Q Manage Health Care
Vol. 13, No. 1, pp. 17–32

Study Designs for PDSA Quality Improvement Research

Theodore Speroff, PhD; Gerald T. O'Connor, PhD, DSc

Objective: *The purpose of this article is to discuss strengths and weaknesses of quasi-experimental designs used in health care quality improvement research. The target groups for this article are investigators in plan-do-study-act (PDSA) quality improvement initiatives who wish to improve the rigor of their methodology and publish their work and reviewers who evaluate the quality of research proposals or published work.*
Summary: *A primary purpose of PDSA quality improvement research is to establish a functional relationship between process changes in systems of health care and variation in outcomes. The time series design is the fundamental paradigm for demonstrating such functional relationships. The rigor of a PDSA quality improvement study design is strengthened using replication schemes and research methodology to address extraneous factors that weaken validity of observational studies.*
Conclusion: *The design of PDSA quality improvement research should follow from the purpose and context of the project. Improving the rigor of the quality improvement literature will build a stronger foundation and more convincing justification for the study and practice of quality improvement in health care.*

Key words: *quality improvement, quasi-experimental design, research methodology*

The scientific method is a process of proposing a study, designing an experiment to collect evidence, arranging the observations to test the hypothesis, and interpreting the results. The model of plan-do-study-act (PDSA) quality improvement uses the scientific method to answer "How will we know that a change is an improvement?"[1,2] (Figure 1). The PDSA model advocates the formation of a hypothesis for improvement (Plan), a study protocol with collection of data (Do), analysis and interpretation of the results (Study), and the iteration for what to do next (Act). Conducting a scientific study is the core concept of PDSA quality improvement that is fundamental to iterative learning.[3–6] Deming recommends that organizations adopt the real-time use of the scientific method, as multiple studies accumulate over time, an organization will then develop a profound knowledge about achieving quality.[3]

The core objective in PDSA quality improvement research is to assess whether a study intervention imposed to change a process produces an improvement in outcome. Like all scientific studies, PDSA quality improvement research can be viewed as probes for knowledge that involve testing interventions by manipulating variables and observing the effect upon

From the Departments of Medicine and Preventive Medicine, Center for Health Services Research, Vanderbilt University Medical Center, Nashville, Tenn (Dr Speroff); and the Departments of Medicine and Community and Family Medicine, Center for the Evaluative Clinical Sciences, Dartmouth Medical School, Hanover, NH (Dr O'Connor).

Corresponding author: Ted Speroff, PhD, Center for Clinical Improvement S1124MCN, 1161 S 21st Ave, Vanderbilt University Medical Center, Nashville TN 37232 (e-mail: ted.speroff@vanderbilt.edu).

Supported by a conference grant from the Agency for Healthcare Research and Quality (1R13HS1008601).

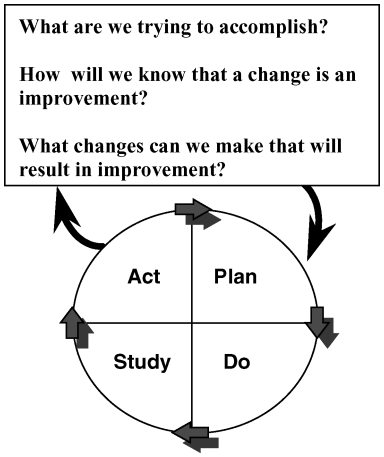

Figure 1. A formal model for PDSA clinical improvement.

other variables. Knowledge derived from this process of science is predicated on research methodology, which is the skill of putting the scientific method into practice. In 1963, Campbell and Stanley published a classic handbook on research methodology that still provides a solid framework for evaluating the validity and generalizability of scientific evidence.[7,8]

The Campbell and Stanley framework classifies research design as preexperimental, experimental, and quasi-experimental based on the basic principles of valid interpretation and replication of findings. Preexperimental designs are the clinical case study, the pretest-posttest design and the static-group comparison design.[7] All the preexperimental designs are weak examples of research because they are subject to extraneous factors that interfere with the conclusion or inferences to be drawn (Table 1). A study free of these sources of error has internal validity because the rival explanations are ruled out as not plausible.

The case study is a report written or told with great detail and precision to illuminate on an interesting case. The report is anecdotal and implies an implicit comparison to an expectation of what would have been had an intervention, exposure, or treatment not occurred or had circumstances been common, typical, or status quo. From case studies, we learn about new phenomena, paradoxes, techniques, and applications. Logical analysis is used to extract principles for case-based learning or to formulate hypotheses from qualitative examples. The problem, of course, is that rival sources of explanation are numerous and uncontrolled. Although the case report makes an interesting, compelling story and communicates powerfully a lesson to remember, it is weak science because there are no controls over how and when the intervention was applied.

The pretest-posttest design (also called the before-after study) is the comparison of observations at baseline to observations that occur after intervention. A difference across 2 points in time is attributed to intervention. The flaw in before-after studies is post-hoc reasoning that jumps to conclusions based on the error in logic called ergo propter hoc (after this, therefore, because of this).[9] Two measurements at single points in time are not sufficient to unequivocally rule out pre-existing trends, regression toward the mean, extraneous factors associated in time, chance fluctuation, or placebo effect.[7,10]

The static-group design includes 1 group that has experienced an intervention and 1 that has not experienced the intervention; the 2 groups are compared on observations collected only following the intervention period. Without knowing baseline equivalence, the groups' differential susceptibility to the experience or self-selection to an exposure confounds and precludes a fair comparison.[7,9]

The experimental design arose to control all the factors that potentially jeopardize internal validity.[7] In the randomized controlled trial, subjects are randomly allocated to the intervention and control groups to promote the equivalence of experience between the 2 groups. If large numbers of subjects are randomly assigned, a balanced representation permits interpretation of the difference between the intervention and control group to the effects of the intervention. In addition, investigators implement rigorous methodological procedures to secure maximum internal validity. A consequence is that the randomized controlled trial may not be applicable to the subjects, environment, or circumstances outside the tight control of the experiment.

