Viewpoint

Enhancing causal interpretations of quality improvement interventions

G Cable

Abstract
In an era of chronic resource scarcity it is critical that quality improvement professionals have confidence that their project activities cause measured change. A commonly used research design, the single group pre-test/post-test design, provides little insight into whether quality improvement interventions cause measured outcomes. A re-evaluation of a quality improvement programme designed to reduce the percentage of bilateral cardiac catheterisations for the period from January 1991 to October 1996 in three catheterisation laboratories in a north eastern state in the USA was performed using an interrupted time series design with switching replications. The accuracy and causal interpretability of the findings were considerably improved compared with the original evaluation design. Moreover, the re-evaluation provided tangible evidence in support of the suggestion that more rigorous designs can and should be more widely employed to improve the causal interpretability of quality improvement efforts. Evaluation designs for quality improvement projects should be constructed to provide a reasonable opportunity, given available time and resources, for causal interpretation of the results. Evaluators of quality improvement initiatives may infrequently have access to randomised designs. Nonetheless, as shown here, other very rigorous research designs are available for improving causal interpretability. Unilateral methodological surrender need not be the only alternative to randomised experiments.
(*Quality in Health Care* 2001;**10**:179–186)

Keywords: causal interpretations; quality improvement; interrupted time series design; implementation fidelity

Key messages
- In an era of chronic resource scarcity it is critical that quality improvement professionals have confidence that their project activities cause measured change.
- A commonly used research design, the single group pre-test/post-test design, provides little insight into whether quality improvement interventions cause measured outcomes.
- Many other quasi-experimental designs can be employed in most contexts instead of single group pre-test/post-test designs, and provide much greater causal interpretability of the findings of the quality improvement project evaluation. One of the most powerful of these is interrupted time series designs.

Implications of these findings for quality improvement
The adoption of more rigorous research designs to evaluate quality improvement efforts will enable quality improvement professionals to know with greater confidence whether the efforts actually work. In turn, projects determined to be successful based on this more compelling evidence can more confidently be diffused as best practices. A latent benefit of improving the level of rigour in the evaluation research design is that the perception of quality improvement projects as "scientific" will be enhanced.

The Quality Institute, Atlantic Health System, 325 Columbia Turnpike, Florham Park, NJ 07932, USA
G Cable, *director of research*

Correspondence to:
Dr G Cable
drcableg@hotmail.com

Accepted 30 March 2001

Actuating beneficial change is the *raison d'être* of quality improvement efforts. In an era when private entities and governments are increasingly less willing to pay for care, quality improvement initiatives must additionally be able to demonstrate superiority over competing strategies, or the option of doing nothing at all.[1–4] In the current milieu, every major quality improvement project should include an evaluation research design that permits rigorous testing of the extent to which improvement efforts actually cause measured change. The most commonly employed research designs—single group pre-test/post-test designs and one shot case studies (that is, a single group design with a post-test only)—provide little evidence of the causal impact of improvement activities.[5] Consequently, methodological rigour is often unnecessarily sacrificed when designs are available which can improve the causal interpretability of the results. The purpose of this paper is to make a case for improving the rigour of research designs in order to enhance

the causal interpretability of quality improvement efforts.

The first section of the paper provides background regarding the comparative virtues of different research designs for making causal statements about intervention effects. This section also contains a description of a rigorous quasi-experimental design that can often be employed in quality improvement studies instead of less rigorous single group pre-test/post-test designs. The second part of the paper contains a comparison of the results of a re-evaluation of a quality improvement project conducted to reduce the rates of bilateral cardiac catheterisations with the results of the original evaluation. The original evaluation used a single group pre-test/post-test design while the quasi-experimental design described in the first part of the paper is employed in the re-evaluation. The comparison shows the value of improving the rigour of the evaluation research design for enhancing the causal interpretability of improvement efforts.

Comparing the internal validity of research designs: randomised trials, single group pre-test/post-test designs, and quasi-experiments

The primary purpose of research design is to provide investigators with evidence regarding the degree to which an intervention *causes* measured outcomes. A causal relationship is one in which four conditions are met: (1) a measurable cause precedes a measurable effect and the timing of the effect is consistent with the nature of the mechanism behind the cause; (2) the magnitude (including duration) of the effect is proportional to that of the cause; (3) the effect is not present in the absence of the action; and (4) all plausible competing explanations for the effect can be ruled out. The degree to which an investigator can infer cause is reflected in the level of internal validity of the research design. Designs with high internal validity provide the investigator with great confidence, *ceteris paribus*, that a manipulated variable—for example, in this context, the activities of a quality improvement effort—caused the observed change in the outcome variable(s).[5 6]

