False Dichotomies and Health Policy Research Designs: Randomized Trials Are Not Always the Answer

Stephen B. Soumerai, ScD[1], Rachel Ceccarelli, MPH[1], and Ross Koppel, PhD FACMI[2]

[1]Harvard Medical School Department of Population Medicine, Harvard Pilgrim Healthcare Institute, Boston, MA, USA; [2]Sociology Department & LDI Wharton & School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

Some medical scientists argue that only data from randomized controlled trials (RCTs) are trustworthy. They claim data from natural experiments and administrative data sets are always spurious and cannot be used to evaluate health policies and other population-wide phenomena in the real world. While many acknowledge biases caused by poor study designs, in this article we argue that several valid designs using administrative data can produce strong findings, particularly the interrupted time series (ITS) design. Many policy studies neither permit nor require an RCT for cause-and-effect inference. Framing our arguments using Campbell and Stanley's classic research design monograph, we show that several "quasi-experimental" designs, especially interrupted time series (ITS), can estimate valid effects (or non-effects) of health interventions and policies as diverse as public insurance coverage, speed limits, hospital safety programs, drug abuse regulation and withdrawal of drugs from the market. We further note the recent rapid uptake of ITS and argue for expanded training in quasi-experimental designs in medical and graduate schools and in post-doctoral curricula.

KEY WORDS: research design; health interventions; quasi-experimental design; randomization.
J Gen Intern Med 32(2):204–9
DOI: 10.1007/s11606-016-3841-9

> Information in administrative data sets is spurious by default.
> John Ioannidis[1]

This statement prolongs the polarizing debate on the trustworthiness and reproducibility of findings from "available data."[2–4] We disagree that observational data are *always* spurious.[5] While many weak observational studies are biased,[6] many valid designs using administrative data produce trustworthy findings. Moreover, RCTs can be infeasible, invalid or not generalizable despite being the "gold standard." Study end points are manipulated, or patients may not be blind to their treatment, resulting in placebo effects or exaggerated beliefs in the study treatment. Furthermore, RCTs are only useful for a fraction of health interventions, such as drugs and medical technologies.[7–9] In addition to national policies, real-life events create other unparalleled research opportunities, e.g., government seatbelt laws, banishing certain drugs from the market, changing highway speed limits,[10] high deductible health insurance,[11] changes or extreme spikes in the cost of drugs,[12,13] antibiotic controls,[14] health outcomes of the UK's pay-for-performance program,[15] anti-indoor smoking regulations,[5] and outcomes of state regulation of psychoactive drug use.[16] These policies produced important health effects, including changes in mortality, that cannot be studied experimentally.

Received June 2, 2016
Revised July 13, 2016
Accepted July 29, 2016
Published online October 18, 2016

THE INNOVATION OF QUASI-EXPERIMENTATION

In 1963, Campbell and Stanley, published their landmark text, "Experimental and Quasi-Experimental Designs for Research,"[17] revised thereafter in 1979 and 2002.[5,18] They showed several quasi-experimental research designs were often resistant to the main threats to validity such as secular trends or history bias (e.g., pre-intervention improvements in acute MI care), selection bias (e.g., study groups already healthier than controls), etc.[17] This and other texts on quasi-experimental designs have expanded the acceptance of non-experimental studies.[5,17–19]

Campbell and Stanley described three main categories of research design:

1. **Randomized Experiments**: These "gold-standard" designs randomly allocate patients or clusters (e.g., health centers) to intervention and control groups. Assuming an adequate sample size, randomization addresses most sources of selection bias and confounding. However, randomized trials can still mislead if they are too small, non-representative or not really double blind.
2. **Strong quasi-experiments**: These designs compare changes in outcomes before and after a study intervention with changes in a comparable control group. Variations include: (1) comparisons of changes in hospitalization rates after a drug safety program with simultaneous changes in multiple control groups[20] and (2) interrupted time series with or without control group(s) that measure abrupt changes from baseline trends (e.g., sudden

increases in the level or slope of emergency room admissions among the chronically mentally ill soon after a cap on public insurance benefits).[21]

3. **Weak "pre-experiments,"**[17]: This group of untrustworthy studies is not included in Cochrane systematic evidence reviews of changes in health policies or programs,[22] e.g., single observations before and after an intervention without any controls or simple cross-sectional designs that merely correlate having an intervention with mortality at a single point in time.[23–25] These study designs cannot distinguish intervention effects from what would have occurred in the absence of the intervention [e.g., they do not address the reality that more profitable and prestigious hospitals are more likely than others to invest the vast sums required for electronic health records (EHRs)]. Such studies have influenced policymakers to spend trillions of dollars on health IT technologies with few demonstrated health benefits.[26–28]

