Sense and sensibility: On the diagnostic value of control chart rules for detection of shifts in time series data

Jacob Anhøj

Centre of Diagnostic Investigation, Rigshospitalet, University of Copenhagen, Denmark

E-mail: [jacob@anhoej.net](mailto:jacob@anhoej.net)

Short title: Diagnostic value of control chart rules for time series data

Revision date: 2018-01-04

# Abstract

The aim of this study was to quantify and compare the diagnostic value of statistical process control chart rules for detection of non-random variation in time series data in order to make recommendations for their application under different conditions.

Contro charts are point-and-line graphs showing a measure over time and employing statistical tests for identification of non-random variation.

Numerous tests have been developed to identify non-random variation. In this study, I investigate the diagnostic value of the four Western Electric Rules and the two Anhoej rules using simulated time series data with and without non-random variation introduced as shifts in process centre over time. The primary outcome is likelihood ratios of individual and combined tests. Likelihood ratios are useful measures of a test’s ability to discriminate between the true presence or absence of a specific condition.

This study confirmed that likelihood ratios are useful measures of the diagnostic value of control chart rules. The Western Electric rules 1-4 perform very well with short data series (< 20 data points). For longer data series, the Anhoej rules alone or in combination with the WE rule 1 may be a better choice. Most importantly, the choice of which and how many rules to apply in a given situation should be made deliberately depending on the specific purpose of the SPC analysis and the number of available data points.

# Introduction

Over the past decade or so, the term “improvement science” has gained attention and sparked debate [1]. In healthcare, where I work, improvement science is viewed by many as the natural successor – or rather supplement – to evidence based medicine. If evidence based medicine is about doing the right things then, improvement science is about doing things right, and one is meaningless without the other [2].

In a systematic review The Health Foundation concludes that: “Improvement science is about finding out how to improve and make changes in the most effective way. It is about systematically examining the methods and factors that best work to facilitate quality improvement” [1].

Following this, change and improvement are closely related in that improvement is always the result of change. However, not all change results in improvement. In order to know that improvement is happening, we must be able to measure the quality characteristics of the processes we are trying to improve. As improvement always happens over time, time is an essential part of the analysis, and since measurement is subject to variation whether or not improvement is happening, the aim of the analysis is to discriminate between naturally occurring variation in data over time (noise) and variation that is the result of changes to a process (signal).

Statistical process control (SPC) comprises a set of tools including control charts, which help to distinguish signal (non-random or special cause variation) from noise (random or common cause variation) in time series data.

## Statistical process control charts

SPC charts are point-and-line graphs showing a measure over time and employing statistical tests for identification of non-random variation.

SPC charts are based on the assumption that, if the process in question is random, i.e. containing only noise, the data points will be randomly distributed around the process centre expressed by the mean or median and nearly all of them will appear between limits estimating the random variation inherent in the process. These limits are called control limits and are added as horizontal lines to the chart. Control limits are usually positioned at a distance of +/- 3 times the estimated within sample standard deviation (SD) from the centre line. Consequently, control limits are also referred to as 3-sigma limits. Fig 1A shows an example of a process containing random variation only.

Fig 1. Example control charts. A: random variation. B: Non-random variation caused by a large, possibly transient, shift in data identified by one data point being outside the upper control limit. C: Non-random variation caused by a sustained moderate shift in data identified by an unusually long run of 13 data points below the centre line (WE rule 4 and Anhoej rule 1) and unusually few crossing (Anhoej rule 2). See text for details.

The calculation of sigma limits depends on assumptions regarding the theoretical distribution of data, and many types of control charts exist for different types of measure and count data.

Control chart theory is a wast area, and the interested reader is recommended to consult the specialist literature. Montgomery provides a comprehensive coverage of the subject with emphasis on the application of SPC in engineering and management [3]. A concise introduction to the application of control charts in healthcare is provided by Mohammed [4]. Wheeler gives a more general introduction to the thinking behind and practical use of SPC [5].

Statistical Process Control is not about statistics, it is not about “process-hyphen-control”, and it is not about conformance to specifications. […] It is about the continual improvement of processes and outcomes. And it is, first and foremost, *a way of thinking* with some tools attached. [5, p. 152]

## Testing for non-random variation

Non-random variation may take many forms depending on the nature of its underlying causes.

Originally, SPC charts were designed to identify sudden, larger (> 2 SD) and possibly transient shifts in data (outliers or freaks). For this purpose testing for one or more data points outside the control limits is sufficient (Fig 1B). However, using this test only, other types of non-random variation may go unnoticed for a long time (Fig 1C).