Randomised controlled trials are widely recognised as having high levels of internal validity primarily due to the creation by the investigator of two or more mathematically equivalent comparison groups.[5–7] Group equivalence is achieved through the process of random assignment, assignment that assures that each case has the same probability of being assigned to the two (or more) arms of the trial. When sample sizes are sufficiently large, random assignment of cases to treatment arms provides great confidence before the implementation of the intervention that the arms are equivalent on all known and unknown factors that might affect the outcome measures employed in the trial. Hence, upon normal completion of the trial investigators can have great confidence, quantifiable by statistical intervals and other measures, that differences between the arms of the trial are due largely to variables they themselves have manipulated.[5–7] In many contexts, however, it is not feasible to design and implement a randomised controlled trial because investigators have little control over how cases are assigned to treatment arms, or an appropriate comparison group simply cannot be assembled. Unfortunately, in quality improvement research this often leads investigators to use the single group pre-test/post-test design or one closely related as the primary fall back design, resulting in a large decrease in the level of internal validity.

In contrast to randomised controlled designs, single group pre-test/post-test designs have comparatively low internal validity in part because they provide the investigator with no counterfactual—that is, no evidence regarding what would have happened in the absence of the intervention. In the absence of a relevant counterfactual, changes in the outcome measure from pre-test to post-test can be attributed to many factors other than the intervention.[8] These factors include, but are not limited to, such things as external events, cyclical variations in the outcome measure, and undiscovered changes in the instrumentation employed to collect outcome. The fall off in the level of internal validity is wholly unnecessary in most instances, given the availability of several rigorous quasi-experimental designs.[5 6 8]

One family of quasi-experimental designs with generally high levels of internal validity is interrupted times series designs. Time series are data collected at equal intervals over time. These data may be collected for any unit of analysis. In medicine, time series data might be collected to capture weekly medication ordering errors in a hospital, to monitor variations in hourly temperature readings for a septic patient, or as monthly rates of bilateral cardiac catheterisations (as in the example presented below). Time series data can then be represented through graphical techniques and mathematical modelling in a way that permits the identification of systematic (and potentially causal) factors and non-systematic factors.[5 9–11]

Interrupted time series, as the name implies, are time series during which an event occurs that is thought to "interrupt" the existing numerical variation in the series in some systematic manner, as when administration of the antibiotics (the event) interrupts the hourly variation in the temperature of the septic patient. Following the event the series is expected to change in level, slope, and/or shape based on some *a priori* understanding of the mechanism through which the event causes change. An interrupted time series design can be depicted as follows with a commonly used notation of research design[5 6]:

$O_1\ O_2\ O_3\ O_4\ O_5\ O_6\ \mathbf{X}\ O_7\ O_8\ O_9\ O_{10}\ O_{11}\ O_{12}\ O_{13}\ O_{14}\ O_{15}\ O_{16}\ O_{17}\ O_{18}$

In this example the series is 18 equal interval periods long. The "O"s depict the collection of data for each period (O for observation). The position of the "X" indicates that the event occurred between periods 6 and 7.

Interrupted time series designs can be used to determine whether the empirical post-event

series is different in some systematic fashion from the series before the event. Powerful mathematical modelling techniques have been developed to determine the effect of the event on the series, independent of other systematic factors such as other discrete events and seasonal or cyclical variations in the series.[10 11] The net effect is a design that provides much greater confidence than single group pre-test/post-test designs that the event "interrupting" the series actually caused post-event changes in the series.

An especially powerful interrupted time series design is one with "switching replications" in which time series of identical length are assembled for two or more non-equivalent (that is, not the product of random assignment) but nonetheless similar groups—for example, two hospitals in the same area. Each series in the design will have experienced the event of interest, but at different periods in the series. The switching replications design for two groups in which the event has occurred is depicted as follows:

group 1: $O_1\ O_2\ O_3\ O_4\ O_5\ O_6\ \mathbf{X}\ O_7\ O_8\ O_9\ O_{10}\ O_{11}\ O_{12}\ O_{13}\ O_{14}\ O_{15}\ O_{16}\ O_{17}\ O_{18}$

group 2: $O_1\ O_2\ O_3\ O_4\ O_5\ O_6\ O_7\ O_8\ O_9\ O_{10}\ O_{11}\ O_{12}\ \mathbf{X}\ O_{13}\ O_{14}\ O_{15}\ O_{16}\ O_{17}\ O_{18}$

In this example the event occurred between periods 6 and 7 in the first group under study and between periods 12 and 13 in the second group. Each series can serve as a comparison—a counterfactual—for the other, because the series are coeval. As a result, the internal validity of this design greatly exceeds that of single group pre-test/post-test designs as well as that of a single group interrupted time series.[5 8]

