ELSEVIER

Journal of Clinical Epidemiology 62 (2009) 959–966

Journal of Clinical Epidemiology

Nonrandomized quality improvement intervention trials might overstate the strength of causal inference of their findings

Linda C. Li[a,b,*], Lorenzo Moja[c], Alberto Romero[d], Eric C. Sayre[e,f], Jeremy M. Grimshaw[g,h]

[a]*Department of Physical Therapy, University of British Columbia, Vancouver, British Columbia, Canada*
[b]*Arthritis Research Centre of Canada, Vancouver, British Columbia, Canada*
[c]*Italian Cochrane Centre, Mario Negri Institute for Pharmacological Research, Milano, Italy*
[d]*Grupo REDEGUIAS, Red_MBE (G03/90), Hospital Universitario Virgen de Valme, Seville, Spain*
[e]*Department of Statistics and Actuarial Science, Simon Fraser University, Vancouver, British Columbia, Canada*
[f]*Arthritis Research Centre of Canada, Vancouver, British Columbia, Canada*
[g]*Clinical Epidemiology Program, Ottawa Health Research Institute, Ottawa, Canada*
[h]*Centre for Best Practice, Institute of Population Health, University of Ottawa, Ottawa, Canada*

Accepted 20 October 2008

Abstract

Objective: To assess the strength of causal inferences reported in randomized and nonrandomized evaluations of quality improvement (QI) interventions in relation to the study design and the direction of results for the primary outcomes.

Study Design and Setting: We searched 11 journals for QI intervention studies that aimed to change clinician behavior. Statements that addressed the causal inference between intervention and outcomes were extracted and were rated by 34 researchers for the strength of causality.

Results: We found 38 randomized controlled trials (RCTs) and 35 non-RCTs, and extracted 68 quotes from the abstracts and 139 from the main text. A significant interaction was found between study design and direction of results for the abstract quotes ($P = 0.022$). The ratings for non-RCTs were higher when the results were mixed, but for RCTs, they were higher if the results were positive or no effect, although none of the differences were statistically significant at $\alpha = 0.05$ after adjusting for multiple comparisons. For the main text quotes, the causality rating was higher by 0.43 for RCTs than for non-RCTs after adjusting for the direction of results ($P < 0.001$).

Conclusion: Authors might have overstated the strength of causal inference in the abstracts of non-RCTs, but appeared to report causality appropriately in the main text.

Keywords: Causality; Research reporting; Quality improvement intervention; Review; Study design; Causal inference

1. Introduction

In spite of good intentions, research findings and practice guidelines are not always translated into clinical practice [1–4]. This has led to the development of quality improvement (QI) interventions to change professional behavior based on research evidence. Randomized controlled trials (RCTs) are the gold standard for evaluating the effect of QI interventions [5,6]; however, they are sometimes methodologically challenging because of political, practical, and ethical barriers. Consequently, many QI studies have used non-RCTs (quasi-experimental or observational studies) [5–7].

Non-RCTs are intrinsically weaker research designs, because they are prone to a wider range of potential risks of bias; for example, the influence of secular trends or sudden changes in uncontrolled before-and-after studies, or other concurrent events in interrupted time series designs [7]. These problems are minimized in RCTs, where the effect of an intervention is estimated through direct comparison with a control group that receives either no treatment or usual care. Randomization ensures that all known and unknown biases are distributed evenly at the start of the study. As such, any differences observed postintervention can be more confidently attributed to the intervention. RCTs are, therefore, the most robust design for making causal inferences in health care interventions [8,9].

Cook and Campbell argue that researchers need to explore plausible rival hypotheses when interpreting the results of non-RCTs [10]. Given that it is frequently

* Corresponding author. Arthritis Research Centre of Canada, 895 West, 10th Avenue, Room 324, Vancouver, British Columbia V5Z 1L7, Canada. Tel.: +604-871-4577; fax: +604-879-3791.

E-mail address: lli@arthritisresearch.ca (L.C. Li).

