Universiteit
Leiden
The Netherlands

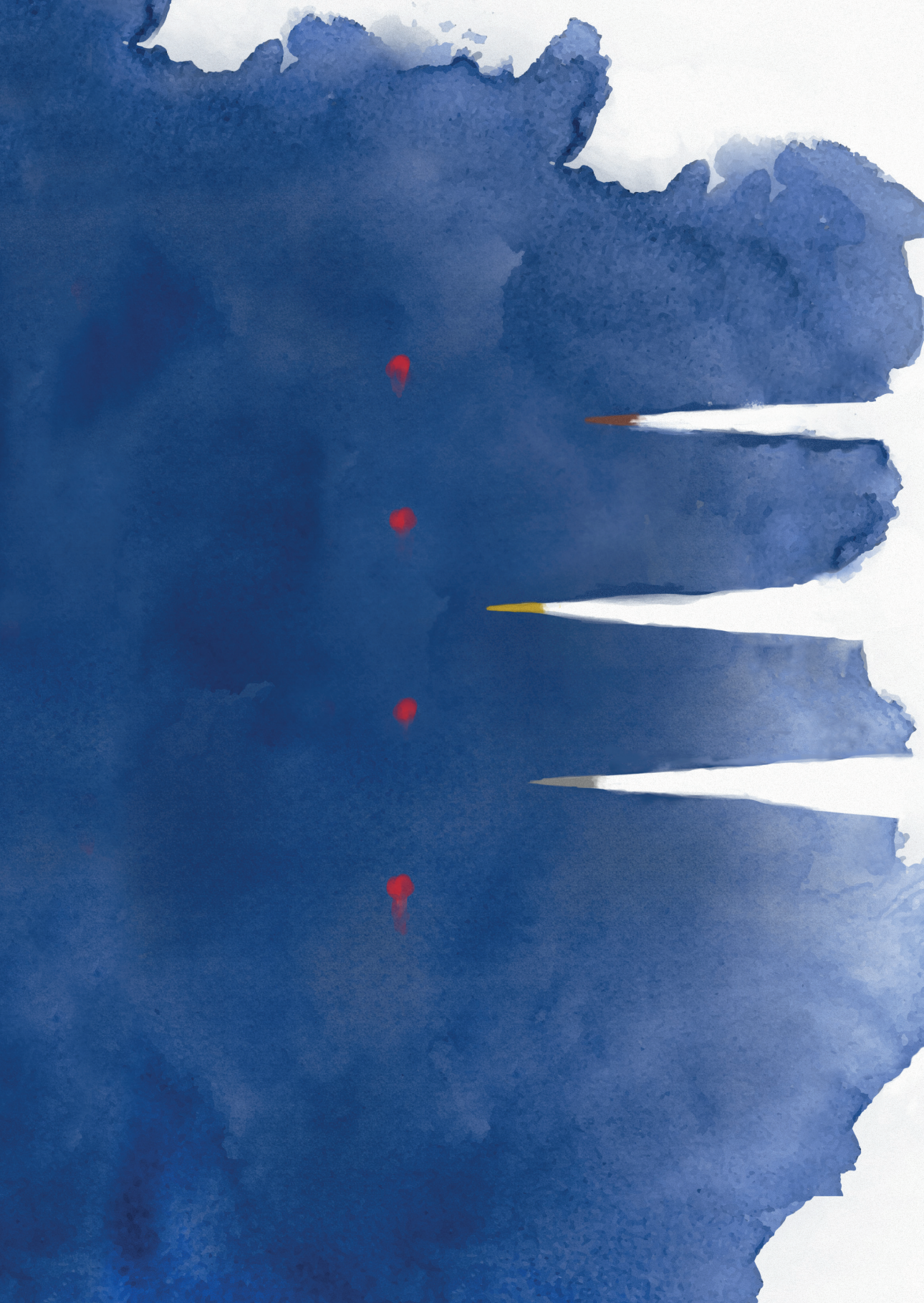
What makes the best performing hospital? the IQ Joint study

Schie, P. van

Citation
Schie, P. van. (2023, November 8). *What makes the best performing hospital?: the IQ Joint study*. Retrieved from https://hdl.handle.net/1887/3656771

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3656771

Note: To cite this publication please use the final published version (if applicable).



Chapter 4

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

Peter van Schie[1,2], Leti van Bodegom-Vos[2], Liza N. van Steenbergen[3] Rob G.H.H. Nelissen[1], Perla J. Marang- van de Mheen[2]

1. Department of Orthopaedics, Leiden University Medical Centre, Leiden, Netherlands;
2. Department of Biomedical Data Sciences, Medical Decision Making, Leiden University Medical Centre, Leiden, Netherlands;
3. Dutch Arthroplasty Register (LROI), 's-Hertogenbosch, Netherlands.

J Bone Joint Surg Am. 2020 Dec 2;102(23):2087-2094

Abstract

Background

Given the low early revision rate after total hip arthroplasty (THA) and total knee arthroplasty (TKA), hospital performance is typically compared using 3 years of data. The purpose of this study was to assess how much earlier worsening hospital performance in 1-year revision rates after THA and TKA can be detected.

Methods

All 86,468 THA and 73,077 TKA procedures performed from 2014 to 2016 and recorded in the Dutch Arthroplasty Register were included. Negative outlier hospitals were identified by significantly higher O/E (observed divided by expected) 1-year revision rates in a funnel plot. Monthly Shewhart p-charts (with 2 and 3-sigma control limits) and cumulative sum (CUSUM) charts (with 3.5 and 5 control limits) were constructed to detect a doubling of revisions (odds ratio of 2), generating a signal when the control limit was reached. The median number of months until generation of a first signal for negative outliers and the number of false signals for non-negative outliers were calculated. Sensitivity, specificity and accuracy were calculated for all charts and control limit settings, using outlier status in the funnel plot as the golden standard.