When a study uses concomitant time series in which the events occur at different periods as in the switching replications design, researchers can more readily detect the potential presence of "history" threats to internal validity. History threats are events or processes other than the event of interest that can affect the variation of the outcome measure(s) under study. For this reason, history threats to internal validity are sometimes referred to as "external" events.[8] History threats often go undetected and potentially confound the investigator's ability to imply cause from the effect of the interventions.[5 6 8] However, when the design includes at least two series as in a switching replications design, a "common" history threat will probably register as a change that occurs at similar periods, and of similar magnitude and duration in all series. Hence, the threat posed by the external event to the causal interpretability of the intervention effect is essentially neutralised because the investigator can detect whether and when it is present in more than one series.[5 8]

In the next section the switching replications design is used to re-evaluate a quality improvement effort implemented in the mid 1990s in one state in the USA. The re-evaluation shows how the use of this design provides greater insight into the extent to which improvement efforts caused the measured outcomes than the single group pre-test/post-test design used in the original evaluation.

> In 1991 the American College of Cardiology and the American Heart Association published guidelines regarding cardiac catheterisation and catheterisation laboratories in response to the increases in the numbers of catheterisations being performed, changes in the reason why they were being performed, and the setting in which they were being performed—that is, other than in hospital based laboratories. In addition to clinical practice issues, the guidelines addressed ethical concerns including patient safety, conflicts of interest related to ownership, operation, self-referrals, and advertising of services.

Box 1 Rationale behind new guidelines for right heart catheterisations.

Comparison of switching replications interrupted time series design with a single group pre-test and post-test design: PRO Cardiac Catheterisation Project

BACKGROUND TO THE ORIGINAL PROJECT

Since the implementation of a programme developed by the Health Care Finance Administration in the mid 1990s to improve care provided to Medicare patients, medical peer review organisations (PROs) have completed hundreds of improvement projects.[12 13] In 1994 a PRO in a north eastern state in the USA initiated a quality improvement project designed to reduce the use of bilateral cardiac catheterisations in the Medicare population (aged 65 and over) within a few participating catheterisation laboratories in the state.[14] The project was initiated in response to new guidelines which held that right heart catheterisations are unnecessary for diagnostic purposes in the absence of specific clinical indications (box 1).[15]

The intervention for each laboratory included the following components: (1) a policy change requiring documentation to perform a bilateral catheterisation; (2) staff education regarding the policy change; (3) a multidisciplinary approach to the development, implementation, and assessment of the improvement plan; and (4) development of plans of action to address policy non-compliance (box 2).[14]

In the original analysis the PRO used a single group pre-test/post-test design to evaluate the success of the improvement efforts in each laboratory. The original evaluation measured change from a pre-intervention period from January to December 1993 (that is, one data point for the pre-intervention period) to each quarter of 1995 and the 1995 period as a whole (one data point for the post-intervention period).[14] No data were analysed from 1994, the year in which interventions were implemented. This design can be depicted for each laboratory as:

$O_{1993}\ \mathbf{X}_{1994}\ O_{1995}$ where the "O"s represent measurement periods and "X" the timing of the intervention. Hence, it is clear from the

- A policy change requiring documentation to perform a cardiac catheterisation on both sides of the heart including a checklist of the indications for the bilateral catheterisation to be part of the patient's medical record
- Staff in-service education regarding the policy change
- A multidisciplinary approach to the development, implementation, and assessment of the improvement plan—specifically, inclusion of a representative from all professions potentially affected by the plan
- A mechanism for monitoring the implementation of the plan
- Development of plans of action to address policy non-compliance (for example, what the administrative steps would be should a physician refuse to comply with policy changes)

Box 2 Elements common to all laboratory interventions.

notation depicting this design that temporal variation in bilateral catheterisations went unmeasured at intervals *within* the pre-intervention period, which included all of 1993 up to the point when interventions were implemented in 1994 in each laboratory. The omission of information could have provided insight into seasonal or cyclical variation in bilateral catheterisation rates, as well as trending evidence.