Table 1 provides a simple hierarchy of common strong and weak designs.[6,29]

INTERRUPTED TIME SERIES (WITH OR WITHOUT A CONTROL GROUP): EXAMPLES OF A QUASI-EXPERIMENTAL DESIGN

Interrupted time series designs allow researchers to control for baseline secular trends, observe a sudden effect of an intervention (a change in level or slope) and assess the stability of the change over time.[30] The design is strongest when researchers can follow another group of patients who have not experienced the intervention, i.e., a control or "comparison series." Accessible descriptions of ITS methods are numerous.[7,30–32]

Even without a perfect comparison group, ITS can be causally persuasive. Figure 1 below shows the effect of a sudden state-imposed Medicaid three-drug reimbursement limit that restricted medications among chronically ill poor patients with cardiac and other chronic illnesses.[33] Medication use plummeted immediately by half.

When advocacy organizations sued for damages, the state suddenly replaced the regulation with a less draconian $1 copayment per prescription after about a year.

Table 1 Hierarchy of Strong and Weak Designs, Based on Capacity to Control for Biases

Strong Design: Often Trustworthy Effects	
Multiple RCTs	The "gold standard" of evidence, incorporating systematic review of all RCTs of an intervention (e.g., random assignment of smoking cessation treatment).
Single RCT	A single, strong randomized experiment, but sometimes not generalizable.
Interrupted time series with control series	Baseline trends often allow visible effects and control for biases. This design has two controls: baseline trend and control group to measure sudden discontinuities in trend soon after an intervention.
Intermediate designs: Sometimes Trustworthy Effects	
Single interrupted time series	Controls for trends, but no comparison group (see above).
Before and after with comparison group	Pre-post change using single observations. Comparability of baseline trend often unknown.
Weak Designs: Rarely Trustworthy Effects (No Controls for Common Biases. Excluded from Literature Syntheses)	
Uncontrolled before and after (pre-post)	Simple observations before and after intervention, no baseline trend or control group.
Cross-sectional designs	Simple correlation, no baseline, no measure of change.

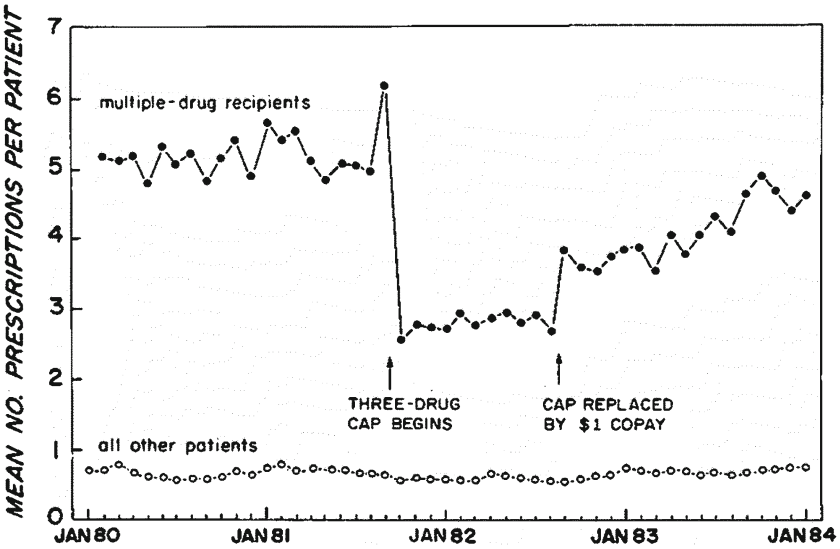

Figure 1 Times series effects of changes in drug benefit limits and cost sharing on the average number of constant-size prescriptions per continuously eligible patient per month among noninstitutionalized New Hampshire patients receiving multiple drugs (n = 860) and other outpatients (n = 8002)[33]

Immediately, the slope of prescription use increased to just below pre-cap levels. The off-on-off design and immediate, marked changes in the levels and slopes of the trend over 48 monthly observations do not allow or require an RCT to infer cause and effect. The graph of the longitudinal data is "worth a thousand p-values". Government documents also reveal no "co-interventions" (simultaneous policies that could cause the outcome) and threaten the validity of such ITS designs.