Results

The funnel plot showed that 13 of 97 hospitals had significantly higher O/E 1-year revision rates and were negative outliers for THA and 7 of 98 hospitals had significantly higher O/E 1-year revision rates and were negative outliers for TKA. The Shewhart p-chart with the 3-sigma control limit generated 68 signals (34 false-positive) for THA and 85 signals (63 false-positive) for TKA. The sensitivity for THA and TKA was 92% and 100% respectively; the specificity was 69% and 51%, respectively; and the accuracy was 72% and 54%, respectively. The CUSUM chart with a 5 control limit generated 18 signals (1 false-positive) for THA and 7 (1 false-positive) for TKA. The sensitivity was 85% and 71% for THA and TKA, respectively; the specificity was 99% for both; and accuracy was 97% for both. The Shewhart p-chart with a 3-sigma control limit generated the first signal for negative outliers after a median of 10 months [Interquartile range (IQR):2 to 18] for THA and 13 months [IQR:5 to 18] for TKA. The CUSUM charts with a 5 control limit generated the first signal after a median of 18 months [IQR:7 to 22] for THA and 21 months [IQR:9 to 25] for TKA.

Conclusion

Monthly monitoring using CUSUM charts with a 5 control limit enables earlier detection of worsening 1-year revision rates with accuracy so that initiatives to improve care can start earlier.

Introduction

Most arthroplasty registries publish annual reports including funnel plots for binary clinical outcomes, with the purpose of monitor hospital performance and providing feedback. Funnel plots are graphical tools to compare outcomes with those of other hospitals and detect hospitals performing significantly better or worse in terms of these outcomes. In orthopaedics, the 1-year revision rate is an important performance indicator to monitor quality of hospital care. Consequences of a revision are dramatic for patients and entail considerable costs. However, due to low event rates for 1-year revision as well as for many orthopaedic performance outcomes, multiple years of outcomes are usually combined in funnel plots to obtain detectable and reliable hospital differences.(1-6) Because arthroplasty registries typically combine 3 years of data, it may take a long time before deteriorating performance is noticed, resulting in late action plans to improve care.(3) Thus, more frequent monitoring of clinical endpoints such as 1-year revision rates is needed, as are reliable and earlier signals if outcomes deteriorate.

Statistical Process Control (SPC) charts such as Shewhart p-charts and Cumulative SUM (CUSUM) charts may offer additional information because the performance is plotted more frequently over time (for example, monthly). Several good clinical studies and the focus to improve the quality of care, led to growing interest in these charts.(7-16) SPC-charts with their control limits can distinguish between an "in-control" process, showing only chance variation within control limits, and an "out-of-control" process showing systematic (special-cause) variation and generating a signal (alert) when the control limit is reached.(17) However, with SPC charts there is a trade-off between the number of false positive and the number of false negative signals, determined by the level at which control limits are set. In practice, minimization of the number of false-positive signals in particular is recommended because they may result in alert and improvement fatigue by clinicians.(18,19)

Various SPC charts are available, but there is uncertainty about which chart and control limit to choose.(20,21) In the present study we opted for Shewhart p-charts and CUSUM charts. The Shewhart p-chart is considered to be accessible, especially with regard to implementation and easy interpretation.(22) However, the CUSUM chart has superior performance in detecting small (<10%) and large (>10%) increases in event rates.(13,22-24) These two SPC charts thus seemed logical candidates to test. The authors of a previous orthopaedic study already described CUSUM charts implementation, but did not address how much earlier a signal was generated or its reliability compared with the more commonly used funnel plot, which seems crucial for these techniques to be accepted in routine clinical practice.(25)

The aim of this study was to assess the extent to which Shewhart p-charts and CUSUM charts enable monitoring such that worsening 1-year revision total hip arthroplasty (THA) or total knee arthroplasty (TKA) rates in Dutch hospitals are detected earlier within a timeframe of 3 years, with good sensitivity, specificity and accuracy, compared with the current method of arthroplasty registries using funnel plots.

Methods

Study design

This observational study used routinely collected data from the nationwide Dutch Arthroplasty Register (Landelijke Registratie Orthopedische Implantaten (LROI)). (6) Data completeness in this register is checked against in-hospital patient records and currently exceeds 98% for primary arthroplasties and 96% for revisions.(26,27)

Study population

All Dutch patients who underwent a primary THA or TKA procedure from January 2014 to December 2016 as recorded in the LROI were included. The following patient characteristics were available: age, sex, body mass index (BMI, kg/m^2), smoking (yes or no), American Society of Anaesthesiologists (ASA) classification (I,II,III-IV), Charnley score (A,B1,B2,C, and not applicable) and diagnosis (osteoarthritis or non-osteoarthritis).(28) Revision within one year (yes or no) was the primary outcome measure (defined as replacement, removal or addition of any component).

Statistical analysis

The between-hospital variation in 1-year revision rates after primary THA and TKA during 2014-2016 was estimated, applying the same method as used by the LROI. For each patient the expected revision risk was calculated using logistic regression analysis, including all patient characteristics described above as independent variables and 1-year revision as the dependent variable. Missing patient characteristic values (<10% for all variables) were imputed with the mean for numeric variables or the mode for categorical variables (meaning that the most frequently occurring category was imputed). All expected revision risks were then summed within a hospital to obtain the aggregated expected number (E) of revisions per hospital. The observed numbers (O) divided by expected numbers were depicted in a funnel plot with 95% control limits. Negative outlier hospitals are those outside the upper limit, meaning that they had significantly higher revision rates than expected given their patient-mix. Positive outlier hospitals are those outside the lower limit, meaning that they had significantly lower revision rates.

