatient Care Facilities

“Little or no difference in”:
- “Compliance with quality assurance”
- “Use of health services”
- “Patients’ health status”

WE'VE DECIDED TO USE THE NEW **PAY** INCENTIVES ON THE PROJECTS WE WERE GOING TO DO ANYWAY.
Closing Comments

“Despite its unfulfilled promise and discouraging evidence, this approach is in widespread use internationally. It is entrenched in the Affordable Care Act.”

“Part of the problem is the explosion in statistical techniques that attempt to “adjust for” or “correct” unquestionably dissimilar study and comparison groups.”
Acknowledgements

- Thomas O. Pyle Fellowship (Dr. Soumerai), Department of Population Medicine, Harvard Medical School, and Harvard Pilgrim Health Care Institute, Boston

- Dr. Soumerai received grant support from the Commonwealth Fund.

- Dr. Soumerai received grant support from CDC’s Natural Experiments for Translation in Diabetes (NEXT-D).

- Dr. Naci, Department of Social Policy London School of Economics and Political Science received no financial support for developing this article.

- Dr. Majumdar was supported as a Health Scholar (Alberta Heritage Foundation for Medical Research).

- We are grateful to Ellen Taratus for outstanding editing of this article.

- Thanks to Caitlin Lupton, Jeanne Madden, and Wendy Drobnyk for editorial support and graphic design.
EDITOR’S CHOICE

How Do You Know Which Health Care Effectiveness Research You Can Trust? A Guide to Study Design for the Perplexed

Stephen B. Soumerai, ScD; Douglas Starr, MS; Sumit R. Majumdar, MD, MPH