Even more important to policy and economic analysis, later time-series studies visibly showed that the sudden loss of medication access substantially increased institutionalization of frail elders and increased acute mental health care use among the severely mentally ill. The cost of hospitalization and nursing home admissions dwarfed the drug savings.[21,34] Indeed, the clearly observable ITS findings strongly contributed to many health policy improvements in the US and other countries, including rejections by many states of strict limits on drug coverage for vulnerable populations, expansion of state-funded pharmacy assistance programs,[35] and the establishment of subsidies to drug coverage under Medicare Part D.[36]

ITS can also *debunk* claimed or false "effects" via elegant and parsimonious illustrations. Figure 2 demonstrates that hospital mortality was not really affected by the nationwide (US) hospital safety program of the Institute for Healthcare

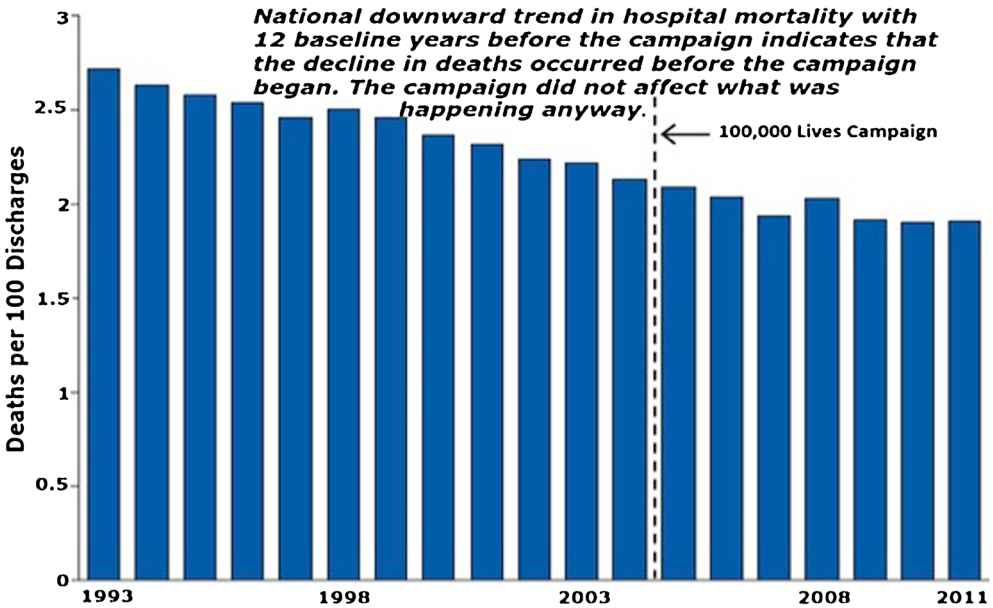

Figure 2 Example of a strong time-series design that controlled for history bias in the Institute for Healthcare Improvement's (IHI) 100,000 lives campaign. Exhibit is based on data from the Agency for Healthcare Research and Quality (HCUP, 2015).[6]

Improvement. The reported mortality decrease appears to evaporate when examined in relation to the ongoing secular trend: a fancy way of saying the investigators did not control for baseline decreases in mortality (history bias) and only focused on post-intervention data.[6] No statistics are needed to seriously question the claims of 122,000 lives saved. Using only administrative data without a control group, it is clear the decline was already happening.

Figure 3 shows increased fatal and injurious car crashes on Arizona highways with a new 65 MPH vs. a previous 55 MPH speed limit. It is an especially powerful example of ITS because the study group data come from only those highways with posted higher speed limits reflecting the new law. The large and marked upward shift immediately after the change in speed limit is obvious. In fact, Fig. 3 also displays fatal and injurious car crashes on AZ highways that did not increase the posted speed limits. In this graph there is no sudden shift in fatal and injurious car crashes. No RCT would be feasible in such a study, and the ITS and control group provide strong data on the impact of this new law.