Second, the extent to which SPC-methods can generate an earlier signal for deteriorating performance within a 3-year time frame was estimated. Risk-adjusted monthly Shewhart p-charts (with 2 and 3-sigma control limits) and risk-adjusted log-likelihood CUSUM charts (with 3.5 and 5 control limit) for 1-year revisions were constructed to detect an odds ratio of 2, for each hospital across 3 years.(22) Figure 1 shows an example of a Shewhart-p-chart, in which the center line indicates the mean hospital performance and the area between both control limits is where variation is considered random (by chance). A value outside control limits is considered a systematic variation and generates a signal. Usually 2 and 3-sigma control limits are used, with the 2-sigma control limit having a higher likelihood of type a 1 error (false-positive signal) and the 3-sigma control limit having a higher likelihood of a type 2 error (false-negative signal). Figure 2 shows an example of a CUSUM chart with 3.5 and 5 control limits. This chart shows the cumulative performance across patients over a period of time. When the chart-statistic reaches the control limit, a signal is generated and the chart resets to zero. Similarly, the control limits are chosen to balance the likelihood of false-positive and false-negative signals, with 3.5 and 5 most commonly used in practice.(22,25) Appendix I gives a more detailed description of the Shewhart p-chart and CUSUM chart.

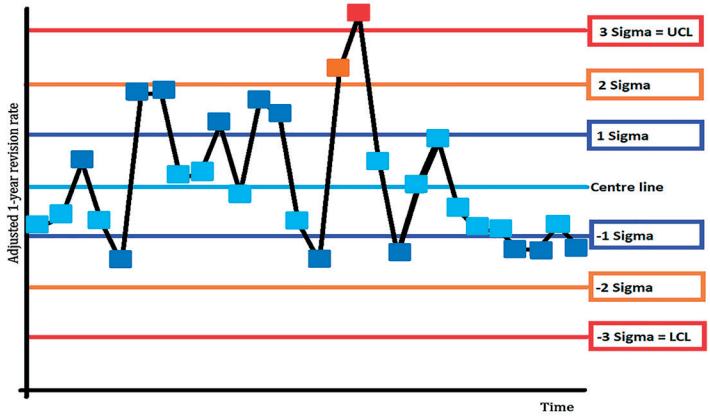


Figure 1 Example of a Shewhart p-chart
See text for explanation of chart.
UCL = upper control limit and LCL = lower control limit.

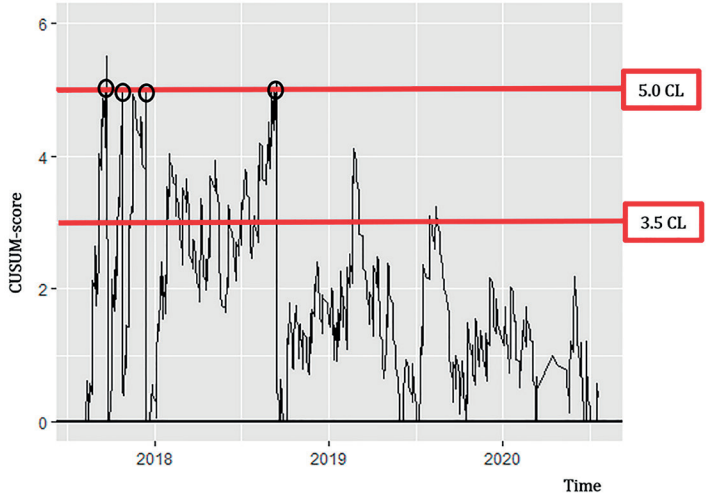


Figure 2 Example of a CUSUM chart
A hypothetical CUSUM chart with 3.5 and 5 control limit (CL). The chart resets to 0 after the 5 control limit is reached. In this example, 4 signals are generated, and the hospitals shows an improvement for this outcome over time. See text for further explanation of chart.

For both charts and control limit settings, we calculated the median number of months needed to generate the first signal for negative outlier hospitals and the number of false signals for other hospitals. Furthermore, we calculated the signals missed for negative outlier hospitals. Additionally, sensitivity, specificity and accuracy for both charts and control limit settings were calculated within the 3-year time frame using the negative outlier status of a hospital in the funnel plots as the ''golden" standard. The accuracy for correctly classifying a hospital was defined as:

$$\frac{(\text{Number of true positive classified hospitals} + \text{Number of true negative classified hospitals})}{(\text{Total number of hospitals})} \text{ X } 100\%.$$

Analyses were performed using SPSS version 25 (IBM). The LUMC Medical Ethical Committee considered the study exempt for ethical approval under Dutch law (CME, G18.140).

Results

The study included 86,468 primary THA procedures from 97 hospitals and 73,077 primary TKA procedures from 98 hospitals. The rate of missing data was <4% for

all variables, except for smoking (<10%). On the patient-level, the average 1-year revision rate was 1.8% for THA and 1.2% for TKA. On hospital-level, the median revision rate was 1.6% (interquartile range (IQR):1.0 to 2.3) for THA and 1.1% (IQR:0.7 to 1.6) for TKA (Table 1).

Table 1 Distribution of patient characteristics and outcomes in participating hospitals

	THA (*n*=97 hospitals)		TKA (*n*=98 hospitals)	
	Median (IQR)	**Range**	**Median (IQR)**	**Range**
Procedures (*n*)	759 (526-1173)	2-2502	699 (463-938)	9-1998
Mean age (years)	69.3 (67.8-70.1)	50.6-71.8	68.8 (67.4-69.7)	56.5-72.2
Gender, female (%)	66.1 (63.3-68.0)	0.0-74.1	65.2 (61.9-67.8)	8.3-100.0
Mean BMI (kg/m^2)	27.3 (27.0-27.8)	25.9-28.6	29.8 (29.3-30.4)	20.5-31.0
Smoking (%)	13.2 (10.7-15.2)	0.0-27.9	9.8 (8.4-11.8)	1.0-20.5
ASA classification (%)				
• ASA I	17.4 (14.2-21.4)	3.3-100	11.8 (9.8-16.0)	3.8-54.5
• ASA II	65.0 (59.8-70.4)	0.0-96.7	68.7 (63.7-73.6)	42.5-91.6
• ASA III-IV	15.6 (11.5-20.4)	0.0-40.1	16.6 (10.8-21.8)	0.0-50.6
Charnley score* (%)				
• A	49.3 (43.7-53.9)	23.7-78.2	45.3 (35.6-52.4)	13.1-100.0
• B1	27.8 (22.9-33.4)	3.6-50.7	33.0 (27.3-40.3)	0.0- 57.8
• B2	20.1 (18.1-22.9)	4.7-28.3	19.4 (16.2-21.5)	0.0-28.0
• C	1.9 (1.0-3.3)	0.0-12.2	2.3 (1.1-4.2)	0.0-17.4
Diagnosis (%)				
• OA	87.1 (83.5-90.8)	42.2-100.0	96.6 (95.5-97.9)	58.6-100.0
• Non-OA**	12.9 (9.3-16.5)	0.0-57.8	3.4 (2.1-4.5)	0.0-41.4
1-year revisions (%)	1.6 (1.0-2.3)	0.0-7.0	1.1 (0.7-1.6)	0.0-16.7

