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

Stephen B. Soumerai, ScD¹, Rachel Ceccarelli, MPH¹, and Ross Koppel, PhD FACMI²

¹Harvard Medical School Department of Population Medicine, Harvard Pilgrim Healthcare Institute, Boston, MA, USA; ²Sociology 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.

THE INNOVATION OF QUASI-EXPERIMENTATION

In 1963, Campbell and Stanley, published their landmark text, “Experimental and Quasi-Experimental Designs for Research,” revised thereafter in 1979 and 2002. 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. This and other texts on quasi-experimental designs have expanded the acceptance of non-experimental studies.

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, RCTs 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, and (2) interrupted time series with or without control group(s) that measure abrupt changes from baseline trends (e.g., sudden

Information in administrative data sets is spurious by default.

John Ioannidis¹

This statement prolongs the polarizing debate on the trustworthiness and reproducibility of findings from “available data.” We disagree that observational data are always spurious. While many weak observational studies are biased, 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. In addition to national policies, real-life events create other unparalleled research opportunities, e.g., government seatbelt laws, banning certain drugs from the market, changing highway speed limits, high deductible health insurance, changes or extreme spikes in the cost of drugs, antibiotic controls, health outcomes of the UK’s pay-for-performance program, anti-indoor smoking regulations, and outcomes of state regulation of psychoactive drug use. 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
increases in the level or slope of emergency room admissions among the chronically mentally ill soon after a cap on public insurance benefits.\textsuperscript{21}

3. **Weak “pre-experiments,”\textsuperscript{17,18}** This group of untrustworthy studies is not included in Cochrane systematic evidence reviews of changes in health policies or programs,\textsuperscript{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.\textsuperscript{23,24} 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.\textsuperscript{26–28}

Table 1 provides a simple hierarchy of common strong and weak designs.\textsuperscript{6,29}

### Table 1 Hierarchy of Strong and Weak Designs, Based on 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 RCTs of an intervention (e.g., random assignment of smoking cessation treatment).</td>
</tr>
</tbody>
</table>

| **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. |

<table>
<thead>
<tr>
<th>Intermediate designs: Sometimes Trustworthy Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single interrupted time series</strong></td>
</tr>
<tr>
<td>Controls for trends, but no comparison group (see above).</td>
</tr>
</tbody>
</table>

| **Before and after with comparison group**         |
| Pre-post change using single observations. Comparability of baseline trend often unknown. |

<table>
<thead>
<tr>
<th>Weak Designs: Rarely Trustworthy Effects (No Controls for Common Biases. Excluded from Literature Syntheses)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Uncontrolled before and after (pre-post)</strong></td>
</tr>
<tr>
<td>Simple observations before and after intervention, no baseline trend or control group.</td>
</tr>
</tbody>
</table>

| **Cross-sectional designs**                             |
| Simple correlation, no baseline, no measure of change. |
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. 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, and the establishment of subsidies to drug coverage under Medicare Part D.

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 Improvement’s (IHI) 100,000 lives campaign. Exhibit is based on data from the Agency for Healthcare Research and Quality (HCUP, 2015).
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. 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.

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.
239 per year. Even this jump substantially undercuts 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. 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” 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.”

Research design is often missing in the medical curriculum. Poorly controlled studies are the rule, not the exception. This confuses the public, policymakers, media and researchers themselves. The countless reports (and reversals of findings) regarding micronutrients and physical activities that grossly exaggerate lives saved is a case in point. 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. 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, while biased ones promote inefficient expenditures for useless programs, cause patient safety dangers and suffering, and jeopardize public health.

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

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

REFERENCES

27. Soumerai SB, Koppel R. Avoiding expensive and consequential health care decisions based on weak research design August 31 ed: Health Aff Blog. 2015.