PATIENT- AND SURGEON-ADJUSTED CONTROL CHARTS FOR MONITORING INDIVIDUAL PERFORMANCE

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-004046
Article Type:	Research
Date Submitted by the Author:	16-Sep-2013
Complete List of Authors:	Maruthappu, Mahiben; University of Oxford, Carty, Matthew; Harvard Medical School, Lipsitz, Stuart; Harvard Medical School, Wright, John; Harvard Medical School, Orgill, Dennis; Harvard Medical School, ; Brigham and Women's Hospital, Division of Plastic Surgery Duclos, Antoine; Université de Lyon,
Primary Subject Heading:	Surgery
Secondary Subject Heading:	Public health
Keywords:	Orthopaedic & trauma surgery < SURGERY, SURGERY, PUBLIC HEALTH

PATIENT- AND SURGEON-ADJUSTED CONTROL CHARTS FOR MONITORING PERFORMANCE

Mahiben Maruthappu, M.A., B.M. B.Ch.[1*]

Matthew J. Carty, M.D.[2*]

Stuart R. Lipsitz, Sc.D., Ph.D.[3]

John Wright, M.D.[4]

Dennis Orgill, M.D., Ph.D.[5]

Antoine Duclos, M.D., Ph.D.[6]

[1]Joan & Richard Doll Scholar, Green Templeton College, University of Oxford, UK

[2]Assistant Professor of Surgery, Harvard Medical School, USA

[3]Professor of Medicine, Harvard Medical School, USA

[4]Assistant Professor of Surgery, Harvard Medical School, USA

[5]Professor of Surgery, Harvard Medical School, USA; Vice-Chairman of Surgery, Brigham and Women's Hospital, USA

[6]Associate Professor of Public Health, Université de Lyon, France

* Both authors equally contributed to the development of this manuscript

Please address correspondence to:

Mahiben Maruthappu

Green Templeton College

Oxfordshire

OX2 6HG

maruthappu@post.harvard.edu

Financial Disclosure and Products Statement

None of the participating authors has a conflicting financial interest related to the work detailed in this manuscript, nor do any of the authors maintain a financial stake in any product, device or drug cited in this report. Ethical and institutional board approval was granted prior to conducting this study. Information on data used in this report is available upon request.

ABSTRACT

Objectives: To determine whether an innovative graphical tool for accurate measurement of individual surgeon performance metrics, adjusted for both surgeon-specific and patient-specific factors, significantly alters interpretation of performance data.

Design: Retrospective analysis of all total knee replacements (TKR) conducted at the host institution between 1996 and 2009. The database was randomly divided into training and testing datasets. Using multivariate generalised estimating equation (GEEs) regression models, the training dataset enabled generation of patient-risk and surgeon-experience adjustment factors. To simulate prospective monitoring of individual surgeon outcomes, the testing dataset was mapped on control charts. Weighted κ statistics were calculated to measure the agreement between patient risk adjusted and fully-adjusted control charts.

Setting: Tertiary care academic hospital.

Participants: All patients undergoing TKR at the host institution 1996-2009.

Main outcome measure: Operative efficiency.

Results: 5,313 procedures were analysed. Adjusted control charts were generated using a training dataset comprised of 3,756 procedures performed by 13 surgeons. Operative time gradually declined by 121 minutes with 25 years of experience ($P<0.0001$). Charts were tested by monitoring 4 other surgeons, performing an average of 389 procedures each. Adjustment for surgeon experience significantly altered the interpretation of operative efficiency ($\kappa = 0.29$ [95% CI, 0.11-0.47], and enhanced assessment of a surgeon's improvement or diminishment in efficiency over time. Specifically, experience adjustment inverted the interpretation of surgeon efficiency from above average to below average, or from improving to declining performance.

Conclusions: Adjustment for surgeon experience is necessary for accurate interpretation of metrics over the course of a surgeon's career. Patient- and surgeon-adjusted control charts provide an accurate method of monitoring individual operative efficiency.

Main text word count: 2533

Key Words: performance curve, surgical performance improvement, performance monitoring, control chart, total knee replacement, operative efficiency

Article focus

- In the UK, surgeons have recently been required to publish their individual performance data. In turn, monitoring of performance data is known to improve outcomes.
- Adjustment of metrics for patient case-mix significantly alters their interpretation. Surgical procedures have learning curves, owing to the technical nature of the specialty, however the effects of this on outcomes are currently not adjusted for.
- Our aim was to develop an innovative graphical tool for accurate measurement and monitoring of individual surgeon performance metrics, adjusted for both surgeon-specific and patient-specific factors.

