How Control Limits Are Worked Out

To identify variations that are not just due to the play of chance and where further investigation is likely to be useful, likelihood methods are used to inform setting upper and lower control limits. This process of assessing the play of chance is common to all areas of science and is termed statistical inference.

There are three main schools of statistical inference (that is, frameworks for assessing the play of chance): frequentist, Bayesian, and likelihood, although there is overlap in that likelihood calculations are used in both frequentist and Bayesian inference, and frequentist inference can be thought of as a particular type of Bayesian inference with a completely non-informative prior.

In any case, frequentist statistical inference (*p*-values and confidence intervals) remains the dominant method of statistical inference in medicine. But it is not well suited to control charts because the data for a particular hospital are looked at multiple times (that is, each time a patient is discharged). This means that *p*-values are difficult to interpret. More specifically, the Type 1 error rate is not constant, but increases with the length of the monitoring period. The probability of eventually signalling an alarm is 1.0 for all sequential tests, so that the Type 1 error rate will eventually be 100 per cent and the Type 2 error rate will eventually be 0 per cent.

Because *p*-values are problematic for control charts, the VLAD uses likelihood methods that have been used for control charts in industry since the 1950s. The statistical characteristics of the charts are defined in terms of the average run length to true or false alarm. Ideally, average run length to false alarm should be long (analogous to a low Type 1 error rate) and the average run length to true alarm should be short (analogous to high statistical power or a low Type 2 error rate). In practice, there is a trade-off. A good choice for a control limit, where the chart is said to signal, is one where the average run length to true alarm is suitably short and the average run length to false alarm is not unacceptably short.

Average run lengths (ARLs) for different control limits can be estimated using simulations, Markov chain analysis, or in certain circumstances by approximating formulae. For the VLADs we used simulations. Briefly, we specified data sets of 10 000 patients (under the null and various alternative hypotheses) and iterated 10 000 times to obtain estimates of the median ARL to true and false alarm.

The control limits are reset each time a trigger point is reached: a hospital that has previously hit a 50 per cent deviation from the average will be flagged, then the control limits are reset. The hospital will be flagged a second time if there is a cumulative run of cases which is again a 50 per cent deviation from the average, starting at the first trigger point.

The control limits indicate when chance has become an unlikely explanation for a hospital's deviation from the state average. At this point it becomes useful to look for causes other than chance, as per the Pyramid Model of investigation. As with any method of statistical inference, we cannot completely rule out chance as an explanation, but when control limits are reached chance is an unlikely explanation.

For more information on the technical background, see

Coory M, Sketcher-Baker K, Duckett S. Using control charts to monitor quality of hospital care with administrative data. *Int J Qual Health Care*, 2008 20: 31–9.

Grigg O, Farewell V, et al. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat Meth Med Res* 2003 12: 147–70.

Sherlaw-Johnson C. A method for detecting runs of good and bad clinical outcomes on variable life-adjusted display (VLAD) charts. *Health Care Management Sc* 2005 8: 61–5.

Steiner S, Cook R, et al. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 2000 1: 441–52.