Grand rounds in methodology: four critical decision points in statistical process control evaluations of quality improvement initiatives

Marang-van de Mheen, P.J.; Woodcock, T.

Grand rounds in methodology: four critical decision points in statistical process control evaluations of quality improvement initiatives

Perla J Marang-van de Mheen [1], Thomas Woodcock [2]

[1]Department of Biomedical Data Sciences, Medical Decision Making, J10-S, Leiden University Medical Center, Leiden, The Netherlands
[2]National Institute for Health Research Applied Research Collaboration Northwest London, Imperial College London, London, UK

Correspondence to
Dr Perla J Marang-van de Mheen, Department of Biomedical Data Sciences, Medical Decision Making, J10-S, Leiden University Medical Center, 2300 RC Leiden, The Netherlands; p.j.marang-van_de_mheen@lumc.nl

Linked

Check for updates

ABSTRACT

Quality improvement (QI) projects often employ statistical process control (SPC) charts to monitor process or outcome measures as part of ongoing feedback, to inform successive Plan-Do-Study-Act cycles and refine the intervention (formative evaluation). SPC charts can also be used to draw inferences on effectiveness and generalisability of improvement efforts (summative evaluation), but only if appropriately designed and meeting specific methodological requirements for generalisability. Inadequate design decreases the validity of results, which not only reduces the chance of publication but could also result in patient harm and wasted resources if incorrect conclusions are drawn. This paper aims to bring together much of what has been written in various tutorials, to suggest a process for using SPC in QI projects. We highlight four critical decision points that are often missed, how these are inter-related and how they affect the inferences that can be drawn regarding effectiveness of the intervention: (1) the need for a stable baseline to enable drawing inferences on effectiveness; (2) choice of outcome measures to assess effectiveness, safety and intervention fidelity; (3) design features to improve the quality of QI projects; (4) choice of SPC analysis aligned with the type of outcome, and reporting on the potential influence of other interventions or secular trends.

These decision points should be explicitly reported for readers to interpret and judge the results, and can be seen as supplementing the Standards for Quality Improvement Reporting Excellence guidelines. Thinking in advance about both formative and summative evaluation will inform more deliberate choices and strengthen the evidence produced by QI projects.

THE SITUATION IN PRACTICE

Many quality improvement (QI) projects start because physicians, nurses or other healthcare professionals encounter a problem in practice that they want to improve. Some common triggers include:

► Outcome data showing variation or disparities in care; for example, indigenous Maya women in Guatemala, who mostly deliver at home, are more than twice as likely to die from obstetric complications than non-indigenous mothers.[1]

► Worrying trends within a particular service or care provider organisation; for example, an increase in newly prescribed benzodiazepines and sedative-hypnotic

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Many tutorials have explained the advantages of statistical process control (SPC) techniques over traditional statistical testing in quality improvement (QI) projects, the basic principles, how to select and construct SPC charts and specific issues such as sampling considerations. Little has been written on how to bring these together in a process for using SPC in QI projects, highlighting critical decision points that are often missed but affect the inferences that can be drawn about the effectiveness of the intervention.

WHAT THIS STUDY ADDS

⇒ Critical decision points that should be explicitly reported for readers to interpret and judge the results include
  1. establishing a stable baseline;
  2. the choice of outcome measures;
  3. QI design features; and
  4. SPC analysis used to draw inferences on the intervention effect.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Thinking in advance about the kind of conclusions that need to be drawn from a QI project will inform explicit choices on design, measurements and SPC analyses and thereby allow for greater generalisability of results.

drugs during patients' hospitalisation, putting them at risk for long-term use and drug-related problems.[2]

► Inefficient processes or unnecessary tests; for example, some laboratory tests commonly ordered in combination with another, adding little value to patient management at significant cost.[3]

► Problems in sustaining improvements made previously; for example, adherence to central line maintenance bundles for every patient every time.[4]

► Patient feedback; for example, patients reporting loss of medication information across care settings.[5]

A common approach is for a team of stakeholders to investigate root causes of the problem and create driver diagrams to develop a programme theory and intervention, for example, using the action effect method as a structured approach to guide this process.[6] Multiple Plan-Do-Study-Act (PDSA) cycles are then commonly used to implement the intervention, adding or adapting elements to make it more effective, guided by targeted process and outcome measures.[7] Particularly if positive results appear, the team may want to disseminate the project findings so that others can benefit from the lessons learnt. The Standards for Quality Improvement Reporting Excellence (SQUIRE 2.0) reporting guidelines state that such reports should include 'qualitative and quantitative methods used to draw inferences from the data'.[8] Very often, the desired inference concerns whether and to what extent the intervention causes improvement in the targeted outcome—the effectiveness of the intervention. This raises the question: has the study been designed such that changes in outcome can really be attributed to the intervention?