The results of the original evaluation indicated that bilateral catheterisation rates declined in each laboratory in 1995 from baseline levels of 1993. In laboratory A the decline from 1993 to 1995 was from 87.9% to 41%, in laboratory B rates fell from 82% to 40.4%, and in laboratory C the decline was from 97.9% to 58.1%.[14] The use of this single group pre-test and post-test design in which no data were collected during 1994, the year in which the projects were implemented in each laboratory, makes it difficult to assess the nature of the impact of the individual interventions. More importantly, this design is poorly equipped to assess whether the improvement efforts *caused* the measured changes. Moreover, the design ignores the potentially unique effect on rates in one laboratory in which it was decided to augment the project activities common to all laboratory interventions by changing the packaging of the catheterisation trays. This component of the intervention required the cardiologist to make a specific request in order to catheterise both sides of the heart which compelled him or her to provide the clinical justification to catheterise the second side of the heart—for example, pulmonary hypertension, mitral or aortic valve disease.[14]

METHODS USED TO RE-EVALUATE THE PRO PROJECT

An interrupted time series design with switching replications was employed to assess the impact of the individual interventions in the three laboratories. In this re-evaluation, monthly bilateral catheterisation rates were analysed for the period from January 1991 to October 1996, 70 consecutive months of bilateral catheterisation rates for each laboratory. The data were assembled from Medicare discharge databases, the same source of data as was used in the original evaluation.[14] This extended duration was chosen to provide sufficiently long pre-intervention and post-intervention periods to examine the impact of the interventions after controlling for potentially confounding systematic variation (discussed in greater detail below). To maintain confidentiality, the three laboratories are referred to hereafter as laboratory A, B and C, respectively. Laboratory A implemented their intervention in July 1994, laboratory B in October, and laboratory C in December, respectively. Our re-evaluation design can be depicted as follows:

Laboratory A: $O_{91,\ Jan}\ O\ O\ O.\ O.\ O_{94,\ Jul}\ \mathbf{X}\ O_{94,\ Aug}\ O\ O\ O\ O\ O\ O\ O\ O.\ O.\ O_{96,\ Oct}$

Laboratory B: $O_{91,\ Jan}\ O\ O\ O.\ O.\ O.\ O.\ O.\ O.\ O_{94,\ Oct}\ \mathbf{X}\ O_{94,\ Nov}\ O\ O.\ O.\ O.\ O.\ O_{96,\ Oct}$

Laboratory C: $O_{91,\ Jan}\ O\ O\ O.\ O.\ O.\ O.\ O.\ O.\ O\ O.\ O.\ O_{94,\ Dec}\ \mathbf{X}\ O_{95,\ Jan.}\ O\ O.\ O_{96,\ Oct}$.

It should be apparent that the switching replications design permits a more rigorous test than the original design of the possibility that the implementation of laboratory A's intervention in July independently affected the bilateral catheterisation rates for laboratories B and C before implementation of their interventions. By extension, this design also facilitates the evaluation of whether laboratory C's rates were independently affected by the implementation of laboratory A's and/or laboratory B's intervention.

STATISTICAL ANALYSIS

The quantitative impact of individual interventions was determined using Box and Tiao's methods. Based on Box and Jenkin's autoregressive integrated moving average (ARIMA) models, Box–Tiao methods allow the modelling of systematic structures in each laboratory's series, including potentially confounding seasonal and cyclical effects.[5 10 11] In other words, this part of the modelling process permits the analyst to filter out the potentially confounding effects of systematic variation in time series in order to determine the magnitude and structure of the effect of the interventions. After first constructing ARIMA models for each laboratory, variables representing the timing of the impact of the intervention in each laboratory were added. Three possible forms of the impact were tested in each laboratory: (1) an abrupt permanent effect; (2) a gradual permanent effect; and (3) an abrupt temporary effect. Statistical hypotheses were tested at α=0.05. Ljung–Box's Q statistic was employed as a goodness of fit measure for the models of each laboratory. The statistic tests the null hypothesis that model residuals are not autocorrelated. Thus, *failure* to reject this hypothesis indicates that the empirical model fits the data well.[16] Analyses were conducted using