0895-4356/09/$ – see front matter
doi: 10.1016/j.jclinepi.2008.10.008

What is new?

- Methodological considerations suggest that researchers should be more circumspect about making causal statements for nonrandomized studies compared with randomized studies.
- In a review of quality improvement (QI) intervention studies, we found that researchers overstated the strength of causal inference in the abstracts of nonrandomized studies, but reported causality appropriately in the main text.
- There is a need for improvement for the reporting of QI intervention studies. We encourage international method groups to develop standards for reporting conclusions.
- We encourage peer reviewers and journal editors to scrutinize the conclusions of QI intervention studies to ensure that their conclusions are appropriate given the research design and results.

difficult to rule out such rival hypotheses in non-RCTs, researchers should be more cautious in their interpretation of the strength of causal inference. However, it is not clear that such concerns are accurately reflected in the wording of conclusions in peer-reviewed publications. For example, in an uncontrolled before-and-after study on the use of a multidisciplinary education program for reducing feeding tube use in patients with dementia, the authors stated that, ''interdisciplinary teamwork and focused educational effort can rapidly produce change in practice'' [11]. This conclusion was subsequently criticized for being overzealous in the inference of causality because of potential rival hypotheses that might have explained the findings [12].

The primary objective of the current study was to explore the strength of causal inference in the abstract and discussion sections of RCTs and non-RCTs evaluating QI interventions. From methodological theory, we anticipated that the strength of causal inference made in non-RCTs should be weaker than those in RCTs after adjusting for the direction of results, because it is difficult to exclude other plausible reasons for the observed findings. Our null hypothesis was that the study design (RCT vs. non-RCT) had no effect on the strength of causality statements after adjusting for the direction of results. Our secondary objective was to examine the vocabulary and tone used by investigators to address causal relationships.

2. Methods

We hand-searched 11 medical or health services research journals between January 1, 2002 and December 31, 2003 to identify evaluations of QI interventions (Table 1). We included all studies evaluating interventions that aim to change health professional behavior based on research evidence [13]. This included various forms of continuing education, quality assurance, and behavioral interventions that can affect the ability of health professionals to deliver services more effectively and efficiently. Studies evaluating strategies that solely aimed to change the behavior of patients were excluded.

We included studies that used experimental designs (RCTs or controlled clinical trials), quasi-experimental designs (controlled or uncontrolled before-and-after studies, interrupted time series), or observational designs (concurrent cohort studies, case–control studies, case series). The taxonomy of quasi-experimental and observational designs by Deeks et al. [14] was used. Further, we added the category ''interrupted time-series'' using the definition by Shadish et al. [7]. The definitions of study designs are summarized in Table 2.

2.1. Data extraction form

Two investigators (L.C.L., A.R., or L.M.) reviewed the eligible studies and recorded the following information: (1) study characteristics, including the type of interventions, the type and number of participants, clinical conditions and settings; (2) the study design (i.e., cluster RCT and RCT = ''RCT''; other designs = ''non-RCT''); (3) the direction of main results (i.e., positive: the primary outcome[s] of the intervention was statistically significantly better; detrimental: the primary outcome[s] of control was statistically significantly better; mixed results: where there were multiple primary outcomes, some showed that the intervention was statistically significantly better and others showed the opposite; or no effect); and (4) all concluding statements addressing the causal relationship of the intervention and outcomes in the abstract and the main text. A concluding statement had to explicitly include the intervention(s) and phrases about the causal relationship, with or without the outcome(s) (see Table 3 for examples). We excluded phrases about implications for practice. The data extraction form was pretested with seven randomly selected QI intervention articles. Disagreements between reviewers were resolved by discussion or by a third reviewer if no consensus was met.

2.2. Assessment of the strength of causal inferences

We assembled a panel of clinical epidemiologists and health service researchers to assess the strength of causal inference suggested by the concluding statements. Through a balance randomization, the quotes were split into seven packages (each consisting of 10–11 articles). We assigned two of the seven packages to each panelist using the biased-coin randomization. Each panelist then received an