QI projects often employ run charts or statistical process control (SPC) charts to monitor a targeted process or outcome measure, to understand and inform successive PDSA cycles. This approach is an example of formative evaluation; that is, ongoing feedback to refine an intervention. Run charts can help identify upward or downward trends and thereby whether the targeted measure is moving in the right direction, but cannot establish whether a process is in control or not. SPC charts can do both as these have control limits, and therefore add the ability to detect when changes are needed to make the process stable. This type of formative evaluation is often not consistent with the methodological requirements to use SPC for summative evaluation, that is, to draw inferences on effectiveness and generalisability of the improvement effort. If one only starts to think about these methodological issues when writing up a project for publication, it will be too late to change the design to allow for robust and generalisable conclusions concerning effectiveness. In many cases, SPC techniques can be used both for formative evaluation and to draw inferences concerning effectiveness, but only if appropriately designed. Inadequate design reduces the chance of publication in a peer-reviewed journal and more importantly could also result in patient harm and wasted resources if incorrect conclusions are drawn.

There have been many tutorials explaining the advantages of SPC over traditional statistical testing in QI projects, outlining the basic principles, how to select and construct SPC charts, or specific issues such as sampling considerations.[9–11] However, little has been written on bringing these together in a process for using SPC in QI projects, highlighting critical decision points that are often missed but can affect how inferences about effectiveness may be drawn. This paper will address this gap by explaining how these critical decision points are inter-related and how they can be addressed by a careful plan for SPC.

NEED FOR A STABLE BASELINE

Many improvement projects fail to establish a stable baseline against which to identify and quantify any improvements made. They may have either insufficient data points for a stable baseline, or encounter special cause variation within their baseline, meaning it is not stable and may be changing even prior to introduction of an intervention. To understand why a stable baseline is crucial to draw inferences from SPC analyses, we first briefly review the principles of SPC, described in more detail elsewhere.[9] SPC is based on the principle that there is variation in any process, but that the variation is predictable if that process is stable ('common cause variation'). Furthermore, for a stable process, we can compute the range of values within which this variation occurs, based on observed data and a hypothesised underlying statistical distribution (eg, Gaussian, binomial or Poisson, depending on the measure of interest).[9] For instance, under common cause variation, average values of a continuous variable at every data point (eg, weekly averages) will tend to be normally distributed, and we can expect that if the process remains stable, 99.7% of future measurements will be within ±3 SD of the mean. Within these limits, we consider the process to be 'in statistical control'; the limits are therefore called control limits. If an intervention is introduced, this may disturb the expected pattern (special cause variation), meaning that the measurements will deviate from the predicted range.

From these principles, it is clear that if we are to attribute evidence of special cause variation to an intervention, we first need to establish that there is a stable process before introducing the intervention, which is referred to as a stable baseline (figure 1A). Establishing a stable baseline will usually take about 20–25 data points for SPC charts,[11 12] before we can test whether subsequent measurements are starting to deviate from what is expected. This can be understood from the fact that each data point is a sample taken, and that there is likely variation from sample to sample. Taking only a few samples will therefore not give a very good representation of the true mean, similar to
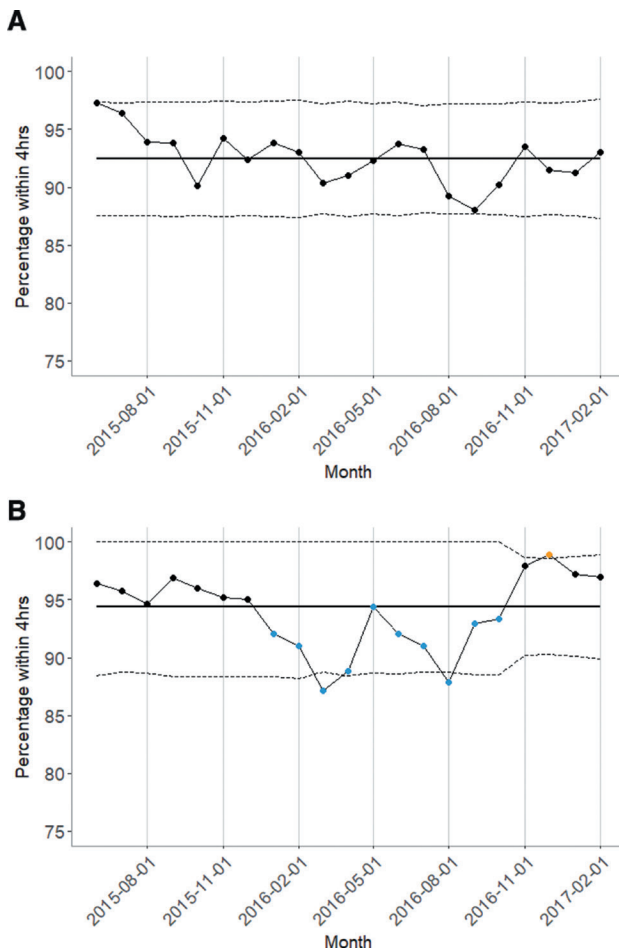
Figure 1 Data are monthly percentages of emergency department attendances admitted to hospital, discharged home or transferred to another provider within 4 hours, for two hospitals in England, taken from figures published by NHS England (adjusted for overdispersion).[31] The baseline process is stable for the hospital depicted in (A), but is unstable for the hospital in (B).