However, since quality improvement often happens gradually or in smaller steps, the focus of this study is the ability to identify persistent shifts in data over time suggesting significant and lasting process improvement or deterioration. For this purpose, a number of additional control chart tests (or rules) have been developed.

### The Western Electric rules

The Best known tests for non-random variation are probably the Western Electric (WE) rules described in the Statistical Quality Control Handbook from 1956 [6]. The WE rules consist of four simple tests that can be applied to control charts by visual inspection to identify unusual, i.e. non-random, patterns in the distribution of data points relative to the control and centre lines:

1. **One** or more points beyound a 3-sigma limit.
2. **Two out of three** successive points beyond a 2-sigma limit (two thirds of the distance between the centre line and the control line).
3. **Four out of five** successive points beyond a 1-sigma limit.
4. **Eight** or more successive points on one side of the centre line.

When using the WE rules, it is generally recommended that control charts should have between 20 and 30 data points. With fewer data points, they loose sensitivity (more false negatives), and with more data points they loose specificity (more false positives).

### The Anhoej rules

Some lesser known tests for non-random variation are the Anhoej rules proposed and validated by me in two previous publications [7, 8]. The Anhoej rules consist of two tests that are based solely on the distribution of data points in relation to the centre line:

1. **Unusually long runs**: A run is one or more consecutive data points on the same side of the centre line. Data points that fall on the centre line do neither break nor contribute to the run. The upper 95% prediction limit for longest run is approximately (rounded to the nearest integer), where is the number of useful data points. For example, in a run chart with 24 data points a run of *more* than 8 would suggest a shift in the process.
2. **Unusually few crossings**: A crossing is when two consecutive data points are on opposite sides of the centre line (ignoring data points on the centre line). In a random process, the number of crossings is expected to follow a binomial distribution, . Thus, in a run chart with 24 useful data points, *fewer* than 8 crossings would suggest that the process is shifting.

Critical values for longest run and number of crossings may be calculated using the formulas provided or looked up in statistical tables [7].

The Anhoej rules were developed to reliably identify persistent shifts in data over time, and while they are useless in detecting transient shifts in data and slower than the WE rules in detecting larger shifts, they have some advantages:

* The Anhoej rules do not depend on sigma limits, and when used with the median as the centre line they are agnostic to the theoretical distribution of data. Therefore they are useful as stand-alone rules with run charts, which are a lot easier to construct than control charts and require pen and paper only.
* The Anhoej rules adapt dynamically to the number of available data points and can be applied to charts with as few as 10 and up to indefinitely many data points without loosing sensitivity and specificity.

### Other rules

Many, many more tests and rule sets have been developed, and in practice there is no limit to the number of ways one could identify unusual patterns in data. However, as the number of applied tests increase, the risk of false positive results increase as well, and especially some some popular rule sets have proven to be at best useless in practice [9, 8].

For these reasons, the decision on which and how many rules to use in a given situation should be made deliberately, preferably before data collection begins, and based on one’s understanding of the processes involved. This study is an attempt to add objectivity and reproducibility to this selection process.

## Likelihood ratios

Traditionally, likelihood ratios are used to evaluate how well clinical tests are able to discriminate between the presence and the absence of specific clinical conditions. In this study, which builds on my previous study [8], I applied likelihood ratios to evaluate how well the WE rules are able to tell random variation from non-random variation in simulated time series.

The questions of interest are: what is the chance that a control chart with a positive test really represents non-random variation?; and what is the chance that a chart with a negative test really represents random variation?

Likelihood ratios are diagnostic measures designed to answer such questions. Assume that a control chart tests positive for non-random variation. A perfect test would mean that the chart would certainly come from a process with non-random variation (true positive, TP). However, some charts with only random variation also test positive (false positive, FP). We therefore correct the true positive rate by the false positive rate by dividing one with the other. Likewise, if a chart tests negative it could be a false negative (FN) rather than a true negative (TN) [10].

A likelihood ratio greater than 1 speaks in favour of the condition being tested for, which in our case is non-random variation in the form of a shift in time series data, while a likelihood ratio less than 1 speaks against a shift. The further a likelihood ratio is from 1, the more or less likely is the presence of a shift.

As a rule of thumb, a positive likelihood ratio (LR+) greater than 10 is considered strong evidence that the condition being tested for is present. A negative likelihood ratio (LR-) smaller than 0.1 is considered strong evidence against the condition [11].