Table 1

EXTRANEOUS FACTORS THAT WEAKEN STUDY VALIDITY

Threats to validity	Description of sources of error
History	Extraneous events that occur over time and between measurements affect the subject's responsiveness. Examples include seasonal variation, special events and ecological occurrences within the surroundings of the study.
Maturation	Changes and growth inherent to the respondents (unit of observation) that occur with passage of time. Examples include fatigue, boredom, needs, spontaneous remission and learning processes.
Testing	The effects or learning that take place with the experience of measurement that carry over to downstream measurements. Examples include practice effects, familiarity with testing, comfort with the test situation, rumination over expectations, and debriefing that occurs with other participants.
Instrumentation	Changes in the calibration of instruments (decay in mechanical parts) or observers (changes in performance) that occur over time. Examples include fatigue of data collectors, breaking-in or learning effects in grading or scoring, drift or a change in criteria during observation over time, observer familiarity with procedure or respondents, or knowledge bias resulting from unblinded procedures.
Statistical regression	Regression-to-the-mean operates when selection of subjects is based on extreme scores. Change from pre to post testing is related to correlation between measures, chance variability and error of measurement. Regression toward the mean is an artifact recognized when those who initially scored the lowest have gained the most during post-testing or vice versa.
Selection bias	Bias occurs when subjects are differentially recruited into comparison groups based on characteristics or prognostic factors that interact with the intervention. Sampling procedures using volunteerism, captive or conveniently available audiences, and self-selection are prone to selection bias. Random allocation to comparison groups avoids selection bias.
Drop out	Differential loss of respondents or drop out between comparison groups can result in bias when comparing outcomes. Intent to treat analysis is performed to lessen the effect of bias.
Reactivity to testing or measurement (external validity)	Subjects can experience heightened sensitivity to intervention due to measurement or awareness of participation (e.g., Hawthorne effect). The process of measuring may change what is being measured. Behavior may change when people realize they are being monitored. Testing or observing becomes a stimulus to change and is not just a passive recording of behavior.
Multiple treatment interference	This category includes all antecedent contributing factors that carry-over to the study. Experience with similar prior treatments can effect the response to the current study intervention. Characteristics of the subject's environment, community or setting that are similar to the context of the study influence the familiarity and comfort to the study situation. All the socio-demographic and behavioral characteristics, readiness for change and preferences of the subjects are cofactors that influence outcomes. These factors are controlled by defining the target population based on eligibility and exclusion criteria. Random sampling from the target population promotes generalization and representativeness of results.

Modified from Campbell and Stanley.[7]

Research designs that lack random allocation but address issues of internal validity are quasi-experimental. The seminal work by Campbell and Stanley was to encourage use of relevant and robust quasi-experimental designs that secure adequate and proper results tempered by knowing which competing interpretations the designs fail to control.[7] Rigorous observational studies are often valid and replicate results found in randomized controlled trials.[11] Because every study is merely a probe for knowledge and not a proof of fact, a study requires scrutiny on its probing value, that is, the contributions of the study relative to its limitations. Campbell and Stanley remind us that quasi-experimental designs are able to overcome internal validity issues of pre-experimental designs by replication schemes that lessen the plausibility of rival hypotheses due to uncontrolled factors or historical effects.[7]

On the one hand, quality improvement is typically conducted in settings where random allocation to groups is often not feasible for ethical reasons or where the environmental condition prevents experimenter manipulation. On the other hand, case reports and before-after studies are common in the health care quality improvement literature.[10,12] One reason that health care quality improvement has gravitated to the case report and the before-after design is to quickly disseminate stories of success where a change in behavior led to a change in outcome. Weak designs, however, do not allow internally valid conclusions and the consequence is that the science of quality improvement will proceed ineffectively in health care. PDSA quality improvement research should avoid these pragmatic shortcuts in order to demonstrate clearly that it was the strategic maneuver that caused improvement. The purpose of this article is to advocate for the use of quasi-experimental strategies to improve the scientific foundation of PDSA quality improvement in health care. The issues essential to this purpose are to distinguish the type of knowledge pursued by PDSA quality improvement research and to map the quasi-experimental research designs appropriate to enhancing the rigor of PDSA quality improvement studies.

A PERSPECTIVE ON PDSA QUALITY IMPROVEMENT RESEARCH

The aim of PDSA quality improvement is to pursue effective changes in process that favorably affect outcomes.[3,13] Several characteristics distinguish the discipline of quality improvement. Quality improvement operates with a systems point of view. A system is an organizational structure of interrelated processes operating to produce an output. In industry, we think of a system as the fabrication and assembly operations to make products. A system, however, may comprise a mass of interactions involving mental activity, attitudes, and behaviors of a single human or among a group of people orchestrated to produce services and outcomes.[14] Quality improvement involves understanding how the system works and sharing that knowledge within the organization.[3,15] Thus, the focus of intervention in PDSA quality improvement is on a system of production that predicts the sources of variation of its output.

The Dartmouth clinical improvement model is a schematic process-outcomes representation of the systems point of view in health care quality improvement.[16–18] In this schematic, patients access a system of interrelated human and technology processes that provide diagnosis, treatment, and follow-up to alter the patient's health. Outcomes are measured using targets selected for a value compass.[19] To show that health outcomes following changes to a system of care are better than the outcomes prior to the change requires tracking of health status and clinical outcomes. The purpose of PDSA quality improvement research is to clearly establish the functional or causal relationship between changes in behavior (interventions on system performance) and impact on outcome, that is, the direct relationship between process changes and variation in outcome. The type of knowledge pursued by PDSA quality improvement research is how to improve the behaviors and capabilities of the process that affect the end product.

Total quality management deals with organizational management, teamwork, systems thinking,

understanding process, and the psychology of change to create an environment for improvement.[3,20–22] Health services management and research on organizational practice typically starts by understanding the pathway of these processes using a variety of qualitative, case analysis tools to identify the set of key conditions and causes which work together.[14,23] Outcomes are considered a function of the processes that comprise the system and every key step of the process is made explicit and examined for its contribution to the desired output. Creative thinking, innovations, and change ideas are solicited to either change or create new structure or modify the existing behaviors that drive the processes to enable better outcomes.[1] These aspects of the quality movement are necessary, but not sufficient, to qualify as PDSA quality improvement. PDSA quality improvement is defined when data are collected to demonstrate that change by intervention resulted in improvement. From the findings of PDSA quality improvement studies, we gain information about system variables and their importance to affecting outcomes.[1,14] From this perspective, PDSA quality improvement requires research methodology to test hypotheses that are derived from theory and to change ideas about improving a process.

The systems approach requires that quality improvement direct intervention at the organizational unit responsible for the processes that produce the outcome of interest. Although outcomes are frequently measured at the individual product or patient level, the intervention is aimed at the organizational unit or microsystem that operationally defines and distinguishes the process of treatment.[15,18,24] Redesign of structure or changes in behavior are meant to improve the human or mechanical production or service which influence the quality of outcomes. Thus, the targets for quality improvement can range from practice management of a clinic, emergency department, operating room, or hospital to clinical management of a single patient or group of patients with a certain disease or need for service. Hence, it is common for PDSA quality improvement studies to have a sample size of 1. The objective of such PDSA quality improvement studies is to show that behavioral or organizational change is functionally related to an improved outcome in the individual case. This mandate for quality improvement to make statements about the individual brings research and practice closer together. Whether the individual case is a single patient or an organization of care such as a clinic or operating room, we will refer to the object receiving intervention as the subject.

Various disciplines have a history of attempting to establish functional relationships between process change and outcome improvement that we designate as PDSA quality improvement research. The present manuscript integrates concepts from behavioral psychology,[22,25–29] single-case methodology,[30–34] and statistical process control.[1,14,35–40] Common to these methods are carefully constructed measures of responses administered repeatedly and regularly over time in an individual to discover the effectiveness of a change idea in particular circumstances. In the following sections, we will present scenarios that illustrate the major quasi-experimental designs that are relevant to PDSA quality improvement research (see Table 2) and factors pertinent to their limitations, addressing both single-case and comparative-group applications.