having only few observations in traditional statistical analysis. Even though it is possible to calculate control limits with fewer data points, the charts become more powerful when at least 20 data points are used.[11] Run charts need a similar stability to enable detection of a changing trend, which usually takes about 10–15 data points,[11] so these are sometimes used if the urgency of improvement outweighs the need for a stable baseline. The sample size for each data point is also important in this context, as this influences variation in the mean (or other statistic of interest) and hence the width of the control limits. Together, these determine the power to detect differences at a certain point in time. Further guidance regarding sampling considerations, the minimum sample size and power calculations is available elsewhere.[11 12] Related to this is the choice of the time unit, for example, whether monthly or weekly measures are used. This will likely depend on the number of eligible patients or available data. The rationale for the number of data points should therefore be reported (similar to other time series

techniques),[13] rather than only the total time before and after the intervention.

Sometimes, the preintervention process is not stable, that is, there is special cause variation in the baseline, for instance, one point outside the control limits (figure 1B). Once this is detected, there are several things that can be done. The first is to look for causes of the special cause variation; for example, this may occur in a particular subgroup of patients and once the data from this subgroup are removed from the analysis, the process is stable for all other patients. Removing a subgroup of patients from the analyses will obviously limit the generalisability, but at least allows inferences to be drawn on those remaining patients. This does not necessarily entail withholding the intervention from any subgroup of patients, and care must be taken to ensure certain groups are not disadvantaged or discriminated against, as this merely relates to the analyses. Another option is to delay the introduction of the intervention, particularly if the special cause variation occurred only early in the baseline or is limited to a single data point, so that adding a few more data points yields a stable baseline. It is important that this is done iteratively until a stable baseline is achieved—the resulting control limits are sometimes known as 'trial limits'—otherwise predictions about future outcomes and thereby inferences about the intervention effect will be invalid and misleading. In practice, improvement teams have to balance the potential benefits of waiting for a stable baseline against the potential harm of delaying onset of improvement efforts. There will not always be an approach that can achieve the best of both worlds—but at least if such questions are considered up front, this will be a conscious decision.

CHOICE OF OUTCOME MEASURES

As for any intervention study, a range of outcome measures is required for a QI study to cover the effectiveness (primary outcome measure), safety (balancing measures) and compliance (intervention fidelity measures) with the intervention, using existing or newly collected data.

The primary outcome measure should capture the key quality or safety issue targeted by the QI project, which can relate to an outcome for patients or a care process. The definition and data collection method for this outcome measure should be the same before and after the intervention, and for all healthcare settings (including control settings if applicable) as slight changes will affect the effect attributed to the intervention. For instance, using retrospective (existing) data before the intervention versus prospective data after the intervention is likely to induce methodological effects unrelated to the intervention. In practice, because a stable baseline needs to be established (which might take time) and the QI team may be keen to start implementing the intervention, this often means that existing (routinely collected) data are used. Examples

include the number of opioid-related oversedation events per 1000 patient-days,[14] percentage of primary care patients lost to follow-up,[15] time to receive antibiotics[16] or number of deliveries between newborns with an Apgar score <7 after 5 min.[17] Even though it is possible to use prospectively collected data, enough data must first be collected to establish a stable baseline before starting the intervention. Using existing data therefore has the advantage of being able to start developing and implementing the intervention more quickly, provided that the baseline process is indeed stable. What we encounter frequently though is that there has been an audit of only a few data points showing that there is a problem of some kind, which then initiated a QI project and implementation of an intervention a year later. Even though a lot of effort has gone into such projects, without a stable baseline and continuous longitudinal data we cannot attribute any change in outcome to the intervention.