* LR+ = TP / FP
* LR- = FN / TN

Likelihood ratios are closely related to other measures of diagnostic value like sensitivity, specificity, and predictive values. But in contrast to these, likelihood ratios are independent of the prevalence of the conditions in the population of interest.

Thus, likelihood ratios allow us to quantify the probability of shifts in time series and are useful quality characteristics of control chart rules.

## Study aim

The aim of this study was to quantify and compare the diagnostic value of the Western Electric control chart rules for detection of non-random variation in time series data in order to make recommendations for their application under different conditions. In addition, I wanted to compare the performance of the WE rules and the Anhoej rules.

# Methods

For the purpose of this study, I used the R programming language v. 3.4.3 [12] to simulate time series data from random normal numbers with known sample averages and a fixed sample standard deviation (SD = 1). I developed custom functions for calculating likelihood ratios and for testing time series data for non-random variation using the WE zone rules and the Anhoej runs rules. For data manipulation and plotting, I used functions from the tidyverse package v. 1.2.1. [13].

To investigate the effect of series length (number of data points) and shift size (sample average) on the diagnostic value of different rules, 10,000 time series were simulated for each combination of series length (10, 12, …, 40 data points) and shift size (0, 0.2, …, 3.0 SD units). In total 2,560,000 time series were simulated and tested in relation to a fixed set of centre line and sigma limits of 0 +/- 1, 2, and 3 SD.

For each series the proportions of true and false positive and negative results were calculated for selected combinations of tests. Positive and negative likelihood ratios were calculated for a shift size of 2 SD and series lengths of 10, 20, and 40 data points respectively.

The R source code is available as supplementary material to this article [S1 R source].

# Results

Figure 2 shows the the probability of detecting a shift in data over time in relation to the shift size for different series lengths and combinations of control chart rules. Other things being equal, the probability increases with increasing series length and shift size. Also, the probability increases with the number of tests combined.

Fig 2. Sensitivity of control chart rules. we1 = Western Electric (WE) rule1, we12 = WE rule 1 or 2 etc. Anhoej = Anhoej rules. k = series length (number of data points).

The intercepts in Figure 2 corresponding to no shift (SD = 0) indicate the probability of getting false positive tests when no shifts are present in data. The false positive rates are explored further in Figure 3, which shows that the risk of getting false positive tests also increases with increasing series lengths and number of tests combined The exception being the Anhoej rules, which fluctuate but stay below 10%.

Fig 3. False alarm rates of control chart rules when a shift of 2 standard deviation units is present.

Figure 4 illustrates the value of positive and negative test results using likelihood ratios for combinations of series lengths and tests when a shift of 2 SD is present in data. As mentioned, a better test is one with a large range, preferably with LR+ above 10 and LR- below 0.1.

Fig 4. Likelihood ratios of control chart rules when a shift of 2 standard deviation units is present.

Other things being equal, the value of a positive test decreases while the value of a negative test increases with more tests and longer series. For short data series (k = 10), all combinations of rules perform well. Especially, if one needs to be able to exclude shifts in data with high reliability, the combination of WE rules 1-4 appear to be of high diagnostic value. For longer data series, the WE rule 1 and the Anhoej rules appear reliable and robust. One should be cautious when using combinations of multiple rules with longer data series as the values of positive tests decrease significantly (LR+ < 10).

# Discussion

To my knowledge, this is the first study to investigate and quantify the value of the Western Electric rules using likelihood ratios on simulated time series data. Traditionally, the diagnostic properties of SPC charts have been described using the average run length (ARL) measure, which is the average number of data points expected until a rule signals non-random variation. ARLs may be calculated for in-control and out-of-control conditions, where in- and out-of-control means random and non-random variation respectively provided that the theoretical distribution of data and the nature of the out-of-control condition are known [3]. ARLs are comparable to sensitivity and specificity measures traditionally used in health sciences in that they express the chance of having a positive test given the condition being present or absent? Likewise, power functions (Fig 2) and risk of false alarms (Fig 3) are commonly used to describe the value of SPC rules [3].

However, like sensitivity and specificity, ARLs and power functions do not answer the questions of interest for the practical use of a test: given a test result, what is the chance that this result truly reflects the presence or absence of the condition?

Predictive values are alternative measures of diagnostic properties, which are designed to answer these kinds of questions. But unlike likelihood ratios they depend on the prevalence of the condition in the population of interest, which is often unknown or even unknowable.