Table 2

STUDY DESIGNS FOR PDSA QUALITY IMPROVEMENT RESEARCH

Study design	Characteristic	Primary concern
Time-series (AB)	Continuous, longitudinal data	Historical control
Equivalent time-series (ABAB)	Replication	Carry over effects
Multiple baseline (AAAB, AABB, ABBB)	Lagging of interventions	Contamination
Factorial	Experimental design	Confounding

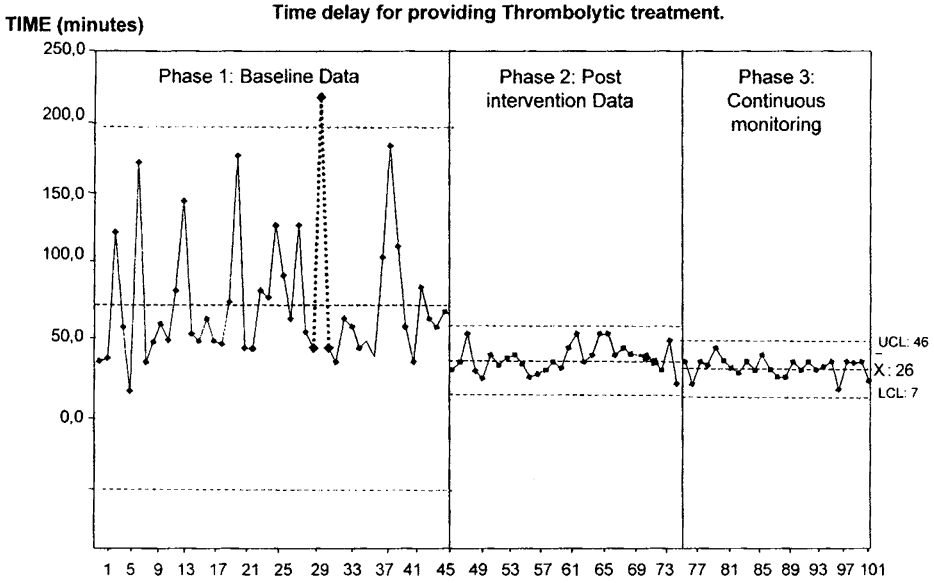

Figure 2. Reducing time delay in the thrombolysis of MI (Reproduced with permission from American Journal of Medical Quality).[41]

TIME SERIES (AB) DESIGN

Saturno et al[41] conducted a study to improve thrombolytic therapy in acute myocardial infarction by reducing the time from the arrival at the emergency department to the provision of the thrombolysis (door-to-needle time). Treatment times for 46 patients comprise the baseline period (see Figure 2) and averaged 72 minutes. Intervention comprised guidelines for the management of acute myocardial infarction patients. Times for the next 56 postintervention patients are shown in Figure 2, the average time improved to 26 minutes. Additional examples of similar published studies include optimizing inhaled corticosteroids or environmental controls to improve peak expiratory flow rates in asthma patients,[42,43] solving, and maintaining safety issues to decrease needlestick injuries[44] and re-engineering physician office tasks to improve productivity and revenue.[29] These studies are examples of the AB design (also called the time-series design).[7,30–33] The architecture of this design consists of a baseline period (phase A) followed by the introduction of intervention or treatment (phase B). The AB design is the most elementary and popular version of single-case design and contains several principles fundamental to the single-case methodology (Table 3) that overcome the pitfalls of the before-after design.[7]

Single-case designs rely on continuous or repeated observations over time. The basic comparison is to demonstrate a change from baseline. Rejection of the null hypothesis requires a discontinuity in results between phases A and B that corresponds with onset of intervention. The assessment requires confidence in the use of baseline as an historical control. A feature critical in statistical process control is that a process undisturbed by extraneous causes tends to repeat itself.[14] A stable process implies that the variation, and central tendency in outcomes will remain predictable within statistically controlled limits unless a fundamental change is made to the factors that control the process. Failure to demonstrate baseline stability is a sign that special causes or unusual variables are exerting influence that can alter the variation and predictability of outcomes. The steps in statistical process control are to examine data for special sources of variation that cause a process to be unstable, segment special from random variation,

Table 3

PRINCIPLES FUNDAMENTAL TO SINGLE-CASE METHODOLOGY

Study characteristic	Principle
Continuous assessment	A reliance on repeated observations of performance over time is needed to make the comparison of interest. The outcome is examined over time to determine whether changes in behavior coincide with the intervention.
Baseline assessment	The pattern and stability of performance at baseline serves as the criterion to evaluate whether the intervention leads to change. If treatment is effective, future performance will differ from the level predicted by repeated observations at baseline.
Stability of performance	Confidence in projecting baseline requires absence of a trend in the data and relatively little variability in performance.
Variability in the data	Variability is fluctuation in performance over time. Wider variation makes it more difficult to draw conclusions about the effects of intervention.

and establish stability in random fluctuation (also called common cause variation) prior to PDSA quality improvement.[14,35,38,40] Constancy or stability during baseline is required to rule out extraneous trends and to document convincingly that intervention is superior to the processes previously responsible for performance.

Once baseline conditions are established for stability, the baseline performance is statistically compared as the reference set to treatment performance for changes in slope (the rate or trend of change) and/or change in level (the incremental gain) related to the timing of intervention.[1,26,27,29] If intervention is effective, performance will differ from the projected level extrapolated from the stable baseline, that is, a large effect on the process alters the pattern of results. Changes that occur immediately after onset of treatment and that have a marked shift in level of performance are more easily attributed to the special introduction of an intervention. When the pattern of results is not dramatic, the influence of alternative explanations is more plausible. Changes that are delayed and take place well after treatment has been applied are more difficult to interpret because of intervening experiences. Furthermore, continuous application of the intervention should produce a stable but altered course of performance and not result in a temporary or transient effect. Thus, we apply the criteria of immediacy, magnitude, and stability to the pattern of results to determine the effectiveness of intervention.

Statistical process control provides relatively uncomplicated statistical rules for evaluating whether the pattern of observations from baseline through intervention differs from the random fluctuation that underlies the natural history of the baseline process.[14,35,39,40] For example, the statistical process control analysis of the thrombolysis guidelines derived the statistical control limits shown in Figure 2.[41] Rules for testing significance provide the statistical argument for an immediate improvement in door-to-needle time attributable to guideline implementation. Had a downward trend been present during baseline, outcomes following intervention may be mere continuation of the trend, making it difficult for Saturno et al to attribute improvement to the guideline intervention. Even if baseline reveals a trend, there are other sophisticated analysis techniques that can test for changes in slope and discontinuity in performance.[45] The disadvantage is that as many as 50 data points may be needed for continuity-discontinuity time-series analysis. The preference in PDSA quality improvement is to establish stability at baseline and use statistical process control methods

to lend support to the idea that intervention is responsible for driving outcomes in the sought-after direction.