Balancing measures are important to include in a QI project to ensure the intervention has no unintended effects, and therefore also require the same data before and after introduction of the intervention. Examples of balancing measures include falls that result in injury,[2] prehospital time[18] and major adverse events.[19] Intervention fidelity measures are used to establish whether the intervention has been delivered as intended, similar to treatment compliance in randomised controlled trials. These are very useful as part of the PDSA cycles, as they are often intermediate processes to improve the primary outcome, and so provide information regarding the need to adapt the intervention. Showing improvements in these intermediate measures therefore also contributes information on the likelihood that the intervention (rather than something else) produced the change in the primary outcome. Examples of intervention fidelity measures include compliance with clinical care bundles,[4 17] the percentage of infants with verbal consent documented[19] or the proportion of clinical care providers receiving an educational intervention.[2] These fidelity measures are mostly collected prospectively during intervention implementation and can add further evidence concerning the causal relationship between the intervention and outcome.

QI STUDY DESIGNS

The design of a QI study is important as it defines the data collection, analysis and inferences that can be drawn, and should therefore be explicitly reported. For instance, an uncontrolled before-after study giving only average estimates in two periods will not be able to control for any secular trends.

First, the use of routinely collected (existing) data versus newly collected (prospective) data is an important decision, similar to cohort studies where retrospective refers to already available data and prospective to newly collected data. This distinction is important as use of prospective data allows for choices on which data to collect and how to measure, including for relevant confounders, whereas these possibilities are more limited when existing data are used.

Second, the time horizon available for a QI project may affect choices on the time unit for the longitudinal data. If, for instance, the QI project needs to be conducted within 1 year, this means that monthly data on the outcomes will not give enough data points to establish a stable baseline, let alone for any evaluation of the intervention effect. Weekly or fortnightly data may give enough data points, provided there is sufficient sample size for each data point, that is, the total number of eligible patients every week or 2 weeks. Alternatively, if a primary outcome can only be assessed monthly, this means that the time horizon of the QI project needs to be longer.

Third, it is important to consider whether a controlled or uncontrolled QI study design is to be used. A controlled study is a stronger design to draw inferences on the effectiveness of the intervention, increasing the likelihood that the intervention produced the change in outcome if it did not change in the control group. There are different types of controls that can be considered for QI studies, which will help to deal with different types of confounding as reported for other time series[20] (table 1). Previous QI studies have used:

- Location-based controls, for example, another ward in the same hospital not exposed to the intervention.[2 21]
- Characteristic-based controls, for example, a study aiming to reduce readmissions specifically for patients with heart failure, used patients with acute myocardial infarction and pneumonia as controls.[22]
- Outcome-based controls, for example, a study targeting hand hygiene to improve healthcare-associated infections attributable to inpatient or outpatient care, used infections attributable to the operating room as the control as these were expected to be less sensitive to changes in hand hygiene.[23]

When selecting a control, it is important to consider the scale of the intervention and risk of contamination. For instance, this might include whether professionals working in another location or another patient group chosen as control could still be exposed to the intervention. In addition, one needs to consider the type of factors that will be controlled for and what cannot be controlled because these factors are uniquely tied to the intervention group or outcome. For example, examining the changes in outcome in a control ward in the same hospital will show the impact of, for example, a new hospital-wide policy on the outcome, which can be taken into account when interpreting the impact of the intervention in the intervention ward. However, a change in ward policy in the intervention ward that affects the outcome cannot be separated from the intervention effect. The choice of control also depends on whether it will be possible to

Table 1 Controlled quality improvement studies—different types of controls take care of different types of confounding

Type of control	Description	What it controls for (and what not)
Location-based controls	The control is another location that is similar to the intervention location but has not received the intervention, for example, another ward in the same hospital or another similar hospital.	Controls for confounding factors that will affect both locations, but not factors uniquely tied to the intervention location.
Characteristic-based controls	If the intervention targets specific patient groups defined by their characteristics, for example, elderly, patients with a specific diagnosis, then non-targeted groups may act as controls.	Controls for confounding factors that will affect both patient groups, but not factors specific for the intervention group. If the characteristic defining who is intervention/control (eg, age) is a confounder, then this cannot be controlled for.
Historical controls	This type of control can only be used if such historical data are available (and with sufficient completeness).	Potentially, it might help to control for seasonal effects, but cannot control for confounding factors that are specific for the period when the intervention was implemented (eg, changes in case mix).
Outcome controls	The control is another outcome in the same population not targeted by the intervention (or expected to be not very sensitive for the intervention).	Controls for those confounding factors that would affect both the intervention and control outcome, but not those unique for the intervention outcome.

collect similar data on the outcome measures as for the intervention group.