Key messages

- Patient- and surgeon- adjusted control charts can be used to accurately assess trends in performance metrics, with the aim of informing professional development.
- Surgeon-specific adjustment is necessary for correct assessment of operative efficiency and performance metrics; failure to do so exposes metrics to statistically significant misinterpretation.

Strengths and limitations of this study

- Use of high-granularity data on over 5000 procedures, across 14 years, performed by 17 surgeons; robust statistical demonstration of the effects of adjusting for surgeon-specific factors.
- Single centre, retrospective, and focused on operative time, which although has clear relevance to operative efficiency and financial costs, is not a clear patient-centred outcome.

INTRODUCTION

Increasing patient demands, costs and emphasis on safety have led to marked interest in performance tracking of individual healthcare providers [1, 2, 3, 4, 5]. While the adoption of techniques to monitor surgeons has lagged behind that of providers in other areas of medicine, we are now witnessing the insinuation of such tools into the surgical sphere. In the United Kingdom, surgeons have recently been required to publish their individual performance data. There has also been a growing interest, amongst both patients and health authorities, to track surgeons' individual performance. Two expressed concerns have been that surgeons will be averse to operating on high-risk patients (those demonstrating unfavorable case-mix), and less likely to take on junior trainees (those with unfavorable experience) during procedures for fear of poor performance results [1, 6, 7, 8]. Without addressing such concerns, it is possible that publication of performance data may result in unwanted changes in practice, and the generation of inaccurate, inequitable data.

Several methods of individual surgeon performance monitoring have been proposed [9, 10, 11], frequently adjusting for patient-specific characteristics, or case-mix, including demographics and medical co-morbidities. The influence of such characteristics upon surgical outcomes has been well-explored and acknowledged [12, 13, 14]. Very recently, the role of surgeon-specific factors such as operative experience and surgical team familiarity, with respect to outcomes, has also been elucidated [15, 16, 17].

The relative importance of adjusting for both patient- and surgeon-specific factors when assessing operative performance has historically been stymied by difficulties in generating research databases of sufficient granularity and robustness to carry out detailed statistical analyses. Such limitations have been ameliorated by the development of large depositories of electronic medical data. Through the parsing of such data, we are now fortunate to have the opportunity to perform inquiries once thought impossible.

While it is important that surgeons are monitored, and clear that measurements result in improvements [3], it is paramount that the 'measuring stick' with which performance is evaluated, is accurate. To address the aforementioned confounders of case-mix and surgeon-experience, we explored the use of patient- and surgeon-adjusted control charts to permit accurate performance tracking of individual surgeons.

Control charts, a tool initially devised in the manufacturing industry, permit iterative improvements in quality by statistical process control [18]. They comprise of mapping a process metric or outcome on a chart with a pre-determined benchmark. Upper and lower limits are placed on the chart, typically at two or three standard deviations from the benchmark value; exceeding these limits indicates an anomalous or unlikely event that is signaled, investigated and when necessary acted upon to improve the charted process. In health care they have demonstrated benefit by permitting identification of causes of variation and enabling safety monitoring [19]. Control charts have been demonstrated to result in improved health outcomes, efficiency and safety [20, 21].

We believe that when appropriately adjusted, control charts may offer similar benefits in the sphere of surgical efficiency and performance. Specifically, our aim was to determine whether such a tool, when adjusted for both surgeon- and patient-specific factors, would significantly alter interpretation of performance data. We present the results of this research endeavor as a proof-of-concept of the potential value of adjusted control charts over more traditional methodologies in performance monitoring.

METHODS

Study design and population

Following institutional review board approval (protocol 2006p000586), we conducted a longitudinal analysis of 5,313 knee replacement procedures performed by 17 surgeons at a single academic tertiary care center (Brigham & Women's Hospital, Boston, MA, USA) between 1996 and 2009. In order to develop the performance chart independently from the data to be monitored, the database was randomly split into training and testing datasets according to surgeon identity [22]. The training dataset included 3,756 procedures performed by 13 surgeons and was used to define the baseline parameters of the performance charts, as well as models for outcomes adjustment. The testing dataset included 1,557 procedures performed by the 4 other surgeons, with the aim of putting the performance charts to the test by checking their application to external data.