An important premise underlying analysis is that measurement is adequate and that observation procedures are reliable.[33] The choice of the target behavior, selection of the intervention tactics and evaluation of treatment effectiveness must be in alignment so that the aim of the PDSA quality improvement initiative, the operational details, and proficiency in carrying out the work focus on the pertinent question. The measurement process should be unobtrusive and procedures embedded as a natural part of the quality improvement environment. In Saturno et al,[41] times were already routinely collected as part of the admission to the emergency department, the electrocardiogram printouts, and medical record documentation for administration of medication. These data were abstracted and entered into the control chart for analysis of changes in door-to-needle time. This methodologic feature of the Saturno study was important because measurement is not only data collection but also an intervention technique. In quality improvement, participants are often the observers. Unblinded, self-monitoring is vulnerable to measurement bias. The potential for observers to become subjectively influenced by participating in their own study requires standardization of the self-monitoring procedures, rigorous training of the observers, reliability checks, correspondence of subjective evaluation with objective measures, and when possible, blindness to the hypotheses. Objective results in PDSA quality improvement are more easily achieved by improving the measurement process and using unobtrusive measures than by blinding procedures.

The salience of measurement, especially self-monitoring, makes before-after designs particularly vulnerable to reactivity and Hawthorne effects (Table 1). The repeated measurement of the time series design monitors these placebo effects during baseline and allows the participants to acclimate to data collection. Allowing the subject to adjust and habituate to data monitoring eliminates reactivity as a viable alternative explanation to the baseline-treatment comparison.[1] Repeated observations also provide a check on whether observers change the manner in which they apply definitions or instrumentation over time. Thus, baseline measurement using continuous, repeated observation and application of control charts provides the opportunity to control measurement error, methods bias, and detection bias[9] so they will not interfere with the conclusions that the investigator would like to draw. Continuous assessment during post intervention for constancy of effect also serves as a check between real and random changes. Showing a stable change from baseline to intervention excludes statistical regression toward the mean as an alternative explanation. All this measurement, of course, requires standardization and process indicators to substantiate that the intervention is carried out as intended. This implies careful and precise description of the interventions applied and a clear statement of the question we want to answer.

The biggest weakness of the time-series design is failure to control history. The major caveat is that performance might change even without intervention or for reasons other than the intervention. Ancillary conditions present at baseline may spill over to the treatment period or extraneous factors in the background but concurrent with intervention may contaminate the treatment period and confound the study results.[7] Critical appraisal of methodology teases out potential alternative explanations.[1,33] For example, the baseline and treatment periods for monitoring asthma might capture different segments of a curve that reflects seasonal variation. The methodologic solution is to be aware of extraneous factors such as cyclic variations and to balance their effects across the baseline and intervention phases. A statistical solution is to adjust for seasonal effects.[40,46] Because quality improvement is conducted in the field, the subject is exposed to surrounding conditions that may occur in parallel with the intervention. For example, a Center for Disease Control report on needlestick transmission of infectious disease might increase motivation for careful handling of needles and confound interpretation of quality improvement results. In addition, treatments that require a delay before appearance of effect become

confounded by maturation and extraneous events that occur following baseline. Thus, interventions with a long latency are poor candidates for a time-series study.

One variation on the time-series design is to add a no treatment control group (AA) to the time series (AB) design. In addition to the within-subject comparison between the intervention (B) to baseline (A) of the time-series design, the multiple time series design compares the intervention group to a control group.

In quality improvement, control charts are often fed back to the team that manages the process. Data feedback is itself an intervention that can enhance motivation for change. Comparison between control and intervention conditions that both receive data feedback can further differentiate intervention from these spurious effects. However, a multiple time series design with a comparison between subjects makes the study vulnerable to selection bias (Table 2). In clinical epidemiology, multivariate statistical techniques are used to adjust for confounding variables. In quality improvement, Alemi et al have introduced the concept of control charts adjusted for covariates.[47] Similar to cohort studies, the multiple time series design must confront methodological limitation due to selection bias and establish risk adjustment methodology. The use of risk-adjusted control charts would allow the investigator to appropriately account for changes in the patient case mix over time and differences between groups.

Another variation on the time-series is the changing criterion design which modifies the intervention phase of the AB design. This design is applicable when intervention can be incrementally modified or titrated to produce a dose-response on outcome. A performance criterion is set to signal success of intervention. Consider a scenario where the hypothesis is that spread and compliance with computerized order entry will decrease medication errors and the quality improvement team sets a stretch goal for the clinicians of the institution to reach benchmark performance. With the changing criterion design, the strategy is to set an initial target, say at 50% of benchmark performance. Rate of success is fed back to the clinicians, for example, by posting a control chart. After performance stabilizes and consistently meets criterion, the level for success is made more stringent, say 60% of benchmark performance. Several iterations occur within the intervention phase until the clinicians reach benchmark performance levels. The purpose is to show that increments in outcome coincide with shifts made in criterion. If a functional relationship between process and outcome exists, then better, more frequent or consistent application of the intervention (computerized order entry) has sufficient control over the process to shape performance to reach higher standards.

When performance does not closely correspond to changes in the criterion, the influence of the intervention is difficult to detect and improvements may be attributed to extraneous factors. A rapid change in performance that exceeds the criterion raises the possibility that extraneous factors coincided with the onset of intervention. In summary, the investigator looks for a close correspondence where performance follows shifts in the criterion over the course of intervention, uses annotated control charts to indicate timing of interventions, and statistically tests the changes in performance. This variation of the time series design is consistent with the quality improvement strategy of continuous improvement cycles to implement incremental process improvement.

Protection of the AB design against spurious results depends on the rigor of the research methodology in eliminating as many potentially confounding sources of variability as possible so that the functional relationship between process changes and outcome can be determined with more precision.

Despite the improvements over the pre-post test design, the time-series design does not allow for an unequivocal demonstration of the controlling effects of the intervention. In clinical epidemiology, a test for false positive results is replication because it is extremely unlikely that extraneous, arbitrary events would consistently covary with new systematically repeated applications of treatment. In similar fashion, enhancements to the time-series design use the principle of replication to overcome its issues with internal validity.

THE EQUIVALENT TIME SERIES (ABAB) DESIGN

The ABAB design replicates the time-series design. The A and B phases of the AB design are repeated to yield 4 phases of time-series data. We expect performance to differ sharply as a function of the different conditions, producing a systematic pattern across baseline and treatment conditions. The effects of quality improvement are clear if performance improves during the first intervention phase (B), reverts or approaches baseline when intervention is withdrawn (replicated A phase) and improves when treatment is reinstated (replicated B phase, see Figure 3). The second A phase tests the projection of the first baseline and resolves concern over use of a historical control. If extraneous factors are involved, the pattern of response will reveal a gradual improvement over time, including during the second baseline (replicated A phase). The replication of the difference between predicted (replicated baseline A) and obtained (replicated treatment B) levels of performance reinforces generalization across time and provides a direct demonstration that intervention effects are responsible for the change in outcomes. Absence of replication indicates that (1) extraneous factors led to the change during the first AB intervention phase, (2) the intervention may have carryover effects that require sufficient time to washout, or (3) the subject was permanently altered by structure or learning. The intervention that changes process must be reversible so that a return to baseline means the underlying process is able to return to the same previous stable condition. The ABAB design is not suitable when the investigator expects carryover or learning effects that prevent reversal to prior conditions or if reversing to baseline is not desirable or ethical.