Even though not included as a separate item in the SQUIRE guidelines, it would be helpful if authors include an explicit description of their study design. Explicit reporting on the data used, time horizon, (un)controlled design, type of control and which confounding factors are controlled in this way, will give more insight into the quality of evidence generated by this particular QI study.

SPC ANALYSES TO DRAW INFERENCES ON THE INTERVENTION EFFECT

Choosing the appropriate type of SPC chart is the first step in SPC analyses, directly linked to the outcome for which the analysis is conducted. Outcomes can be based on different types of variables (eg, continuous, percentage) that follow different statistical distributions, which subsequently determines calculation of control limits. Just as the appropriate regression analysis is chosen based on the type of outcome, the same is true for choice of SPC chart (table 2) as described in more detail elsewhere.[10 12] Although continuous measures contain more information than dichotomised equivalents, many QI projects have primary outcomes expressed as percentages/proportions and hence the p-chart is most frequently used. More complex charts such as the exponentially weighted moving averages and cumulative sum charts use accumulated information over time and are thereby able to detect small

Table 2 Choosing the appropriate type of statistical process control chart aligned with the type of outcome data

Type of chart	Type of outcome data with example
X-bar S-charts	For continuous data. Under only common cause variation, the means will follow the normal distribution, such as length of stay[32] or time to antibiotics. These are used in combination where the X-bar chart will investigate whether the intervention has improved the mean, and the s-chart whether it has improved the SD (ie, less variability) of the outcome. Both are needed as the X-bar control limits are calculated under the assumption that the SD is in control, and one may go out of control independently from the other.
P-Charts	For discrete event data. Under only common cause variation, the proportions follow the binomial distribution, such as percentage of patients prescribed a new sedative[2] or percentage of patients not seen in the last year.[15]
C-Chart or U-chart	For count data. Under only common cause variation, the counts follow the Poisson distribution. c-Charts are used when the variable counts the total number of an event such as the total number of patient falls in a week (provided a stable 'area of opportunity', eg, consistent number of patients on the ward across weeks). u-Charts are used when the variable is an average rate of an event, adjusted for a denominator size (or 'area of opportunity'), such as the number of opioid-related oversedation events per 1000 patient-days.[14]
G-Chart	For variables that follow the skewed geometric distribution under only common cause variation. Can be used to investigate time (or number of patients) between outcomes, particularly useful if the outcome occurs infrequently, such as the number of deliveries between newborns with an Apgar score <7 after 5 min.[17]
(EW)MA chart	For different types of data, plots moving averages (MA) over time and thereby includes more measurements in every plotted data point. MA charts give equal weight to all previous measurements within a rolling window of a specified size, whereas exponentially weighted (EW)MA charts place the greatest weight on the most recent observation. Particularly useful to detect small or gradual shifts in the outcome, such as small increases in surgical site infections and thereby earlier detection of an outbreak.[33 34]
CUSUM chart	For different types of data, the cumulative sum chart statistic summarises the extent to which an outcome shifts away from a baseline rate with higher values corresponding to stronger evidence of a change. Particularly useful for earlier detection of small increases in event rates, such as a worsening in 1-year revision rates after total hip and knee replacements.[35]
CUSUM, cumulative sum.	

changes more quickly than the p-chart.[24] U-Charts are appropriate for rates, where the number of events is adjusted for the time at risk, for example, the rate of central line infections per 1000 days in situ. C-Charts monitor counts, for example, the number of falls per week. If the correct chart type is chosen, deviations from the underlying distribution will show up as special causes. An incorrect choice of chart may result in poor chart performance, such as an increased risk of false positives.

The next step is to plot the observed outcome data over time and to calculate the centre line and control limits, to allow detection of special cause variation introduced by the intervention. Most QI projects use ±3 SD control limits to limit the risk of a type I error—incorrectly detecting special cause variation (ie, false-positive signals). Where traditional statistical techniques mostly accept a 5% type I error risk to test one hypothesis at a time and use ±2 SD for clinical decisions, control charts consist of many data points and therefore multiple points contributing to the overall false-positive risk. As shown elsewhere, a control chart of 25 data points using 3 SD limits would have an overall false-positive rate of 6.5% whereas using 2 SD limits would increase this to 27.7%.[9] To detect special cause variation, several rules are commonly used that can be summarised as data falling outside the control limits or displaying abnormal patterns within the control limits, that would not be expected under common cause variation[9 10]:

▶ One point outside the upper or lower control limit, as this is outside the range of predicted values for common cause variation.