The values under ''Median (IQR)" indicate the mean or the percentage of the median hospital. The values under ''Range" indicate the highest and lowest means or percentages among the hospitals. *The Charnley score was used to evaluate comorbidity in relation to levels of activity. **All diagnoses except osteoarthritis (fracture, osteonecrosis, rheumatoid arthritis, inflammatory arthritis, etc).

Outlier hospitals

Based on 3-year funnel plots, 13 hospitals performing THA were negative outliers with a median O/E (observed divided by expected) ratio of 1.9 (IQR:1.5 to 2.5) compared with 0.9 (IQR:0.5 to 1.1) for the other hospitals. For TKA, there were 7 negative outliers with a median O/E ratio of 2.3 (IQR:2.3 to 2.8) compared with 0.8 (IQR:0.6 to 1.2) for the other hospitals (Table 2 and Appendices II and III; red dots). Two hospitals were negative outliers for both THA and TKA. Eighteen hospitals were positive outliers for THA with a median O/E ratio of 0.4 (IQR:0.3 to 0.5) and 14 hospitals were positive outliers for TKA with a median O/E ratio of 0.3 (IQR:0.2 to 0.5) (Appendices II and III; green dots).

Table 2 Outlier hospitals with significantly more revisions than expected during 2014-2016

Hospital	Negative outliers	
	THA (*n*=13 hospitals)	TKA (*n*=7 hospitals)
	2014-2016 O/E	2014-2016 O/E
4	1.4	
6	1.5	
9	2.5	2.2
13	1.5	
14	1.4	
21	2.1	
28	1.8	
33	2.1	
37	1.6	
35		2.3
39		2.0
41		2.3
52	1.9	
87	2.7	2.8
88	3.3	
89		2.7
90	2.6	
95		13.3
Median (IQR) negative outliers	1.9 (1.5-2.5)	2.3 (2.3-2.8)
Median (IQR) all other Dutch hospitals	0.9 (0.5-1.1)	0.8 (0.6-1.2)

An O/E ratios is provided only for negative outlier during the 3-year period.

Earlier signals compared with false signals using two SPC methods

I. Shewhart p-chart

For THA, 195 signals of worsening performance were generated for 70 hospitals at the 2-sigma (similar to 2 standard deviation in hypothesis testing) control limit with all 13 negative outlier hospitals alerted, but also 57 hospitals incorrectly alerted (sensitivity 100%, specificity 32%, accuracy 41%). At the 3-sigma control limit, 68 signals were generated for 38 hospitals, with 12 negative outlier hospitals alerted (sensitivity 92%, specificity 69%, accuracy 72%). At 3-sigma, the first signal for negative outliers was generated after a median of 10 months (IQR:2 to 18), which should be considered against 34 false-positive signals for other hospitals. For 1 negative outlier hospital, no signal was generated. More than 1 signal was generated for 9 negative outliers and 7 other hospitals (table 3).

For TKA, 214 signals were generated for 85 hospitals at the 2-sigma control limit, with all 7 negative outlier hospitals alerted (sensitivity 100%, specificity

14%, accuracy 20%) and 85 signals were generated for 52 hospitals at 3-sigma (sensitivity 100%, specificity 51%, accuracy 54%). At 3-sigma, the first signal for negative outliers was generated after a median of 13 months (IQR:5 to 18), which should be considered against 63 false-positive signals. All negative outlier hospitals were alerted. More than 1 signal was generated for 6 negative outliers and 14 other hospitals (table 3).

II. CUSUM chart

For THA, 33 signals were generated for 16 hospitals at 3.5 control limit (sensitivity 85%, specificity 94%, accuracy 93%) and 18 signals were generated for 12 hospitals at 5 control limit, correctly alerting 11 of 13 negative outliers (sensitivity 85%, specificity 99%, accuracy 97%). At the 5 control limit, the first signal for negative outliers was generated after a median of 18 months (IQR:7 to 22), which should be considered against one false-positive signal for other hospitals. Two negative outlier hospitals were not alerted. More than 1 signal was generated for 4 negative outliers and none for other hospitals (table 3).

For TKA, 16 signals were generated for 12 hospitals at the 3.5 control limit (sensitivity 71%, specificity 92%, accuracy 91%) and 7 signals were generated for 6 hospitals at 5 control limit with 5 of the 7 outliers correctly alerted (sensitivity 71%, specificity 99%, accuracy 97%). At the 5 control limit, the first signal for negative outliers was generated after a median of 21 months (IQR:9-25) which should be considered against one false-positive signal. Two negative outliers were not alerted. More than 1 signal was generated for 1 negative outlier and none for the other hospitals (table 3).