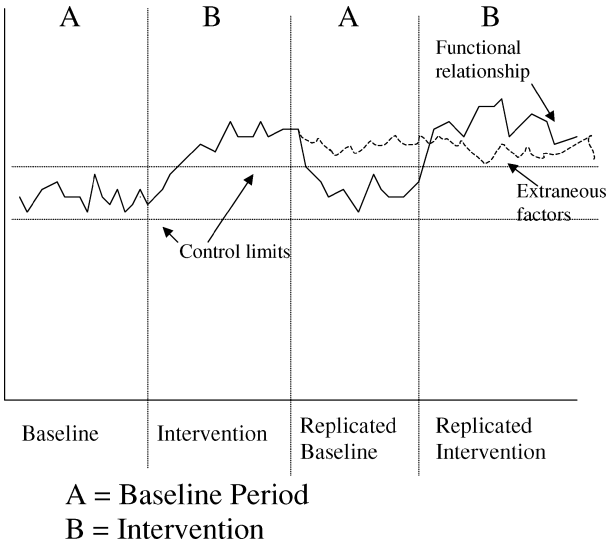

Figure 3. Equivalent times series (ABAB) pattern of response.

Use of the replication strategy in quality improvement research is the counterpart to a control group in factorial design. The replication strategy renders history as an unlikely explanation but elicits vulnerability to several other sources of bias. The alternation of conditions might coincide with cyclic extraneous variables that produce a time dependency for the ABAB pattern of response. This confounder may be ruled out by representative sampling of baseline and intervention conditions across time. On the other hand, knowledge or perception of when the B condition is in force may produce reactivity to anticipated intervention. Spurious effects generated by awareness are usually short-term and dissipate over time. If participants are not blinded, a placebo effect may result from enthusiasm and subjective awareness that will confound actual treatment effects. At the very least, demonstration in the change in performance should be stable and maintained beyond transient effects.

An example of the equivalent time series design is an ongoing study at Vanderbilt University Medical Center to improve medical record documentation. Residents in general surgery proposed a template containing prompts for standardized and complete daily progress notes. The primary objective is accurate documentation of patient complications and comorbidities. The measurement indicator is the independent coding of diagnoses by the medical records personnel, who were not involved and were blind to the activities and communications on the ward. A time-series (AB) design compares the coding of complications from the baseline (A) period to an intervention (B) period when the structured progress note is used by the resident physicians. The duration of the baseline and intervention periods coincides with the

education rotation of residents to different service areas of training, yielding independent samples of data collection. Although the source of measurement is from many residents and patients, the unit of intervention is the general surgery service, thus this study uses single-case methodology. If use of the structured progress note makes documentation of complications more salient to the resident work process, we anticipate a shift in the rate of recording patient complications from the A to the B phase of the control chart (see Figure 3).

Personnel in medical records, however, are dependent on chart documentation. To obtain a reference standard of patient-based information, experienced nurse utilization managers on the general surgery service were requested to independently complete a structured progress note on all patients. This information is collected to show that the patient case-mix and severity is actually equivalent across the baseline and intervention periods. Now, however, statistically significant trend in documentation of complication rate from baseline A to intervention B might result from nurses and residents on the same ward discussing the patient. A second AB phase is used to test whether changes are due to contamination effects (extraneous factors) or caused by use of the structured note (strategic intervention). A return to baseline condition is induced by withdrawing use of the structured progress note by residents. If the control chart shows maintenance of the documentation rate during replicated baseline (A), the findings will support a contamination effect mediated by nurse-resident interaction. If the documentation rate reverts to the original baseline, the 2 baselines are equivalent and the findings do not support contamination. As a further test of transient false positive results and concerns over historical controls, reinstitution of the structured note (replicated B phase) tests the functional relationship between use of the structured note (change in process) and outcome. A replication in improvement in outcome establishes a convincing argument that use of the structured progress note by residents impacts documentation of patient complications.

A published example of the ABAB design is provided by Pfadt[26] in the evaluation of the clinical effect of behavioral treatment of individual clients. In the baseline (A) phase, reinforcement of target behavior was withheld and during intervention, reinforcement was contingent on performance. The pattern of response was similar to Figure 3 and statistical limits on the control chart demonstrated that behavior was a function of reinforcement contingencies.

More complex applications of the equivalent time series design have been used to analyze for effects of multiple treatments and their interactions.[30,32,48,49] For example, the AB AB AC AC A(BC) A(BC) design examines the functional relationship of B and C interventions on outcomes and the effects of the BC interaction. These more complicated designs are beyond the scope of the present article but are mentioned to acknowledge the utility of single-case methodology. A cardinal rule in single-case methodology is to change 1 variable at a time when proceeding from one phase to the next and consider any counterbalancing and randomization of treatment order that controls for potential confounding by spurious time dependencies. Equivalent time series designs are suitable when interventions bring about rapid effects with little or no carry-over when intervention is suspended. Ultimately, however, findings are specific to the single subject. Thus, the single-case methodologies are sometimes modified for multiple-subject and comparison group designs.

The caveat to the ABAB design is a carry-over effect that sustains the impact of intervention and prevents the return to baseline conditions. Interventions with carry over effects are those that induce changes in the structure of the organization, implement irreversible procedures or states of condition, or produce an unwillingness to revert to baseline conditions. In this situation, the procedure of AB replication is unwarranted and the multiple baseline design should be considered.

THE MULTIPLE BASELINE DESIGN

In this design, an intervention is replicated by segmenting its application within the same subject onto separate targets. For example, consider physician orders on the medical ward using automated computer entry in place of handwritten orders. The

intervention might first be implemented on only laboratory tests, then on imaging, and then pharmacy. The study is begun with establishing concurrent stable baselines for laboratory, imaging, and drug errors. The first AB design is applied when computer entry of laboratory tests replaces handwritten orders and baseline measures continue for imaging and pharmacy. After performance stabilizes for the laboratory errors, the computer entry intervention is applied to the imaging tests and baseline continues for the drug orders. After the second AAB design stabilizes, the computer entry intervention is applied to the drug orders (the AAAB design). The multiple baseline design is when intervention (B) is sequentially introduced to different baselines (A) at different points in time. As in the time-series design, there needs to be stability of the behaviors during the baseline phases. The time-lagged replication controls for history, avoids problems with carryover effects and attends to the importance of preserving gains (no reversals).