Often the most powerful evidence is a graph that simply and reliably shows the trend and the effects of an intervention. While not infallible, ITS designs can often supplement, replicate or replace some RCTs.[7]

DISCUSSION

Between 1996 and 2015, the number of studies in PubMed identified as "interrupted time series" increased from 12 to

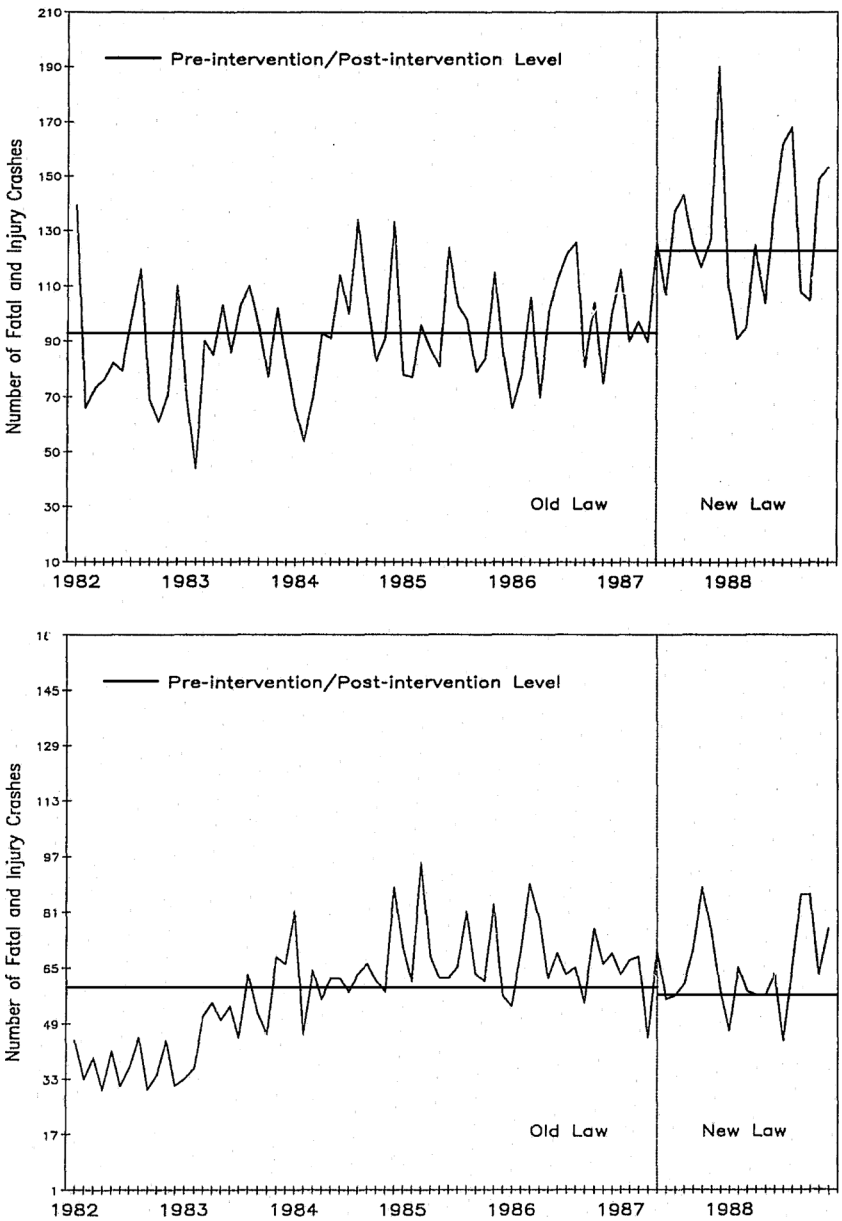

Figure 3 Upper graph shows fatal and injurious crashes on Arizona interstate highways with the increase to 65 MPH maximum speed limit. The lower graph indicates fatal and injurious crashes on Arizona interstate highways with no change in the 55 MPH maximum speed limit[10]

239 per year. Even this jump substantially undercounts such studies because many are described simply as "time series."

We hope the increasing use of this common and useful design is accompanied by an expanding acceptance of other strong non-experimental designs by medical journals and scholars.[5] As teachers we have an obligation to explain quasi-experiments to future medical researchers, along with the difference between strong and weak research designs in evaluating system-wide innovations affecting health. RCTs can only address a small proportion of interventions affecting the cost, quality and outcomes of medical and health policy interventions.