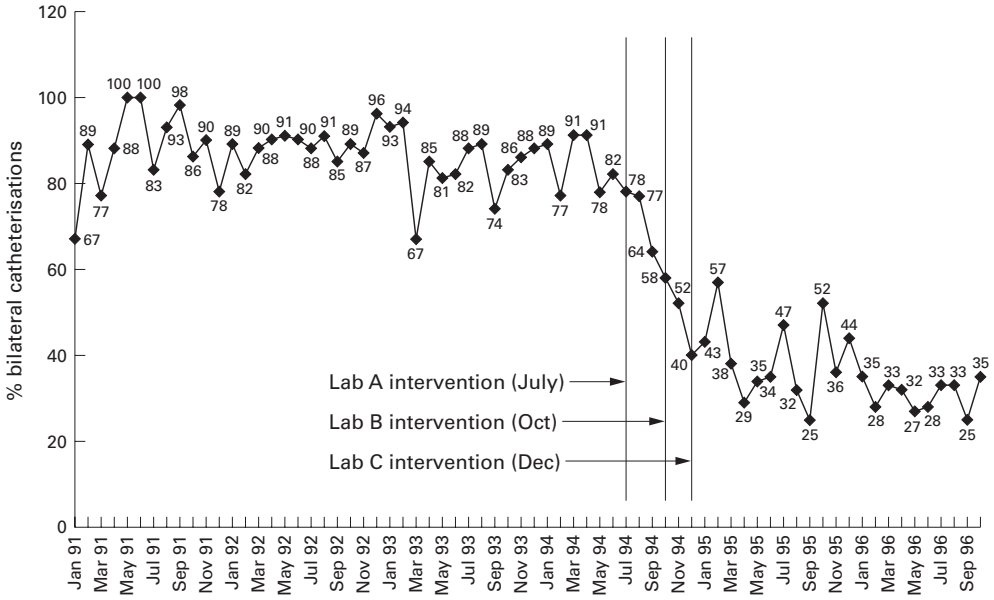

Figure 1 Percentage bilateral catheterisations at laboratory A monthly from January 1991 to October 1996, with delineation of pre-intervention and post-intervention periods for laboratory A, and laboratories B and C.

SAS ETS version 6.12. Graphics were produced in Excel version 5.0.

RESULTS OF THE RE-EVALUATION

Monthly bilateral catheterisation data are depicted for laboratories A, B, and C in figs 1, 2, and 3, respectively. Three vertical lines were drawn in each figure to indicate the timing of the intervention in each laboratory. Laboratory A's data (fig 1) reveal a steep fall in bilateral catheterisation rates immediately following intervention, a decline that had no precedent during the 42 month pre-intervention period in this laboratory. Laboratory B's data (fig 2) suggest that the post-intervention declines in this laboratory are a continuation of the decrease in bilateral catheterisation rates that began in about November 1993, 11 months before implementation of the intervention in October and 8 months before the July implementation of the intervention in laboratory A. Figure 3 shows that the decline in bilateral catheterisation rates in laboratory C appear to start as early as June or July 1994 and level off in the subsequent months leading up to the intervention in December. However, the declines appear to have resumed at an accelerated rate following implementation of the intervention in laboratory C. At this point it is clear that the original evaluation could have used a simple plot of monthly bilateral catheterisation rates in each laboratory to reveal systematic elements of variation that may have been present before implementation of the quality improvement activities. This alone would have significantly improved the internal validity of the original evaluation design.

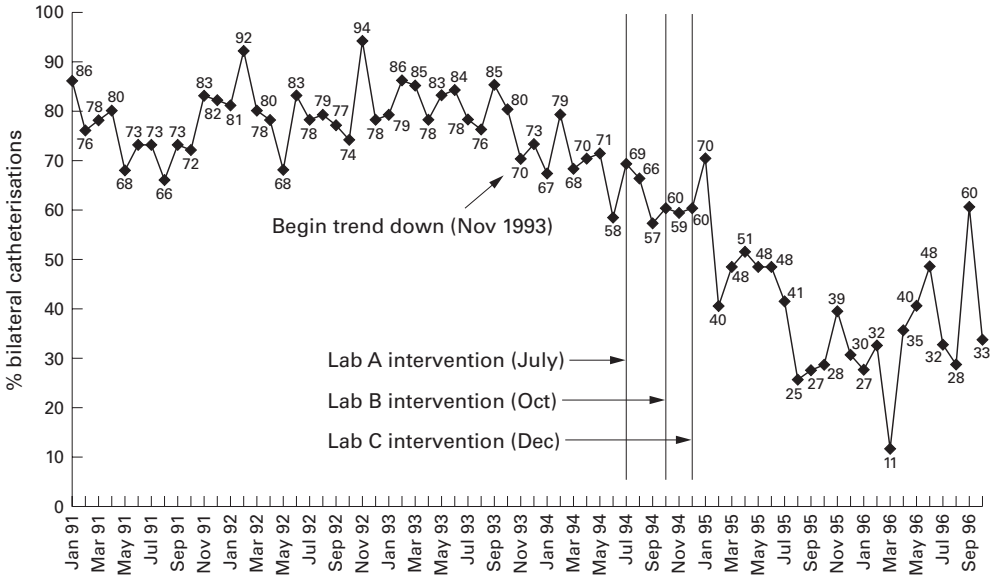

Figure 2 Percentage bilateral catheterisations at laboratory B monthly from January 1991 to October 1996, with delineation of pre-intervention and post-intervention periods for laboratory B, and laboratories A and C.