▶ Trend: a run of 8 successive points[10] (some prefer fewer points[9]) trending up or down.

▶ Shift: 8 successive points on one side of the centre line (again, some prefer a different cut-off).

▶ Two out of 3 successive points beyond 2 SD on the same side of the centre line. One point beyond the 2 SD limit may be an extreme value within the range of predicted values, but two consecutive extreme values are not very likely under random variation.

Which rules are employed should be decided a priori and specified in a QI report. It is helpful to annotate the SPC charts with the different PDSA cycles to show when different elements were added or adapted, and to show when special cause variation occurred (eg, by a different colour).

Certain types of special cause variation, notably shift rule breaks, may indicate an improvement (or indeed a deterioration) in the process being measured. If such a change persists, a new centre line with new control limits should be established to encapsulate the expected variation in this new process. This may happen more than once if a project is successful. For instance, in a recently published QI report the mean monthly percentage of infants receiving timely hepatitis B vaccination increased from a baseline mean

of 45% to 76%, and then to 95%.[19] Other projects have used centre lines before versus during intervention implementation to show the intervention effect,[4] before versus after the intervention ended[15] or during sustainability phases.[16] In such cases, where the new limits are established based on a time period criterion rather than on special cause variation, it is important to note that differences in the mean between the two periods may be due to common cause variation alone, that is, the intervention may not have caused a change. Regardless of the criteria used to determine when new limits are established, each set of limits must be based on sufficient data, as described for the baseline period above. For instance, if at least 25 data points are required to establish baseline limits, it is logical to use at least 25 data points to form control limits for a new, improved process.

There may be factors other than the intervention affecting the outcome and thereby introducing special cause variation. Such confounding can be dealt with by statistical adjustment or stratification as in other research. It is important to report explicitly about any other interventions occurring in the same time period that may have affected the outcome, particularly for uncontrolled QI studies where this could be due to a secular trend. For controlled studies, such reporting should focus on other interventions specific for the intervention group such as a new ward policy (when using a location-based control) or a policy affecting specific outcomes only (when using an outcome-based control). The controlled design will not be able to separate the effect of these factors from those attributable to the intervention itself, highlighting the importance of selecting a control carefully and reporting explicitly on the type of confounding it will control for. This also re-enforces that intervention fidelity measures need to be chosen carefully to show <u>how</u> the intervention may have produced its results, and evaluate whether changes in these intermediate variables are concurrent with changes in the primary outcome. The combination of effects shown for the primary outcome measure in the intervention group (vs control) and intervention fidelity measures will determine our understanding of how the intervention has worked and the likelihood it has produced the changes in outcomes.

DISCUSSION AND RECOMMENDATIONS

Even though not every QI project sets out to establish evidence on the effectiveness of an intervention, many end up seeking to do so and it is therefore important to plan for both formative and summative evaluation needs. By thinking in advance about the kind of conclusions you would like to be able to draw, this will inform explicit choices on the design, measurements and SPC analyses. The present paper has highlighted the critical decision points and shown how these are inter-related with many aspects of a QI study. Explicit reporting on these critical decision points can be seen

Table 3 Critical decision points that need explicit reporting for a quality improvement study

Critical decision point	Reporting requirements
Need for a stable baseline to draw inferences	Rationale for number of data points, the time unit and sample size for each data point
Choice of outcome measures	The same primary outcome before and during intervention implementation, including definition and data collection. Balancing and intervention fidelity process measures report on safety and compliance with the intervention.
QI design features	Explicit reporting on several design features: ► Routinely collected (existing) data or newly collected (prospective) data. ► Time horizon. ► Controlled or uncontrolled design, type of control and for what type of confounding factor this controls (and what not). For instance, the study design in the title could be: a controlled 1-year QI study with routinely collected data. Additional detail can be reported in the methods.
SPC analysis	Explicit reporting on: ► Type of chart and which control limits. ► Which specific rules for special cause variation. ► Calculation of new centre line and control limits. ► Rationale for number of data points to establish a new stable process (sustainability). ► Other interventions or secular trends that may affect the outcome measures.