Table 3 Characteristics of statistical process control charts

	THA (97 hospitals; 13 outliers)				TKA (98 hospitals; 7 outliers)			
	Shewhart p-chart		CUSUM chart		Shewhart p-chart		CUSUM chart	
	2-sigma	**3-sigma**	**3.5 C.L.**	**5 C.L.**	**2-sigma**	**3-sigma**	**3.5 C.L.**	**5 C.L.**
Total number of signals	195	68	33	18	214	85	16	7
• Number of good signals* (%)	76 (39%)	34 (50%)	27 (82%)	17 (94%)	36 (17%)	22 (26%)	9 (56%)	6 (86%)
• Number of false signals** (%)	119 (61%)	34 (50%)	6 (18%)	1 (6%)	178 (83%)	63 (74%)	7 (44%)	1 (14%)
Total number of hospitals with signal	70	38	16	12	85	52	12	6
• Number of good signals^{+} (%)	13 (19%)	12 (32%)	11 (69%)	11 (92%)	7 (8%)	7 (13%)	5 (42%)	5 (83%)
• Number of false signals^{++} (%)	57 (81%)	26 (68%)	5 (31%)	1 (8%)	78 (92%)	45 (87%)	7 (58%	1 (17%)
Total number of hospitals with >1 signal	43	16	9	4	54	20	2	1
• Negative outliers with >1 signal (%)	13 (30%)	9 (56%)	8 (89%)	4 (100%)	7 (13%)	6 (30%)	2 (100%)	1 (100%)
• Other hospitals with >1 signal (%)	30 (70%)	7 (44%)	1 (11%)	0 (0%)	47 (87%)	14 (70%)	0 (0%)	0 (0%)
First signal for outliers (months + IQR)	5 [2-10]	10 [2-18]	16 [4-18]	18 [7-22]	5 [3-13]	13 [5-18]	15 [7-22]	21 [9-25]
Sensitivity	100%	92%	85%	85%	100%	100%	71%	71%
Specificity	32%	69%	94%	99%	14%	51%	92%	99%
Accuracy^	41%	72%	93%	97%	20%	54%	91%	97%

*Number of signals generated for negative outliers. **Number of signals generated for other hospitals. $^{+}$Number of negative outliers that received a signal. $^{++}$Number of other hospitals that received a signal. ^Accuracy for correctly classifying a hospital (number of true-positive classified hospitals + number of true-negative classified hospitals) / (total number of hospitals) x 100%.

Discussion

Most arthroplasty registers report revision rates after THA and TKA, as well as differences between hospitals using funnel plots to detect hospitals with significantly worse performance than others (negative outlier hospitals).(1-6) Because of the low event rate, this is typically done by combining multiple years of data. The present study shows that monthly monitoring of THA and TKA revision rates using CUSUM charts with the 5 control limit detected worsening performance earlier than did the funnel plots, with good accuracy within a 3-year time frame; the first signal for negative outliers was generated at a median of 18 months for THA and 21 months for TKA. Using CUSUM charts to monitor deteriorating patterns for revision rates thus makes it possible to initiate improvement initiatives earlier rather than waiting for the results to appear in the funnel plot after 3 years.

Some limitations of this study should be noted. First, given the LROI privacy protocol, we could not confirm that the the negative outlier hospitals were actually being audited for worse performance by the Dutch Orthopaedic Association. However, since we and the Dutch Orthopaedic Association used both the same data source and the same statistical code to generate the outlier status in a funnel plot, it seems highly unlikely that our identification of negative outliers would have differed. Second, the number of months that the signal generation by the CUSUM chart was earlier than the signal generation by the funnel plot may not be directly generalizable to other countries, but it is likely that the differences in favor of the CUSUM chart are generalizable, particularly because the benefits have been shown previously.(25,29) Third, there is a possibility of insufficient adjustment for differences in patient-mix between hospitals because we could control only for those patient characteristics that were collected. However, this limitation would be expected to be similar for both the funnel plot and SPC charts, so it seems unlikely that it affected our conclusions regarding which method is best to detect changing performance. Fourth, registry data are self-reported by orthopaedic surgeons who may not register all revisions, but given the completeness of the Dutch register we do not believe that this affected our results considerably.(26,27) Fifth, surgeons may postpone revisions, resulting in hospitals having low 1-year revision rates but higher revision rates beyond one year. Therefore, using registries to monitor performance reflects daily practice as well as physician's behaviour. We recommend monitoring long-term revision rates (such as at 2 to 5 years) as a balancing measure to check for such occurrences.

There are few examples in orthopaedics of using SPC-methods for quality improvement.(25,29) The Scottish Arthroplasty Project reported using CUSUM chart with the 5 control limit to identify hospital variation in complications.(25) When a signal was

generated by exceeding the control limit, surgeons had to submit a review of their complications for assessment by the Scottish Orthopaedic Association. A reduction in complication rates was observed over the last years since the introduction of this quality improvement strategy. However, due to lack of a control group a causal relationship between CUSUM chart implementation and reduction in complications could not be demonstrated, as a general time trend due to other factors could have been responsible for this reduction. To our knowledge, no empirical studies have been performed to investigate how much earlier worsening performance could be detected using SPC methods before that worsening appeared as an outlier in funnel plots. These empirical data from daily practice are what the present study adds to the simulations in previous studies that already pointed to more rapid detection of small changes in performance with CUSUM charts. This is relevant for (for example) registries and scientific associations deciding whether to implement such SPC charts in their hospital feedback to initiate quality improvement.(22) By examining patient outcomes over time, SPC charts were able to detect deviating performance even when performance had been "in control" in the past, which may be difficult for a funnel plot to detect, because it uses the average outcome over a 3 year period. In addition, the CUSUM chart can be employed to examine the effect of quality improvement initiatives. Using SPC charts thus seems to add relevant information to act upon in daily practice and improve quality of care.

Similar to our study, another study showed the possibility of earlier detection of surgical site infections (SSI) outbreaks using SPC charts.(30) The Shewhart p-charts and exponential weighted moving average (EWMA) charts (another SPC chart) in that study both detected 8 out of 10 SSI outbreaks (including all 4 orthopaedic related outbreaks). In each case, a signal was generated prior to signal generation by the traditional detection methods, with a specificity of 70% and 90% for the Shewhart-p-chart and EWMA chart, respectively.