Figure 4 shows the anticipated staggered pattern for control chart results. Repeated demonstration of change when and only when the intervention is applied provides a convincing demonstration that intervention was responsible for change. The pattern

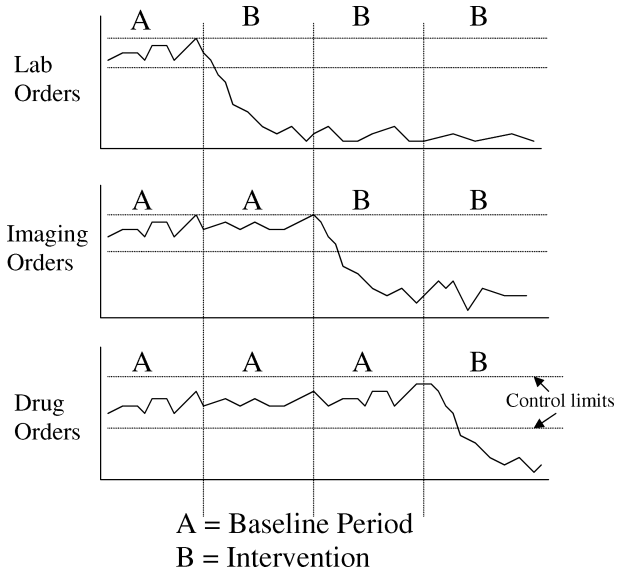

Figure 4. Multiple baseline pattern of response: within subject.

of results rules out performance bias, incidental or conjugate activities that could serve as cointerventions and coincidental historical events that could confound the interpretation of the times-series (AB) design. The strength of replication and criteria of immediacy, magnitude, and stability rule out alternative hypotheses. Methodologic problems occur if the target behaviors within the subject covary, the intervention effect from the (ABBB) branch will spill over to the lagged branches (AABB and AAAB) and contaminate the baseline periods.

A variation of the multiple baseline design is to implement an intervention on the same target behavior but replicate the time-series by segmenting the application across independent subjects or settings with staggered implementation. This is especially suited to the situation in which a particular process or set of behaviors in need of change is constant across different persons or settings. The design with replications staggered over time is consistent with the quality improvement strategy in which intervention is implemented on a small scale first before being deployed widely. An example is a recently funded project to evaluate ELSIE at Vanderbilt University Medical Center. ELSIE is a computer interface that connects computerized physician orders with nursing pathways of inpatient care. Nursing pathways are used to provide acute and disease management cues and prompts to the order entry program. The target behavior is congruence of patient orders to the pathway of care. Among 20 service areas in the medical center, 10 were randomly selected for implementation of ELSIE and paired to the remaining 10 which serve as control sites. The 10 intervention sites will have ELSIE implemented with multiple baselines as follows: site 1 (AB), site 2 (AAB), site 3 (AAAB), until all 10 sites have the integration between computerized orders and pathways. The remaining sites provide concurrent controls, a feature that elaborates on the standard multiple baseline design. The anticipated pattern of control chart results is shown in Figure 5 for the first 3 sites.

The methodologic concern in this application of multiple baseline is contamination or spill-over between settings that would produce vicarious effects

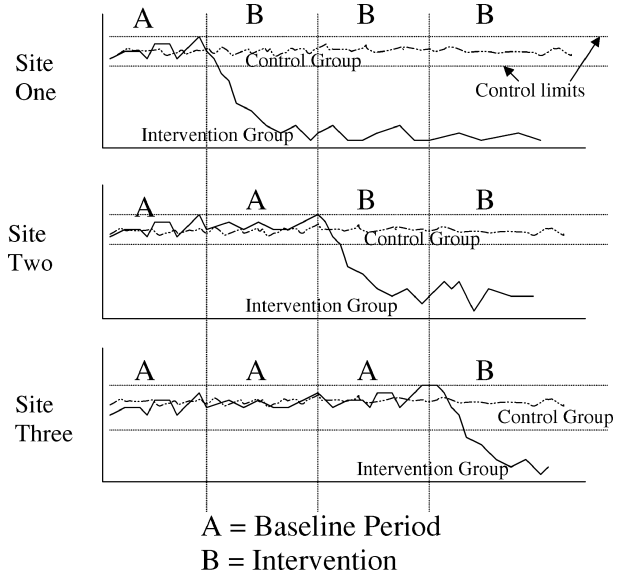

Figure 5. Multiple baseline pattern of response: across subjects.

and ambiguous results. The settings must remain independent and the pattern of results sustained across the multiple baselines. Another methodologic concern relates to prolonging baseline. Long lags due to postponement of intervention increases risk of drop out from the study. Prolonged baselines also open a study to threats of future competing opportunities that can extraneously influence behavior and impact on outcomes. The ELSIE multiple baseline has a 2-year rollout, therefore, concurrent controls were added to the multiple baseline design to address the possible confounding associated with prolonged baselines.

EXPERIMENTAL DESIGN

Single-subject randomized controlled trial

In 1988, Guyatt et al[50] published a "clinician's guide for conducting randomized trials in individual patients." These *N* of 1 RCTs are collaborative experiments between clinician and patient to evaluate the effectiveness of treatment. For example, a clinician and patient cooperated in designing a study on the effectiveness of amitriptyline hydrochoride in treating fibrositis.[51] The purpose of the RCT was not efficacy but whether therapy was warranted in the individual patient. A randomized sequence was arranged where the patient would either take the drug or a placebo. The drug has a rapid onset and stops acting soon after it is discontinued, allowing for a crossover of treatment in the individual patient. The clinician and physician were blind to treatment conditions but a pharmacist was available to help supply the capsules and to monitor compliance. Outcomes may include objective clinical measures but in this trial the patient rated fatigue, aches, pains and sleep disturbance. Among 14 of these *N* of 1 RCTs in fibrositis patients, 6 trials confirmed treatment effect. The *N* of 1 RCT strategies of repeated measures and replication to evaluate treatment of chronic, stable conditions are features similar to the ABAB design. Traditional statistical methods such as the paired-*t* test and sign test for runs have been used for analysis but statistical process control methods could easily be used to assess the *N* of 1 design.

Downward and upward phases that patients are likely to go through with their illness could create serial dependency in longitudinal measurement of outcome. The relationship from one observation to the next is called autocorrelation. Obtaining a series of measures from the same individual increases the likelihood of autocorrelated data. No evidence of autocorrelation has been observed in *N* of 1 trials[50,51] and in statistical process control, Wheeler has reported that control charts with autocorrelations below 0.7 are not severely affected.[40] However, the presence of strong autocorrelation among the repeated measures can have a large impact on traditional control charts[52,53]; the control limits are too narrow, producing false positive interpretations. In these cases, a times series model is fitted to the observations to adjust the limits of the control charts.[40,52,53] Statistical process control methods have been around since the 1930s and have been proven just as effective and grown just as sophisticated as classical statistics.[54]

Factorial

Although used rarely in health care, fractional factorial designs have been widely used in industry to study sources of variation. These planned

experiments with randomization are designed to identify important variables and investigate how their interactions affect outcomes. Fractional factorial designs are a proportion of a full factorial design that allows the size of an experiment to be kept practical while still enabling the estimation of important effects.[2] These methods are of particular interest when more than one change is implemented at the same time. Strategies are used to counterbalance treatments in a way that confounds higher order interaction effects to tease out contributing factors as primary effects.