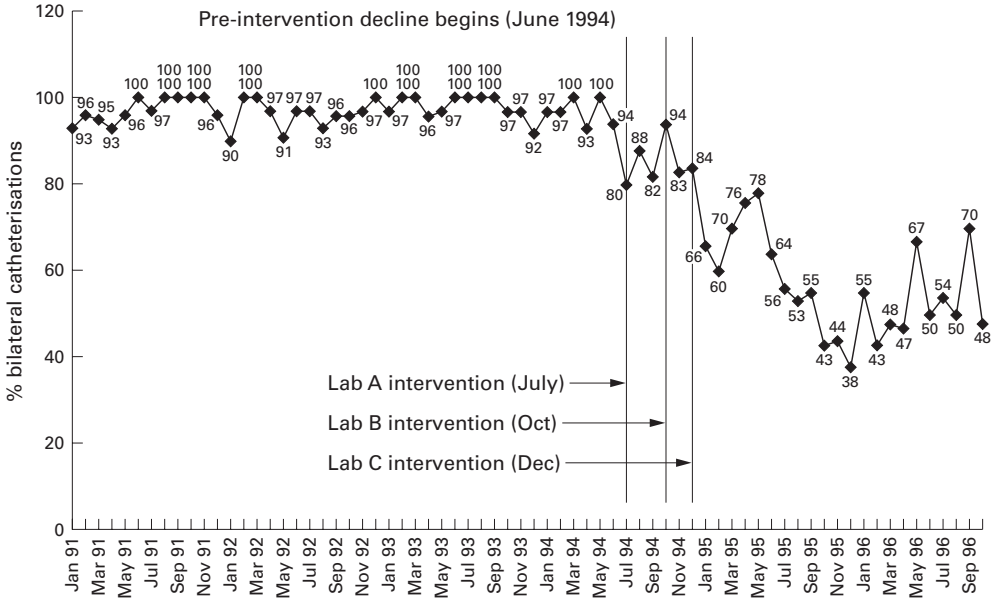

Figure 3 Percentage bilateral catheterisations at laboratory C monthly from January 1991 to October 1996, with delineation of pre-intervention and post-intervention periods for laboratory C, and laboratories A and B.

Some evidence exists of a "local" history threat to the internal validity of the re-evaluation in which an external event or events has affected bilateral catheterisation rates in only one of the laboratories.[5 8] Specifically, in laboratory B, but not A or C, something other than the intervention appears to have initiated a long monotonic decline in rates beginning in about November 1993 that continued throughout the remainder of the period of study. This could be in part the result of a change in the composition of patients in the service area of laboratory B (for example, a younger healthier population), practice changes within the laboratory, or a change in how bilateral catheterisations were measured. Since we do not have these data for laboratory B, or any data for laboratories near it, we are unable to determine whether this downward trend was something unique to laboratory B or the result of event(s) common to other laboratories in proximity to it. Regardless of this, it is evident that the identification of this change in bilateral catheterisation rates in laboratory B beginning in November 1993 was also not possible with the original PRO research design.

The results of Box and Tiao modelling indicate that the intervention in laboratory A was associated with a mean decrease in bilateral catheterisations of almost 2% each month (t=–0.40, p=0.04) in the post-intervention period as part of an ARIMA (0,1,1) model (Ljung–Box Q, p=0.95, that is, fail to reject the hypothesis that model residuals are not autocorrelated, that is, the model fits the data well). The data from laboratory B were first transformed into natural logarithmic form to achieve homogeneity of variance in the series (a necessary prerequisite in Box–Jenkins models).[10 11] The intervention in laboratory B was associated with a not statistically significant and small percentage decrease of 0.0005 per month in the post-intervention period as part of an ARIMA (0,1,1) model (Ljung–Box Q, p=0.80). The data from laboratory C were also modelled using a natural logarithm transformation. The intervention at laboratory C was associated with a mean percentage decrease of 0.03 per month (t=–4.81, p=0.02) in an ARIMA (0,1,1) model without a constant (Ljung–Box Q, p=0.99).

Finally, laboratories A and C are in close geographical proximity to each other (a few miles apart within the same city), raising the possibility that "cross talk" regarding laboratory A's intervention occurred which began a de facto intervention in laboratory C before the implementation of the formal intervention in laboratory C. We therefore examined whether the implementation of laboratory A's intervention had a measurable effect on laboratory C's bilateral catheterisation rates, independent of laboratory C's formal intervention. Although fig 3 provides some graphical evidence in support of this hypothesis, the evidence from the Box–Tiao modelling process suggests that no statistically significant decrease occurred in the data in laboratory C as a result of the implementation of the intervention in laboratory A. Moreover, no statistical evidence existed that the October intervention in laboratory B had an independent effect on the rates in laboratory C. Similarly, no evidence was seen that the intervention in laboratory A had an independent effect on bilateral catheterisation rates in laboratory B.