The English hospital mortality surveillance system generates CUSUM charts, on monthly-collected hospital administrative data.(7,8) After implementation of CUSUM charts, the average risk of death fell by 61% in the 9 months following a signal and reached the level of expected risk within 18 months.(7) It could be that signals were triggered by random variation and subsequent reductions occurred due to regression to the mean (a phenomenon in which extreme outcomes are likely to be followed by a fall in subsequent outcomes).(31) This may overestimate the effect of a signal. However, findings could also be explained by hospitals monitoring their own performance and taking action before a signal is generated.(7)

In contrast, one study showed no improvement in incidence rates of ward-acquired methicillin-resistant Staphylococcus aureus (MRSA) after implementation of monthly SPC feedback (with or without diagnostic tools).(14)

In 2017, the Dutch Orthopaedic Association, in collaboration with the LROI started to identify negative outlier hospitals using funnel plots including 3 years of data, with the aim of providing insight into their clinical practice compared with other Dutch hospitals.(32) This study showed that SPC charts should be included as additional hospital feedback information to provide earlier alerts if performance deteriorates and to provide hospitals with the opportunity to introduce quality improvement initiatives earlier to improve patient care. Further research must be performed to determine whether using SPC charts in daily practice will in fact initiate more quality improvement initiatives, which is the focus of an ongoing randomised controlled trial.(33) Crucial for the effectiveness is that professionals can trust the signals from the SPC chart to be reliable, as was demonstrated by data in this study, and therefore known that they warrant subsequent actions to be taken. Using SPC charts allows initiatives to be introduced earlier than is possible if hospitals wait to become an outlier in a funnel plot.

References

1. Swedisch Hip Arthroplasty Register. Annual Report 2017. https://registercentrum.blob.core.windows.net/shpr/r/Eng_Arsrapport_2017_Hoftprotes_final-Syx2fJPhMN.pdf.
2. Swedisch Knee Arthroplasty Register. Annual Report 2018. http://www.myknee.se/en/.
3. Dutch Arthroplasty Register (LROI). Annual report 2018. www.lroi-report.nl.
4. Danish Hip Arthroplasty Register. Annual Report 2018. http://danskhoftealloplastikregister.dk/en/dhr/.
5. Australian Orthopaedic Association National Joint Replacement Registry. Annual report 2018. https://aoanjrr.sahmri.com/annual-reports-2018.
6. van Schie P, van Steenbergen LN, van Bodegom-Vos L, Nelissen R, Marang-van de Mheen PJ. Between-Hospital Variation in Revision Rates After Total Hip and Knee Arthroplasty in the Netherlands: Directing Quality-Improvement Initiatives. J Bone Joint Surg Am. 2019.
7. Cecil E, Bottle, A., Esmail, A., Wilkinson, S., Vincent, C., Aylin, P. P.. Investigating the association of alerts from a national mortality surveillance system with subsequent hospital mortality in England: an interrupted time series analysis. BMJ quality & safety. 2018;27(12):965-73.
8. Cecil E, Wilkinson, S., Bottle, A., Esmail, A., Vincent, C., Aylin, P. P.. National hospital mortality surveillance system: a descriptive analysis. BMJ quality & safety. 2018;27(12):974-81.
9. Dyrkorn OA, Kristoffersen, M., Walberg, M.. Reducing post-caesarean surgical wound infection rate: an improvement project in a Norwegian maternity clinic. BMJ quality & safety. 2012;21(3):206-10.
10. Benning A, Ghaleb, M., Suokas, A., Dixon-Woods, M., Dawson, J., Barber, N., Franklin, B. D., Girling, A., Hemming, K., Carmalt, M., Rudge, G., Naicker, T., Nwulu, U., Choudhury, S., Lilford, R.. Large scale organisational intervention to improve patient safety in four UK hospitals: mixed method evaluation. BMJ. 2011;342:d195.
11. Nicolay CR, Purkayastha, S., Greenhalgh, A., Benn, J., Chaturvedi, S., Phillips, N., Darzi, A.. Systematic review of the application of quality improvement methodologies from the manufacturing industry to surgical healthcare. Br J Surg. 2012;99(3):324-35.
12. Woodall WH, Fogel, S.L., Steiner, S.H.. The monitoring and improvement of surgical-outcome quality. J Qual Technology. 2015(47):383-99.
13. Grigg O, Farewell, V.. An overview of risk-adjusted charts. J R Stat Soc Ser A Stat Soc. 2004;167:523-39.
14. Curran E, Harper, P., Loveday, H., Gilmour, H., Jones, S., Benneyan, J., Hood, J., Pratt, R.. Results of a multicentre randomised controlled trial of statistical process control charts and structured diagnostic tools to reduce ward-acquired meticillin-resistant Staphylococcus aureus: the CHART Project. J Hosp Infect. 2008;70(2):127-35.
15. Cohen ME, Liu, Y., Ko, C. Y., Hall, B. L.. Improved Surgical Outcomes for ACS NSQIP Hospitals Over Time: Evaluation of Hospital Cohorts With up to 8 Years of Participation. Ann Surg. 2016;263(2):267-73.
16. Hollesen RVB, Johansen, R. L. R., Rorbye, C., Munk, L., Barker, P., Kjaerbye-Thygesen, A.. Successfully reducing newborn asphyxia in the labour unit in a large academic medical centre: a quality improvement project using statistical process control. BMJ quality & safety. 2018;27(8):633-42.
17. Woodall WH. The use of control charts in health-care and public health surveillance. J Qual Technology. 2006;26:89-104.