Given the influence research can have on policy, it is distressing that so much research is untrustworthy because of faulty research designs. This unease is the subject of a recent article in the US Centers for Disease Control's *Preventing Chronic Disease* entitled, "*How do you Know Which Health Care Effectiveness Research You Can Trust? A Guide to Study Design for the Perplexed*"[6] http://www.cdc.gov/pcd/issues/2015/15_0187.htm. Similarly, the National Institutes of Health (NIH) are deeply concerned about the phenomenon of "the non-reproducibility of research."[37]

Research design is often missing in the medical curriculum. Poorly controlled studies are the rule, not the exception.[38] This confuses the public, policymakers, media and researchers themselves. The countless reports (and reversals of findings)[39] regarding micronutrients and physical activities that grossly exaggerate lives saved is a case in point.[39] Accompanying the increase in what is viewed as flip-flopping research, we see a marked rise in media and researcher websites devoted to uncovering what is viewed as biased or fraudulent research.

Research design may well be the first consideration in addressing the trustworthiness of research findings.[6] Medical and graduate school curricula should emphasize the weaknesses of uncontrolled or cross-sectional designs and should include both experimental and strong quasi-experimental designs. Well-controlled and -designed studies can save lives,[40] while biased ones promote inefficient expenditures for useless programs, cause patient safety dangers and suffering, and jeopardize public health.[6]

Acknowledgments:

Contributors: *We are grateful to Caitlin Lupton for editorial assistance, her careful analysis of numerous articles and graphic design.*

Corresponding Author: *Stephen B. Soumerai, ScD; Harvard Medical School Department of Population MedicineHarvard Pilgrim Healthcare Institute, Boston, MA, USA (e-mail: ssoumerai@hms.harvard.edu).*

Compliance with ethical standards:

Funders: *This project was supported by a Developmental Research Design grant (Dr. Soumerai and Ms. Ceccarelli) from the Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute. Dr. Soumerai received grant support from the Centers for Disease Control and Prevention's Natural Experiments for Translation in Diabetes (NEXT-D). Dr. Koppel's work was in part supported by the Intel-NSF Partnership for Cyber-Physical Systems Security and Privacy.*

Conflict of interest: *The authors declare that they do not have a conflict of interest.*

REFERENCES

1. **Ioannidis JP**. Are mortality differences detected by administrative data reliable and actionable? JAMA. 2013;309(13):1410–1.
2. **Ioannidis JP**. Why most published research findings are false. PLoS Med. 2005;2(8):e124.
3. **Lehrer J**. The Truth Wears Off. The New Yorker; 2010. http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off. Accessed June 14, 2016.
4. **Freedman DH**. Lies, damned lies and medical science. Atlantic. 2010.
5. **Shadish W, Cook T, Campbell D**. Experimental and quasi-experimental designs for generalized causal inference. Belmont: Wadsworth Cengage Learning; 2002.
6. **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.
7. **Fretheim A, Zhang F, Ross-Degnan D, et al**. A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation. J Clin Epidemiol. 2015;68(3):324–33.
8. **Briesacher BA, Madden JM, Zhang F, et al**. Did Medicare Part D affect national trends in health outcomes or hospitalizations? A time-series analysis. Ann Intern Med. 2015;162(12):825–33.
9. **Soumerai SB, Ross-Degnan D, Gortmaker S, Avorn J**. Withdrawing payment for nonscientific drug therapy. Intended and unexpected effects of a large-scale natural experiment. JAMA. 1990;263(6):831–9.
10. **Epperlein T**. The Impact of the 65 MPH Speed Limit in Arizona. Arizona: Arizona Statistical Analysis Center; 1989.
11. **Wharam JF, Landon BE, Galbraith AA, Kleinman KP, Soumerai SB, Ross-Degnan D**. Emergency department use and subsequent hospitalizations among members of a high-deductible health plan. JAMA. 2007;297(10):1093–102.
12. **Ornstein C**. New hepatitis C drugs are costing Medicare billions. The Washington Post; 2015. https://www.washingtonpost.com/national/health-science/medicare-spent-45-billion-on-new-hepatitis-c-drugs-last-year-data-shows/2015/03/29/66952dde-d32a-11e4-a62f-ee745911a4ff_story.html. Accessed June 14, 2016.
13. **Roehrig C**. The Impact of New Hepatitis C Drugs on National Health Spending. 2016: Health Affairs Blog; 2015.
14. **Arnold SR, Straus SE**. Interventions to improve antibiotic prescribing practices in ambulatory care. Cochrane Database Syst Rev. 2005;4:CD003539.
15. **Serumaga B, Ross-Degnan D, Avery AJ, 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.
16. **Wagner AK, Ross-Degnan D, Gurwitz JH, et al**. Effect of New York State regulatory action on benzodiazepine prescribing and hip fracture rates. Ann Intern Med. 2007;146(2):96–103.
17. **Campbell D, Stanley J**. Experimental and quasi-experimental designs for research. Boston: Houghton Mifflin Company; 1963.
18. **Cook TD, Campbell DT**. Quasi-Experimentation: Design & Analysis Issues for Field Settings. Boston: Houghton Mifflin; 1979.
19. **McCall WA**. How to experiment in education. New York: The Macmillian Company; 1926.
20. **Lee GM, Kleinman K, Soumerai SB, et al**. Effect of nonpayment for preventable infections in US hospitals. N Engl J Med. 2012;367(15):1428–37.
21. **Soumerai SB, McLaughlin TJ, Ross-Degnan D, Casteris CS, Bollini P**. Effects of a limit on Medicaid drug-reimbursement benefits on the use of psychotropic agents and acute mental health services by patients with schizophrenia. N Engl J Med. 1994;331(10):650–5.
22. Effective Practice and Organisation of Care (EPOC). What study designs should be included in an EPOC review? EPOC Resources for review authors. http://epoc.cochrane.org/epoc-specific-resources-review-authors. Accessed June 14, 2016.

