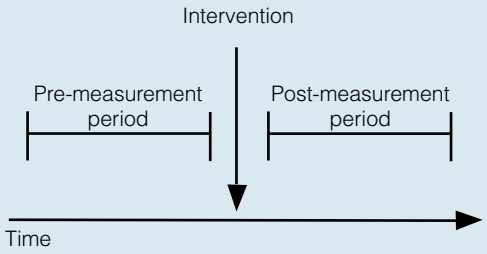
A Primer on Before–After Studies: Evaluating a Report of a "Successful" Intervention

It can be difficult to rigorously evaluate a clinical management or quality improvement intervention. Because these interventions generally occur at a system level (i.e., throughout the clinic, the hospital, or the health plan), it may not be practical to obtain suitable concurrent controls (clinics, hospitals, or plans not exposed to the intervention). As illustrated below, a common approach is to measure outcomes before the intervention is implemented and compare them with outcomes measured afterward—an approach often called a *before–after study* (or a *pre–post study*).



Although academics can easily criticize the lack of a concurrent control group, managers still need to make decisions on the basis of data available to them. This primer is intended to provide guidance on how to think critically about a report of a "successful" intervention obtained from a before–after study.

As with any report of "success," readers should start by asking three questions: Is the outcome unimportant? Is the magnitude of the change trivial? Were critical outcomes ignored? If the reader is comfortable that the answer to each is no, then he or she must go on to challenge the fundamental inference: that the "success" is a consequence of the intervention. The validity of this inference is threatened with an affirmative response to any of the following questions:

Would all participants in the "before group" be eligible for the "after group"? A typical before–after study compares the outcomes of hospitalized patients before and after some system intervention. Thus, different patients are often involved (e.g., patients admitted with pneumonia in June are compared with patients admitted with pneumonia in July). If only certain patients are eligible for the intervention, however, an inference about the success of the intervention can be seriously flawed. Consider a study of the effect an outpatient low-molecular-weight heparin program (which, by necessity, excludes the sickest patients) on the average length of stay of patients with deep venous thrombosis (DVT). A comparison of cost between all patients who have DVT (before) and patients who have DVT and are eligible for the outpatient program (after) would dramatically overestimate the effect of the intervention. The best estimate of the intervention's effect would be to compare all patients with DVT (before) with all patients with DVT (after), including both those who are eligible and those who are ineligible for the program. The comparability of patients in the before group and the after group is particularly relevant in assessments of the effect of guidelines (which generally apply to select patient subgroups).

Is there evidence for a prevailing "temporal trend"? Many outcomes change over time, regardless of whether an intervention has been applied. Consider a before–after study testing an intervention to reduce length of stay in the hospital. The average length of stay is 5 days before the introduction of the intervention but is 4.7 days after introduction. It is tempting to believe that the intervention caused the change. On the other hand, there is a prevailing temporal trend: Length of stay has been decreasing everywhere across time (at least until recently). The same problem would arise in a before–after study that tested an intervention to increase the use of aspirin in patients who have had a myocardial infarction. It would be difficult to untangle whether the observed change is the result of the intervention or dramatic television advertising. Because many forces are likely to be acting on outcomes that people care about, it is important to question whether an intervention is truly responsible for "success," particularly if outcomes are improving everywhere.

Were study participants selected because they were "outliers"? Understandably, some before–after studies target "problem areas" and select persons who are "outliers"—that is, participants who have extreme values in some measure. These studies may follow the same participants over time and face another threat to validity: regression to the mean. Examples could include a study of case management in patients who have had high utilization in the past or a study of an intensive communication tutorial in physicians who have been judged by their patients to have poor communication skills. Even if there is no intervention, participants selected because of extreme values will, on average, be found to have less extreme values with repeated measurement. Extremely high utilization in 1 year tends not to be so high the next (some patients may have had a major heart attack, stroke, or other catastrophic event that does not occur again in the next year); a group of physicians with extremely poor communication skills will tend to improve (some may have had a personal crisis that resolves in the ensuing year). Note that in neither case are the participants expected to return to the mean; they just become less extreme. Regression to the mean sets the stage to ascribe changes to a case management program or a communication tutorial when they actually represent the natural course of events.

Although it is always possible that a change observed in a before–after study is a consequence of the intervention, affirmative responses to any of the preceding questions make the inference more tenuous. Alternatively, the inference is strengthened when investigators paid careful attention to the comparability of the participants. Inferences are further strengthened when the observed change is substantial, unique, and occurs quickly after the intervention—in other words, when it is difficult to ascribe the finding to temporal trends. The confusing effect of regression to the mean can be avoided if participants are not selected because they are outliers. Nonetheless, inferences from a before–after study should be seen as being based on circumstantial evidence. If the accuracy of the inference is important, readers and researchers alike must ask whether there is a reasonable opportunity to test the intervention by using concurrent controls.