18. Anhoj J, Hellesoe, A.B.. The problem with red, amber, green: the need to avoid distraction by random variation in organisational performance measures. BMJ quality & safety. 2017;26(1):81-4.
19. Schmidtke KA, Watson, D. G., Vlaev, I.. The use of control charts by laypeople and hospital decision-makers for guiding decision making. Q J Exp Psychol (Hove). 2017;70(7):1114-28.
20. Schmidtke KA, Poots, A. J., Carpio, J., Vlaev, I., Kandala, N. B., Lilford, R. J.. Considering chance in quality and safety performance measures: an analysis of performance reports by boards in English NHS trusts. BMJ quality & safety. 2017;26(1):61-9.
21. Thor J, Lundberg, J., Ask, J., Olsson, J., Carli, C., Harenstam, K. P., Brommels, M.. Application of statistical process control in healthcare improvement: systematic review. Qual Saf Health Care. 2007;16(5):387-99.
22. Neuburger J, Walker, K., Sherlaw-Johnson, C., van der Meulen, J., Cromwell, D. A.. Comparison of control charts for monitoring clinical performance using binary data. BMJ quality & safety. 2017;26(11):919-28.
23. Montgomery DC. Introduction to Statistical Quality Control. 6th ed New York: Wiley & Sons, 2009.
24. Spiegelhalter D, Sherlaw-Johnson C, Bardsley M. Statistical methods for healthcare regulation: rating, screening and surveillance. J R Stat Soc Ser A Stat Soc. 2012;175:1-47.
25. Macpherson GJ, Brenkel, I. J., Smith, R., Howie, C. R.. Outlier analysis in orthopaedics: use of CUSUM: the Scottish Arthroplasty Project: shouldering the burden of improvement. J Bone Joint Surg Am. 2011;93 Suppl 3:81-8.
26. van Steenbergen LN, Denissen GA, Spooren A, van Rooden SM, van Oosterhout FJ, Morrenhof JW, et al. More than 95% completeness of reported procedures in the population-based Dutch Arthroplasty Register. Acta Orthop. 2015;86(4):498-505.
27. *LROI website. Completeness of Registering Hospitals and Completeness of Registered Arthroplasties in the LROI Based on the Hospital Information System in 2016,* http://www.lroi-rapportage.nl/data-quality-coverage-and-completeness, accessed febr 2019.
28. Jasper LL, Jones, C. A., Mollins, J., Pohar, S. L., Beaupre, L. A.. Risk factors for revision of total knee arthroplasty: a scoping review. BMC Musculoskelet Disord. 2016;17:182.
29. Biau DJ, Milet A, Thevenin F, Anract P, Porcher R. Monitoring surgical performance: an application to total hip replacement. J Eval Clin Pract. 2009;15(3):420-4.
30. Baker AW, Haridy, S., Salem, J., Ilies, I., Ergai, A.O., Samareh, A., Andrianas, N., Benneyan, J.C., Sexton, D.J., Anderson, D.J.. Performance of statistical process control methods for regional surgical site infection surveillance: a 10-year multicentre pilot study. BMJ quality & safety. 2018;27(8):600-10.
31. Marang-van de Mheen PJ, Abel, G. A., Shojania, K. G.. Mortality alerts, actions taken and declining mortality: true effect or regression to the mean? BMJ quality & safety. 2018.
32. CommisionQuality. Commision Quality (Collaboration between Dutch Orthopaedic Association and Dutch Arthroplasty Registry). Protocol: Quality procedure, 2017. www.orthopeden.org.
33. ClinicalTrial.gov. https://clinicaltrials.gov/ct2/show/NCT04055103?term=Arthroplasty&cntry=NL&city=Leiden&draw=2&rank=2.

Supplemental data

Appendix I Description of the Shewhart-p-chart and CUSUM-chart.

Introduction and theory

In recent years, Statistical Process Control (SPC)-methods have gained growing interest in healthcare as a method to monitor quality of care and evaluate quality improvement initiatives.[1-3] In this study we opted for Shewhart-p-charts and CUSUM-charts, but other types of SPC-charts exist e.g. the exponentially weighted moving average (EMWA)-chart, and the g-chart. The general theory behind SPC-charts is that random variation is inherent in all processes, caused by common causes. A process is in-control when there is only random variation (common cause variation). However, situations may arise that cause a process to become out-of-control, due to the particular causes of this situation (special cause variation). SPC-charts with a control limit intend to distinguish between common cause variation and special cause variation, with the intention to investigate for possible causes when special cause variation is detected. The advantage of a SPC-chart over, for example, the funnel-plot where data of multiple years are taken together, is that the time variable is added by plotting the outcomes over time, showing the possible effect of changes in practice nearly real-time rather than that these remain hidden in the pooled data over a longer period.

Shewhart-p-chart

The Shewhart-p-chart generally uses a standard format, as shown in Figure 1 in the manuscript. The x-axis indicates time, e.g. weeks, months or quarters. Because it is a p-chart, the y-axis displays a proportion of a certain outcome (e.g. revision rate). The chart thus presents e.g. the weekly proportion of patients with a certain outcome over time. Three horizontal lines are depicted: the center line (CL), the upper control limit (UCL) and the lower control limit (LCL). The center line represents the average or median level of performance over a certain period. Given the random variation, an outcome will usually vary across this central tendency line and remain within the control limits, assuming that the long-term rate of that outcome does not change and will only present some random variation over time. Usually 2 and 3-sigma control limits are used, with a 2-sigma control limit having higher likelihood of type 1 error (false positive signal) and a 3-sigma control limit a higher likelihood of type 2 error (false negative signal). Control limits are computed statistically based on probability distributions such as the Gaussian ('normal' distribution), similar to hypothesis testing. In general, 95% of data will fall within ±2 standard deviations (SD) or 2

sigma and 99,7% within ±3 SD or 3 sigma. Values that fall outside the chosen upper and lower control limits exceed that range of most values, making it unlikely that this is due to random variation but rather reflects a true difference, in this study indicating that the revision rate has doubled.