QI, quality improvement; SPC, statistical process control.

as an add-on to the SQUIRE guidelines as routinely used for QI reports, and needed for editors, reviewers and readers to interpret and judge the results (table 3). This paper also adds more specific detail to the features of a high-quality measurement plan guiding all phases of the project, including planning the (SPC) analysis, handling missing data and possible confounders.[25]

Making more deliberate choices and explicit reporting on these choices will strengthen QI projects and thereby the evidence they can produce. The need for rigorous methodology and identification of sources of bias in QI projects has been argued previously, including that statements to the effect that data 'are for quality improvement' and not 'research' may run the risk of promoting less rigorous standards.[26] The primary aim at the start of many QI projects may be to improve a particular problem in daily practice, rather than to produce new knowledge as in research.[27] SPC charts are then often used as an approach to measurement, where data will give feedback and inform further development of the intervention. Previous studies have shown that using SPC charts as part of the intervention (formative evaluation) can be effective to improve patient outcomes.[28 29] Even though then primarily a tool for monitoring rather than to draw inferences about the effect of each implemented improvement initiative, this also requires planning the feedback, defining what constitutes special cause variation and so on. Similarly, when used as a statistical tool to draw inferences about the effect of the intervention (summative evaluation), this requires planning as otherwise decisions taken during the process might affect the ability to draw generalisable conclusions. In learning health systems, there will be continual QI projects to evaluate whether interventions to improve care are effective and reduce harm, which require the best possible methods and strongest possible designs to ensure the evidence is as rigorous as possible even if randomisation may not always be

possible.[30] Ultimately, evaluation of QI projects cannot be an afterthought as we need to know which of our efforts have worked to improve care, to ensure that patients will benefit.

ORCID iDs
Perla J Marang-van de Mheen http://orcid.org/0000-0003-1439-0989
Thomas Woodcock http://orcid.org/0000-0002-4735-4856

REFERENCES

1 Austad K, Juarez M, Shryer H, *et al*. Obstetric care navigation: results of a quality improvement project to provide accompaniment to women for facility-based maternity care in rural Guatemala. *BMJ Qual Saf* 2020;29:169–78.

2 Fan-Lun C, Chung C, Lee EHG, *et al*. Reducing unnecessary sedative-hypnotic use among hospitalised older adults. *BMJ Qual Saf* 2019;28:1039–45.

3 Strauss R, Cressman A, Cheung M, *et al*. Major reductions in unnecessary aspartate aminotransferase and blood urea nitrogen tests with a quality improvement initiative. *BMJ Qual Saf* 2019;28:809–16.

4 Kamity R, Grella M, Kim ML, *et al*. From kamishibai card to key card: a family-targeted quality improvement initiative to reduce paediatric central line-associated bloodstream infections. *BMJ Qual Saf* 2021;30:72–81.

5 Garfield S, Furniss D, Husson F, *et al*. How can patient-held lists of medication enhance patient safety? A mixed-methods study with a focus on user experience. *BMJ Qual Saf* 2020;29:764–73.

6 Reed JE, McNicholas C, Woodcock T, *et al*. Designing quality improvement initiatives: the action effect method, a structured

approach to identifying and articulating programme theory. *BMJ Qual Saf* 2014;23:1040–8.

7 Brandrud AS, Nyen B, Hjortdahl P, *et al*. Domains associated with successful quality improvement in healthcare - a nationwide case study. *BMC Health Serv Res* 2017;17:648.

8 Ogrinc G, Davies L, Goodman D, *et al*. Squire 2.0 (standards for quality improvement reporting excellence): revised publication guidelines from a detailed consensus process. *BMJ Qual Saf* 2016;25:986–92.

9 Benneyan JC, Lloyd RC, Plsek PE. Statistical process control as a tool for research and healthcare improvement. *Qual Saf Health Care* 2003;12:458–64.

10 Mohammed MA, Worthington P, Woodall WH. Plotting basic control charts: tutorial notes for healthcare practitioners. *Qual Saf Health Care* 2008;17:137–45.

11 Perla RJ, Provost LP, Murray SK. Sampling considerations for health care improvement. *Qual Manag Health Care* 2013;22:36–47.

12 Benneyan JC. Design, use, and performance of statistical control charts for clinical process improvement. Boston, September 16, 2001. Available: https://www.researchgate.net/publication/228941925

13 Ramsay CR, Matowe L, Grilli R, *et al*. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. *Int J Technol Assess Health Care* 2003;19:613–23.

14 Vermaire D, Caruso MC, Lesko A, *et al*. Quality improvement project to reduce perioperative opioid oversedation events in a paediatric Hospital. *BMJ Qual Saf* 2011;20:895–902.