DISCUSSION OF RE-EVALUATION RESULTS

Our results strongly suggest that the quality improvement interventions reduced bilateral catheterisation rates in two of the three laboratories, laboratories A and C. The evidence indicates that the intervention in laboratory A resulted in an immediate fall in bilateral catheterisation rates that continued during the period of study. The effect of the intervention in laboratory C was somewhat delayed and smaller in magnitude. Visual evidence suggests

that near the end of the period of study the declines might have ceased in laboratory C. Laboratory A was the laboratory that augmented the activities common to all of the laboratory interventions by changing the packaging of the catheterisation trays, thus requiring that cardiologists make special requests to catheterise both sides of the heart. The immediate large magnitude of the effect resulting from intervention in laboratory A suggests that this innovation may have increased the success of the intervention. Perhaps of greater significance for our purposes is the fact that the success of the change in packaging of catheterisation trays could not have been discovered using the original single group pre-test/post-test design. Hence, the importance of using an appropriately rigorous research design is brought into greater relief.

Our findings contrast clearly with those of the original evaluation which described large numerical decreases "following" the quality improvement efforts—that is, between 1993 and 1995—in all of the catheterisation laboratories.[14] The original analysis did not, however, tell the entire story of the quality improvement efforts. Rates did, indeed, fall in all the laboratories between 1993 and 1995, but more rigorous evidence from the re-evaluation suggests that the improvement efforts were efficacious in only two of the three laboratories.

The design employed in our re-evaluation therefore provides demonstrably greater confidence in the causal effect of the improvement efforts in laboratories A and C, as well as evidence regarding the magnitude and duration of the effects (through the Box and Tiao models). Moreover, in contrast to the findings of the original evaluation, the switching replications design provides greater evidence that the intervention in laboratory B was not effective.

Potentially undetected external events threaten our ability to attribute cause with even greater confidence to the interventions in laboratories A and C. Specifically, there may have been events external to the quality improvement intervention that occurred ahead of, or at the same time as, the implementation of the improvement activities in laboratories A and C that can explain post-intervention changes in the series. For example, in March 1994 the US government's Agency for Health Care Policy and Research and National Heart, Lung, and Blood Institute jointly released practice guidelines for the diagnosis and management of unstable angina.[17] The guideline contains recommendations regarding the use of cardiac catheterisation. We cannot therefore rule out the possibility that some part of the decline in bilateral catheterisation rates in both laboratories can be attributed to the momentum for change in practice created by the release of these guidelines. However, this threat also would have affected the original evaluation. Moreover, the single group pre-test/post-test design used in the original evaluation is subject to a panoply of additional threats.[5 6 8] In sum, we believe our evaluation research design has radically improved the accuracy and the causal interpretability of the findings. Moreover, the re-evaluation provides tangible evidence in support of our position that more rigorous designs can and should be more widely employed to improve the causal interpretability of quality improvement efforts.

Conclusions

This paper has attempted to make a case for improving the rigor of evaluation designs used in quality improvement projects in order to enhance the causal interpretability of the results. We have focused on a particular interrupted time series design—switching replications—as one time series design that can be used in many contexts in place of designs with lower internal validity. We recognise that in some contexts a switching replications design is not feasible because the intervention cannot be implemented in a second group. When working under this constraint, single group interrupted time series will still provide greater ability to make causal inferences than the seemingly ubiquitous single group pre-test/post-test design.[5 6]

Two especially compelling interrupted time series alternatives to a single group pre-test/post-test design are the single group interrupted design with a removed treatment and the single group interrupted design with a non-equivalent dependent variable.[5] The removed treatment design can be used when the investigators have the ability to institute the intervention and then subsequently remove it, preferably at a randomly selected post-intervention period.[5 8] An example of the design is depicted below for a time series of 18 periods:

$O_1\ O_2\ O_3\ O_4\ O_5\ O_6\ \mathbf{X}\ O_7\ O_8\ O_9\ O_{10}\ O_{11}\ O_{12}\ \mathbf{R}\ O_{13}\ O_{14}\ O_{15}\ O_{16}\ O_{17}\ O_{18}$

Again, "X" indicates the timing of the implementation of the quality improvement intervention. "R" depicts the timing of the removal of the intervention. The design increases causal interpretability over a simple interrupted time series by providing the investigator with a more powerful test of both the presence and absence of the intervention.[5]

The interrupted time series design with a non-equivalent dependent variable can be depicted as follows:

Outcome measure: $O_1\ O_2\ O_3\ O_4\ O_5\ O_6\ \mathbf{X}\ O_7\ O_8\ O_9\ O_{10}\ O_{11}\ O_{12}\ O_{13}\ O_{14}\ O_{15}\ O_{16}\ O_{17}\ O_{18}$

Related measure: $O_1\ O_2\ O_3\ O_4\ O_5\ O_6\ \mathbf{X}\ O_7\ O_8\ O_9\ O_{10}\ O_{11}\ O_{12}\ O_{13}\ O_{14}\ O_{15}\ O_{16}\ O_{17}\ O_{18}$

This is a single group interrupted time series design because both series are data collected from the same unit of analysis—for example, floor, patient, hospital. The first series are data for the outcome measure expected to be affected by the quality improvement intervention. The second is a time series coeval with the first that normally varies in the same way as the outcome measure, but is not expected to change after the intervention. Thus, if the improvement effort works as designed, investigators would expect to see changes only in the post-intervention series of the outcome measure.

These variants of the interrupted times series design provide powerful alternatives to single

group pre-test/post-test designs. Given the near chronic paucity of resources available to provide care and conduct research, it is imperative that quality improvement projects are able to demonstrate their effectiveness.[1 2] It is not sufficient to show that measures changed for the better "following" interventions and then to assume that the change was caused by the intervention, *post hoc*, *ergo proptor hoc*. The presence of favourable post-intervention changes in outcome measures may lead an organisation to believe that the quality improvement effort was successful, regardless of the internal validity of the evaluation design, and to continue to expend resources on the "successful" quality improvement programme. As long as outcomes are favourable, the inference that the improvement effort worked will have no real consequences. If outcomes deteriorate, however, no mechanism would exist to identify the causes of the deterioration since the original evaluation design was ill suited to determine whether the intervention initially worked.

Evaluation designs for quality improvement projects should be constructed to provide a reasonable opportunity, given available time and resources, for causal interpretation of the results. Evaluators of quality improvement initiatives may infrequently have access to randomised designs. Nonetheless, as we have shown here, other very rigorous research designs are available for improving causal interpretability. Unilateral methodological surrender need not be the only alternative to randomised experiments.

1 Smith S, Freeland M, Heffler S, *et al*. The next ten years of health spending: what does the future hold? The Health Expenditures Projection Team. *Health Affairs* 1998;**17**:128–40.
2 Iglehart J. The American health care system—expenditures. *N Engl J Med* 1999;**340**:70–6.
3 Kuttner R. The American health care system—employer-sponsored health coverage. *N Engl J Med* 1999;**340**:248–52.
4 Kuttner R. The American health care system—Wall Street and health care. *N Engl J Med* 1999;**340**:664–8.
5 Cook TD, Campbell DT. *Quasi-experimentation: design and analysis issues for field settings*. Boston: Houghton Mifflin, 1979.
6 Campbell DT, Stanley JC. *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin, 1963.
7 Meinert C. *Clinical trials: design, conduct and analysis*. New York: Oxford University Press, 1986.
8 Mohr L. *Impact analysis for program evaluation*. Chicago: The Dorsey Press, 1988.
9 Tukey J. *Exploratory data analysis*. Reading, MA: Addison-Wesley, 1977.
10 Box GEP, Tiao GC. Interventions analysis with applications to economic and environmental problems. *J Am Stat Assoc* 1975;**70**:70–9.
11 Pankratz A. *Forecasting with dynamic regression models*. New York: John Wiley and Sons, 1983.
12 Bodenheimer T. The American health care system: the movement for improved quality in health care. *N Engl J Med* 1999;**340**:488–92.
13 Jencks SF, Wilensky GR. The health care quality improvement initiative: a new approach to quality assurance in Medicare. *JAMA* 1992;**268**:900–3.
14 New Jersey Peer Review Organization (PRO). *Reducing the use of combined right/left heart catheterisation in Medicare patients with uncomplicated coronary heart disease*. New Jersey: PRO of New Jersey, 1997.
15 Pepine CJ, Allen HD, Bashore TM, *et al*. American College of Cardiology/American Heart Association guidelines for cardiac catheterization and cardiac catheterization laboratories. Ad Hoc Task Force on Cardiac Catheterization. *Circulation* 1991;**84**:2213–47.
16 Ljung GM, Box GEP. On a measure of lack of fit in time series models. *Biometrika* 1978;**65**:297–303.
17 Agency for Health Care Policy and Research/ National Heart, Lung and Blood Institute. *Unstable angina: diagnosis and management. Clinical practice guideline*. Washington: Public Health Service, 1994.