23. **Amarasingham R, Plantinga L, Diener-West M, Gaskin DJ, Powe NR**. Clinical information technologies and inpatient outcomes: a multiple hospital study. Arch Intern Med. 2009;169(2):108–14.
24. **Cebul RD, Love TE, Jain AK, Hebert CJ**. Electronic health records and quality of diabetes care. N Engl J Med. 2011;365(9):825–33.
25. **Sanghavi P, Jena AB, Newhouse JP, Zaslavsky AM**. Outcomes of basic versus advanced life support for out-of-hospital medical emergencies. Ann Intern Med. 2015;163(9):681–90.
26. **Soumerai S, Koppel R**. A major glitch for digitized health-care records. Wall Street J. 2012. http://www.wsj.com/articles/SB10000872396390443847404577627041964831020. Accessed June 14, 2016.
27. **Soumerai SB, Koppel R**. Avoiding expensive and consequential health care decisions based on weak research design August 31 ed: Health Aff Blog. 2015.
28. **Koppel R, Soumerai S**. Designing good research to support good policy: the case of health IT vol 2015: Health Aff Blog; 2015.
29. **Ackermann RT, Kenrik Duru O, Albu JB, et al**. Evaluating diabetes health policies using natural experiments: the natural experiments for translation in diabetes study. Am J Prev Med. 2015;48(6):747–54.
30. **Zhang F, Wagner AK, Soumerai SB, Ross-Degnan D**. Methods for estimating confidence intervals in interrupted time series analyses of health interventions. J Clin Epidemiol. 2009;62(2):143–8.
31. **Gillings D, Makuc D, Siegel E**. Analysis of interrupted time series mortality trends: an example to evaluate regionalized perinatal care. Am J Public Health. 1981;71(1):38–46.
32. **Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D**. Segmented regression analysis of interrupted time series studies in medication use research. J Clin Pharm Ther. 2002;27(4):299–309.
33. **Soumerai SB, Avorn J, Ross-Degnan D, Gortmaker S**. Payment restrictions for prescription drugs under Medicaid. Effects on therapy, cost, and equity. N Engl J Med. 1987;317(9):550–6.
34. **Soumerai SB, Ross-Degnan D, Avorn J, McLaughlin T, Choodnovskiy I**. Effects of Medicaid drug-payment limits on admission to hospitals and nursing homes. N Engl J Med. 1991;325(15):1072–7.
35. **Soumerai SB, Ross-Degnan D, Fortess EE, Walser BL**. Determinants of change in Medicaid pharmaceutical cost sharing: does evidence affect policy? Milbank Q. 1997;75(1):11–34.
36. **Tunis SR**. Letter from chief medical officer at the Center for Medicare & Medicaid Services, DHHS, to Stephen Soumerai. 2005.
37. 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 June 14, 2016.
38. **Urquhart C, Currell R, Grant MJ, Hardiker NR**. Nursing record systems: effects on nursing practice and healthcare outcomes. Cochrane Database Syst Rev. 2009;1:CD002099.
39. **Ioannidis JP**. Implausible results in human nutrition research. BMJ. 2013;347:f6698.
40. **β-Blocker Heart Attack Trial Research Group**. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. JAMA. 1982;247(12):1707–14.