15 Harris AT, Hoover C, Cmolik B, *et al*. Improving access in a VA primary care clinic using an innovative panel retention tool: a quality improvement report. *BMJ Qual Saf* 2020;29:947–55.

16 Dandoy CE, Hariharan S, Weiss B, *et al*. Sustained reductions in time to antibiotic delivery in febrile immunocompromised children: results of a quality improvement collaborative. *BMJ Qual Saf* 2016;25:100–9.

17 Hollesen RVB, Johansen RLR, Rørbye C, *et al*. Successfully reducing newborn asphyxia in the labour unit in a large academic medical centre: a quality improvement project using statistical process control. *BMJ Qual Saf* 2018;27:633–42.

18 Ajmi SC, Advani R, Fjetland L, *et al*. Reducing door-to-needle times in stroke thrombolysis to 13 min through protocol revision and simulation training: a quality improvement project in a Norwegian stroke centre. *BMJ Qual Saf* 2019;28:939–48.

19 Hayashi M, Grover TR, Small S, *et al*. Improving timeliness of hepatitis B vaccine administration in an urban safety net level III NICU. *BMJ Qual Saf* 2021;30:911–9.

20 Lopez Bernal J, Cummins S, Gasparrini A. The use of controls in interrupted time series studies of public health interventions. *Int J Epidemiol* 2018;47:2082–93.

21 Leis B, Frost A, Bryce R, *et al*. Altering standard admission order sets to promote clinical laboratory stewardship: a cohort quality improvement study. *BMJ Qual Saf* 2019;28:846–52.

22 Amarasingham R, Patel PC, Toto K, *et al*. Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. *BMJ Qual Saf* 2013;22:998–1005.

23 Kirkland KB, Homa KA, Lasky RA, *et al*. Impact of a hospital-wide hand hygiene initiative on healthcare-associated infections: results of an interrupted time series. *BMJ Qual Saf* 2012;21:1019–26.

24 Neuburger J, Walker K, Sherlaw-Johnson C, *et al*. Comparison of control charts for monitoring clinical performance using binary data. *BMJ Qual Saf* 2017;26:919–28.

25 Woodcock T, Adeleke Y, Goeschel C, *et al*. A modified Delphi study to identify the features of high quality measurement plans for healthcare improvement projects. *BMC Med Res Methodol* 2020;20:8.

26 Berenholtz SM, Needham DM, Lubomski LH, *et al*. Improving the quality of quality improvement projects. *Jt Comm J Qual Patient Saf* 2010;36:468–73.

27 Portela MC, Pronovost PJ, Woodcock T, *et al*. How to study improvement interventions: a brief overview of possible study types. *BMJ Qual Saf* 2015;24:325–36.

28 Duclos A, Chollet F, Pascal L, *et al*. Effect of monitoring surgical outcomes using control charts to reduce major adverse events in patients: cluster randomised trial. *BMJ* 2020;371:m3840.

29 van Schie P, van Bodegom-Vos L, Zijdeman TM, *et al*. Effectiveness of a multifaceted quality improvement intervention to improve patient outcomes after total hip and knee arthroplasty: a registry nested cluster randomised controlled trial. *BMJ Qual Saf* 2022. doi:10.1136/bmjqs-2021-014472. [Epub ahead of print: 22 Jun 2022].

30 Horwitz LI, Kuznetsova M, Jones SA. Creating a learning health system through rapid-cycle, randomized testing. *N Engl J Med* 2019;381:1175–9.

31 NHS England. A&E Attendances and Emergency Admissions. Available: https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting-times-and-activity/

32 Uspal NG, Rutman LE, Kodish I, *et al*. Use of a dedicated, non-physician-led mental health team to reduce pediatric emergency department lengths of stay. *Acad Emerg Med* 2016;23:440–7.

33 Baker AW, Haridy S, Salem J, *et al*. Performance of statistical process control methods for regional surgical site infection surveillance: a 10-year multicentre pilot study. *BMJ Qual Saf* 2018;27:600–10.

34 Ilieş I, Anderson DJ, Salem J, *et al*. Large-scale empirical optimisation of statistical control charts to detect clinically relevant increases in surgical site infection rates. *BMJ Qual Saf* 2020;29:472–81.

35 van Schie P, van Bodegom-Vos L, van Steenbergen LN, *et al*. Monitoring hospital performance with statistical process control after total hip and knee arthroplasty: a study to determine how much earlier worsening performance can be detected. *J Bone Joint Surg Am* 2020;102:2087–94.