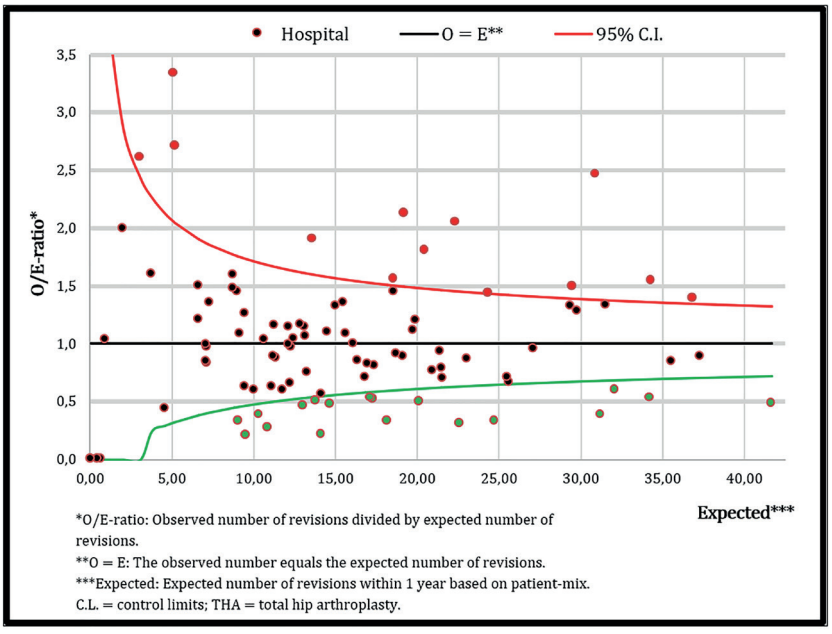
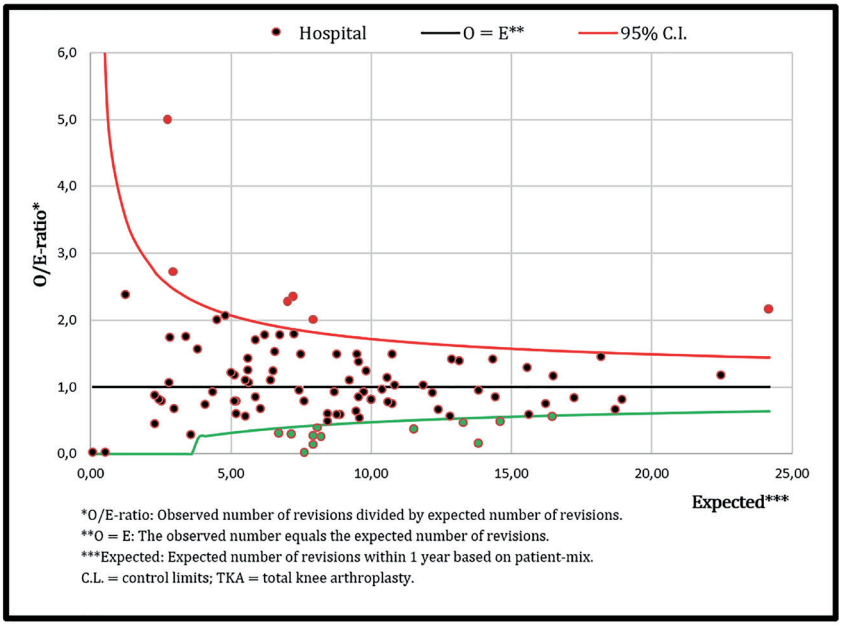
CUSUM-chart

Where the Shewhart-p-chart works with aggregated data over weeks, months of quarters, the CUSUM-chart uses every patient to plot the graph chronologically. For each patient undergoing an operation the expected chance on e.g. a revision is calculated based on certain patient characteristics and compared with the observed outcome, whether this patient has a revision or not. The line in the CUSUM-chart declines when ''good" outcomes occur (e.g. no revisions) representing better performance than expected and increases when 'unfavorable' outcomes occur (e.g. revisions) representing worse performance than expected (Figure 2 in manuscript). When performance is in balance, an increase in the line in the CUSUM-chart because of an ''unfavorable" outcome is counteracted by many small decreases in the line in the CUSUM-chart resulting from ''good" outcomes. Regardless of the use of the CUSUM-chart for detecting a better or worse outcome, the baseline always indicates that a surgeon or hospital is performing as expected. The more the CUSUM-chart line drifts away from the baseline, the more this proves that a surgeon or hospital is performing better or worse than expected. A signal for better or worse performance is generated when the control limit is exceeded, in this case to detect a doubling of the revision rate. Similar to the Shewhart-p-chart, control limit setting of CUSUM-charts allow us to balance the risk of false positive and false negative signals. The control limits in CUSUM-charts are most commonly set at 3.5 or 5, with the 3.5 having higher likelihood of false-positive signals but the 5 having higher likelihood of false negative signals.[4,5] The CUSUM-chart is reset to zero when the control limit is reached. For a detailed description of the Shewhart-p-chart and CUSUM-chart formulas, we refer to Neuburger et al and Benneyan.[4,6]

Literature

1. Benning A, Ghaleb, M., Suokas, A., Dixon-Woods, M., Dawson, J., Barber, N., Franklin, B. D., Girling, A., Hemming, K., Carmalt, M., Rudge, G., Naicker, T., Nwulu, U., Choudhury, S., Lilford, R.. Large scale organisational intervention to improve patient safety in four UK hospitals: mixed method evaluation. *BMJ.* 2011;342:d195.
2. Nicolay CR, Purkayastha, S., Greenhalgh, A., Benn, J., Chaturvedi, S., Phillips, N., Darzi, A.. Systematic review of the application of quality improvement methodologies from the manufacturing industry to surgical healthcare. *Br J Surg.* 2012;99(3):324-335.
3. Woodall WH, Fogel, S.L., Steiner, S.H.. The monitoring and improvement of surgical-outcome quality. *J Qual Technology.* 2015(47):383-399.
4. Neuburger J, Walker, K., Sherlaw-Johnson, C., van der Meulen, J., Cromwell, D. A.. Comparison of control charts for monitoring clinical performance using binary data. *BMJ quality & safety.* 2017;26(11):919-928.
5. Macpherson GJ, Brenkel, I. J., Smith, R., Howie, C. R.. Outlier analysis in orthopaedics: use of CUSUM: the Scottish Arthroplasty Project: shouldering the burden of improvement. *J Bone Joint Surg Am.* 2011;93 Suppl 3:81-88.
6. Benneyan JC, Lloyd RC, Plsek PE. Statistical process control as a tool for research and healthcare improvement. *Qual Saf Health Care.* 2003;12(6):458-464.

Appendix II. Funnel-plot of between-hospital variation in 1-year revisions after THA during 2014-2016



Appendix III. Funnel-plot of between-hospital variation in 1-year revisions after TKA during 2014-2016