Likelihood ratios are easy to compute provided one has access to test results together with information on the true presence or absence of the condition of interest for each test. With the ubiquity of personal computers and free, open source software, large datasets are easily simulated for studies like this comparing known outcomes to combinations of different tests under different conditions.

Likelihood ratios are also easy to understand: given a specific test result, how many times more (or less) likely is it that the condition is present? For example, in a control chart with 10 data points that tests positive on WE rule 1, a shift in the order of 2 SD is about 30 times more likely than no shift. If the same chart tests negative on WE rule 1, a shift of 2 SD is about 5 times (1 / 0.2) *less* likely than no shift (Fig 4).

This study has two important limitations. First, the results are not to be extrapolated outside the conditions being tested. Second, since the results come from simulated data series, they should not be taken as exact values rather than indicators of how different conditions affect the diagnostic value of SPC charts.

Regarding extrapolating the results: This study was designed to specifically investigate the effect of series length and combinations of SPC rules in SPC charts when the process centre and spread are known in advance before the introduction of a persistent shift in the process centre. In practice, SPC charts are often used without prior knowledge of process centre and spread. In such cases, the purpose of the chart may actually be to estimate these properties. Also, changes in real life data come in many more forms than persistent shifts of 2 SD.

In my practice (hospital infections, drug usage, procedure compliance, etc.), sudden shifts are less common than long term trends, waves, and individual outliers. Trends and waves are often identified by the Anhoej rules long before any of the WE rules. Outliers are often picked up quickly by WE rule 1. However, to quantify the diagnostic value of SPC charts for other patterns, one must design studies for the specific purpose.

Regarding the use of simulations and in extension of the previous paragraph: No simulated data can truly reflect the properties of real life data, and the results should be interpreted cautiously. Specifically, sudden, persistent shift of exactly 2 SD that does not change over time, as used in my model, may never happen in reality, and my results are merely suggestive of what is expected to happen when data series grow long and more and more tests are applied.

# Conclusions

This study confirmed that likelihood ratios are useful measures of the diagnostic value of SPC rules. The Western Electric rules 1-4 perform very well with short data series (< 20 data points). For longer data series, the Anhoej rules alone or in combination with the WE rule 1 may be a better choice. Most importantly, the choice of which and how many rules to apply in a given situation should be made deliberately depending on the specific purpose of the SPC analysis and the number of available data points.

All other things being equal, the more data points, the fewer rules should be applied if one wants to maintain satisfactory values of positive tests. On the other hand if the purpose of the SPC analysis is to exclude non-random variation with a high degree of certainty, and if the cost of false positive tests is low, more rules may be useful.

# References

1. Health Foundation. Evidence scan: Improvement science. Health Foundation 2011. Available from: <http://www.health.org.uk/publication/improvement-science>.
2. Glasziou P, Ogrinc G, Goodman S. Can evidence-based medicine and clinical quality improvement learn from each other? <doi:10.1136/bmjqs.2010.046524>.
3. Montgomery DC. Introduction to Statistical Quality Control. 6th ed. USA: John Wiley & Sons; 2009.
4. Mohammed MA, Worthington P, Woodall WH. Plotting basic control charts: tutorial notes for healthcare practitioners. <doi:10.1136/qshc.2004.012047>.
5. Wheeler DJ. Understanding Variation – The Key to Managing Chaos. 2nd ed. Knoxville: SPC Press; 2000.
6. Western Electric Company. Statistical Quality Control Handbook. 2nd ed. Easton: Mack Printing Company; 1958.
7. Anhøj J, Olesen AV. Run Charts Revisited: A Simulation Study of Run Chart Rules for Detection of Non-Random Variation in Health Care Processes. <doi:10.1371/journal.pone.0113825>.
8. Anhøj J. Diagnostic Value of Run Chart Analysis: Using Likelihood Ratios to Compare Run Chart Rules on Simulated Data Series. <doi:10.1371/journal.pone.0121349>.
9. Davis RB, Woodall WH. Performance of the Control Chart Trend Rule Under Linear Shift. Journal of Quality Technology 1988;20: 260-262.
10. Attia J. Moving beyond sensitivity and specificity: using likelihood ratios to help interpret diagnostic tests. Australian Prescriber 2003;26: 111-113.
11. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ 2004;329: 168-169.
12. R Core Team. R: A Language and Environment for Statistical Computing. Version 3.4.3 [software]. Available from: <https://www.R-project.org/>
13. Wickham H. tidyverse: Easily Install and Load the ‘Tidyverse’. R package. Version 1.2.1 [software]. Available from: <https://CRAN.R-project.org/package=tidyverse>