In industries implementing fractional factorial planned experiments, many important interactions have been found and many are waiting to be found.[55] In addition, *N* of 1 studies can be sequentially arranged to build a fractional factorial to analyze across and between group sources of variance.[51] *N* of 1 and fractional factorial designs allow the investigators to learn as they go (PDSA cycles) and decide what should be done next for iterative scientific problem solving.[55] An application, for example, would be the investigation of an intervention program with multiple factors (such as, diet, education and screening) that might influence the improvement in the management of diabetic patients. To learn more about fractional factorial design of experiments see Sloan[56] and Moen et al.[2]

Several randomized control trials have been conducted where quality improvement has been compared to alternative interventions.[57–60] In Dietrich, O'Connor et al,[61] 98 ambulatory practices were randomized to office system, quality improvement intervention, educational intervention or the combination of education plus quality improvement. This study is an example of a complete 2 × 2 factorial experimental design that examines main and interaction effects. The results found that practices using quality improvement did better in providing cancer prevention and detection services.

CONCLUSION

PDSA quality improvement involves the systematic study of a process to discover assignable causes of variation and to solve problems.[14] PDSA quality improvement research is outcomes driven and outcomes that define value of care include cost, quality of life, clinical outcomes, and satisfaction. The science of quality improvement is intended to generate knowledge concerning prediction and control of a health care system to produce better outcomes. The challenge for health care PDSA quality improvement is to provide a thorough and specific test of change that satisfies scientific principles. The use of robust study designs with rigorous compliance to sound data collection and protocol implementation can provide understanding of how PDSA quality improvement interventions work in the real world setting. The extent to which randomization is necessary depends on the stability of the process. Stable or chronic processes are candidates for quasi-experimental designs and control charts for measuring intervention. When experimental designs must be considered for understanding the contributions of various aspects of a complex intervention, *N* of 1 RCTs and fractional factorial designs are practical techniques. Key features of these research designs are times series measurement, choosing an appropriate baseline, testing the stability of the baseline, use of replication and reversals, constancy of the treatment effect, and statistical techniques for evaluating effects. Use of more rigorous design for PDSA quality improvement research and the analytic technique of statistical process control will promote more convincing claims for the effectiveness of improving outcomes and a solid foundation for producing iterative learning.

The percentage of hospitals reporting continuous quality improvement initiatives was 69% in 1993 and climbed to 93% in 1998[62] but the lack of a robust, scientific literature has inhibited the acceptance of quality improvement methods among health professionals.[5,12] The fundamentals of quality improvement that call upon the rigor of the scientific method are highly compatible with the professional value of being persuaded through evidence. Use of robust study design increases the likelihood that the work will be published in peer reviewed journals. Greater dissemination in journals will help build

the scientific foundation that justifies the practice of quality improvement and reinforces a career activity in pursuit of improvement knowledge. Moreover, PDSA quality improvement research is a scientific paradigm that promotes merging of the clinical, operational, research, and educational disciplines within a health care enterprise that brings research and practice close together.

REFERENCES

1. Langley GJ, Nolan KM, Nolan TW, Norman CN, Provost LP. *The Improvement Guide: A Practical Approach to Enhancing Organizational Performance*. San Francisco: Jossey-Bass; 1996.
2. Moen RD, Nolan TW, Provost LP. *Quality Improvement Through Planned Experimentation*. 2nd ed. New York: McGraw-Hill; 1999.
3. Deming W. *Out of the Crisis*. Cambridge, Mass: MIT Center for Advanced Engineering Study; 1986.
4. Box G, Bisgard S. The scientific context of quality improvement. *Qual Prog*. 1987;4:54–62.
5. Blumenthal D, Kilo CM. A report card on continuous quality improvement. *Milbank Q*. 1998;76(4):625–648.
6. Morrison E, Mobley D, Farley B. Research and continuous improvement: the merging of two entities? *Hosp Health Serv Adm*. 1996;41(3):359–372.
7. Campbell DT, Stanley JC, Gage NL. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally College Pub Co; 1966.
8. Cohen J. The earth is round ($p < .05$). *Am Psychol*. 1994; 49:997–1003.
9. Feinstein AR. *Clinical epidemiology: the architecture of clinical research*. Philadelphia: Saunders; 1985.
10. Pellegrin KL, Carek D, Edwards J. Use of experimental and quasi-experimental methods for data-based decisions in QI. *Jt Comm J Qual Improv*. 1995;21(12):683–691.
11. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *NEJM*. 2000;342(25):1878–1886.
12. Shortell SM, Bennett CL, Byck GR. Assessing the impact of continuous quality improvement on clinical practice: what it will take to accelerate progress. *Milbank Q*. 1998;76(4):593–624.
13. Walton M. *Deming Management at Work*. New York: Putnam Publishing Group; 1991.
14. Western Electric. *Statistical Quality Control Handbook*. 2nd ed. Indianapolis, Ind: AT&T Technologies; 1958.
15. Batalden PB, Mohr JJ, Nelson EC, et al. Continually improving the health and value of health care for a population of patients: the panel management process. *Qual Manag Health Care*. 1997;5(3):41–51.
16. Batalden P, Nelson E, Roberts J. Linking outcomes measurement to continual improvment: the serial V way of thinking about improving clinical care. *Jt Comm J Qual Improv*. 1994;20(4):167–180.
17. Nelson EC, Batalden PB, Plume SK, Mohr JM. Improving health care: Part 2. Clinical improvement worksheet and users manual. *Jt Comm J Qual Improv*. 1996;22(8):531–548.
18. Nelson EC, Splaine ME, Godfrey MM, et al. Using data to improve medical practice by measuring processes and outcomes of care. *Jt Comm J Qual Improv*. 2000;26(12):667–685.
19. Nelson EC, Mohr JJ, Batalden PB, Plume SK. Improving health care. Part 1, The clinical value compass. *Jt Comm J Qual Improv*. 1996;22(4):243–256.
20. Luthans F, Thompson KR. Theory D and O.B. Mod.: Synergistic or opposite approaches to performance improvement? *J Organ Behav Manage*. 1987;9(1):105–124.
21. Mawhinney TC. OBM, SPC, and Theory D: a brief introduction. *J Organ Behav Manage*. 1986;8(1):89–105.
22. Redmon WK, Dickinson AM. A comparative analysis of statistical process control, theory D, and behavioral analytic approaches to quality control. *J Organ Behav Manage*. 1987;9(1):47–65.
23. Plsek PE. Techniques for managing quality. *Hosp Health Serv Adm*. 1995;40(1):50–79.
24. Weinstein J, Brown P, Hanscom B, Walsh T, Nelson EC. Designing an ambulatory clinical practice for outcomes improvement: from vision to reality—The Spine Center at Dartmouth—Hitchcock, year one. *Qual Manag Health Care*. 2000;8(2):1–20.
25. Mawhinney T. Total quality management and organizational behavior management: an integration for continual improvement. *J Appl Behav Anal*. 1992;25(3):525–543.
26. Pfadt A, Cohen IL, Sudhalter V, et al. Applying statistical process control to clinical data: an illustration. *J Appl Behav Anal*. 1992;25(3):551–560.
27. Pfadt A. Using control charts to analyze baseline stability: an illustrative example with "real time" data. *J Organ Behav Manage*. 1999;18(4):53–60.
28. Redmon W. Opportunities for applied behavior analysis in the total quality movement. *J Appl Behav Anal*. 1992;25(3):545–550.
29. Gikalov AA, Baer DM, Hannah GT. The effects of work task manipulation and scheduling on patient load, revenue, eyewear turnover, and utilization of staff and doctor time. *J Organ Behav Manage*. 1997;17(1):3–33.
30. Barlow DH, Hersen M. *Single Case Experimental Designs: Strategies for Studying Behavior Change*. 2nd ed. New York: Pergamon Press; 1984.
31. Kazdin AE, Tuma AH. *Single-Case Research Designs*. San Francisco: Jossey-Bass; 1982.
32. Kazdin AE. *Single-Case Research Designs: Methods for Clinical and Applied Settings*. New York: Oxford University Press; 1982.
33. McReynolds LV, Thompson CK. Flexibility of single-subject experimental designs. Part I, Review of the basics of single-subject designs. *J Speech Hear Disord*. 1986;51:194–203.
34. Mawhinney TC, Austin J. Speed and accuracy of data analysts' behavior using methods of equal interval graphic data charts, standard celeration charts, and statistical

process control charts. *J Organ Behav Manage*. 1999;18(4):5–45.
35. Benneyan JC. Statistical quality control methods in infection control and hospital epidemiology, Part II, Chart use, statistical properties, and research issues. *Infect Control Hosp Epidemiol*. 1998;19(4):265–283.
36. Grant E, Leavenworth R. *Statistical Quality Control*. New York: McGraw-Hill; 1980.
37. Levi AS, Mainstone LE. Obstacles to understanding and using statistical process control as a productivity improvement approach. *J Organ Behav Manage*. 1987;9(1):23–32.
38. Nolan TW, Provost LP. Understanding variation. *Qual Prog*. May 1990;23(5):70–78.
39. Pyzdek T. *Pyzdek's Guide to SPS: Vol 2, Applications and Special Topics*. Milwaukee, Wis: ASQC Quality Press; 1992.
40. Wheeler DJ, Chambers DS. *Understanding Statistical Process Control*. 2nd ed. Knoxville, Tenn: SPC Press; 1992.
41. Saturno PJ, Felices FF, Segura J, Vera A, Rodriguez JJ, and the ARIAM Project Group. Reducing time delay in the thrombolysis of myocardial infarction: an internal quality improvment project. *Am J Med Qual*. 2000;15(3):85–93.
42. Boggs PB, Hayati F, Washburne WF, Wheeler DA. Using statistical process control charts for the continual improvement of asthma care. *J Qual Improv*. 1999;25(4):163–181.
43. Gibson P, Wlodarczyk J, Hensley M, Murree-Allen K, Olson L, Saltos N. Using quality-control analysis of peak expiratory flow recordings to guide therapy for asthma. *Ann Intern Med*. 1995;123(7):488–492.
44. Burnett L, Chesher D. Application of CQI tools to the reduction in risk of needlestick injury. *Infect Control Hosp Epidemiol*. 1995;16(9):503–505.
45. Trochim WMK. The regression-discontinuity design. Paper presented at: Research methodology. Strengthening causal interpretations of nonexperimental data; May 1990; Washington, DC.
46. Deppen S. Understanding seasonality in healthcare: improving pediatric asthma care. Paper presented at: Eight International Scientific Symposium on Improving the Quality and Value of Healthcare; 2002; Orlando, Fla.
47. Alemi F, Rom W, Eisenstein E. Risk adjusted control charts for health care assessment. *Ann Oper Res*. 1996;67:45–60.
48. Kearns KP. Flexibility of single-subject experimental designs. Part II, Design selection and arrangement of experimental phases. *J Speech Hear Disord*. 1986;51:204–214.
49. Connell PJ, Thompson CK. Flexibility of single-subject experimental designs. Part III, Using flexibility to design or modify experiments. *J Speech Hear Disord*. 1986;51:214–225.
50. Guyatt GH, Sackett D, Adachi JD, et al. A clinician's guide for conducting randomized trials in individual patients. *Can Med Assoc J*. 1988;139:497–503.
51. Guyatt GH, Heyting A, Jaeschke R, Keller J, Adachi JD, Roberts RS. N of 1 randomized trails for investigating new drugs. *Control Clin Trials*. 1990;11:88–100.
52. Lu C-W, Reynolds MRJ. EWMA control charts for monitoring the mean of autocorrelated processes. *J Qual Technol*. 1999;31(2):166–186.
53. Lu C-W, Reynolds MRJ. Control charts for monitoring the mean and variance of autocorrelated processes. *J Qual Technol*. 1999;31(3):259–275.
54. Woodall WH, Montgomery DC. Research issues and ideas in statistical process control. *J Qual Technol*. 1999;31(4):376–386.
55. Box GEP. Statistics as a catalyst to learning by scientific method. Part II, A discussion. *J Qual Technol*. 1999;31(1):16–29.
56. Sloan MD. *Using Designed Experiments to Shrink Health Care Costs*. Milwaukee, Wis: ASQC Quality Press; 1997.
57. Fischer LR, Solberg LI, Kottke TE. Quality improvement in primary care clinics. *Jt Comm J Qual Improv*. 1998;24(7):361–370.
58. Solberg LI, Reger LA, Pearson TL, et al. Using continuous quality improvement to improve diabetes care in populations: the IDEAL model. Improving care for Diabetics through empowerment, active collaboration and leadership. *Jt Comm J Qual Improv*. 1997;23(11):581–592.
59. Goldberg HI, Wagner EH, Fihn SD, et al. A randomized controlled trial of CQI teams and academic detailing: can they alter compliance with guidelines? *Jt Comm J Qual Improv*. 1998;24(3):130–142.
60. Curley C, McEachern JE, Speroff T. A firm trial of interdisciplinary rounds on the inpatient medical wards: an intervention designed using continuous quality improvement. *Med Care*. 1998;36(8 suppl):AS4–AS12.
61. Dietrich AJ, O'Connor GT, Keller A, Carney PA, Levy D, Whaley FS. Cancer: improving early detection and prevention. A community practice randomized trial. *BMJ*. 1992;304(6828):687–691.
62. Hospitals make significant progress in CQI efforts between 1993 and 1998. *Health Care Strateg Manage*. 1999;17:6.