Outcome measures and data collection

Data were culled from a combination of electronic medical records, an electronic operative time tracking application, and physician employee databases. Operative time, the primary outcome measurement, was measured in minutes and defined as the time elapsed from skin incision to skin closure. Operative time was used as a proxy for operative efficiency. For each procedure, the length of experience of each participating surgeon was calculated. The operative experience of the attending surgeon was calculated as the difference between the date of the procedure and that of the surgeon's completion of training.

Performance curve modeling and case-mix adjustment

We used the training dataset to determine the adjusted performance curve of surgeons during their career based on a multivariate generalised estimating equation (GEE) regression model, taking into consideration the clustering of patients by surgeon [23]. Operative time was the outcome of interest, while surgeon's experience was the predictor and patient case mix (patient age, sex, smoking status and the presence of comorbidities – type II diabetes mellitus, coronary artery disease and chronic obstructive pulmonary disorder) was considered as a covariate in the final model. Because the operative time curve may not necessarily be a linear function of surgeon experience, a number of possible shapes of performance curves were tested. In order to obtain the best fitting shape, surgeon experience was entered as both a linear term and a quadratic term in the final model [24]. An adjusted performance curve was drawn versus the number of years since surgeon graduation. Also, the reduction in operative time independently associated with the attending surgeon's experience was plotted. Model estimates were obtained using the GENMOD procedure in SAS™ 9.2 (SAS Institute Inc., Cary, NC, USA); all tests were 2-tailed, and *p*-values <0.05 were considered significant.

Design and comparison of patient-risk and fully-adjusted control charts

Surgical outcomes for the testing dataset were further adjusted using model estimates that were previously generated from the training dataset. For each surgical procedure, the expected operative time was computed by controlling for patient case-mix. For every surgeon, adjusted outcomes at a given year of experience were calculated as the ratio between the observed and the expected operative time multiplied by the overall mean operative time. Once adjusted, operative time was then plotted on a Shewhart control chart to simulate prospective outcome monitoring for every individual surgeon over the course of his/her career [25]. Each data point depicted the surgeon operative time per year since graduation. The central line value of the patient-risk adjusted chart was constant and was determined based on the overall mean operative time, while the central line value of the fully-adjusted chart varied and depicted the adjusted performance curve of surgeons as a function of their previous experience that was previously generated from the training dataset. Control and warning limits were set at 3 SD and 2 SD around the central line, respectively, to indicate whether a particular surgeon's performance differed significantly from this goal. The detection of a significantly high or poor performance was defined as a single point outside the control limits, respectively, or two out of three successive points between a warning limit and a control limit on the same side of the central line [26]. Underperforming surgeons were positioned above the upper limits (i.e. longer operative time), while surgeons with unusually good results were below the lower limits (shorter operative time) [27].

The agreement between patient-risk and fully-adjusted charts in detecting indicator variations was measured using the weighted Cohen k statistic [28]. The positions of the data points for surgeon individual performance were compared in terms of 5 ordinal levels based on warning and control limits.

RESULTS

A total of 5,313 total knee replacement procedures performed by 17 surgeons were analysed (Table 1). Median surgical experience was 17 years, ranging from 1 to 35 years in practice since graduation. The mean operative time was 109 minutes. A substantial decline in operative time was observed over the course of surgeon's career, resulting in a concave shaped performance curve ($P<0.0001$, Figure 1; to maintain anonymity, the number of years of experience has been removed from the x-axis of all figures).

Figure 2 demonstrates the patient-risk adjusted operative times of 4 surgeons for TKR with respect to the expected 'benchmark' performance curve over time, with slower operative times being placed above the benchmark, indicating reduced operative efficiency, and faster operative times being placed below the benchmark, indicating improved operative efficiency. Inspection of each surgeon's performance curve revealed that surgeon A displayed better operative efficiency than surgeon B, with lower operative times. Furthermore, surgeon A's operative efficiency, unlike surgeon B's, was better than expected, demonstrating operative times below the benchmark performance curve. Surgeons C and D demonstrated similar operative times. With respect to the benchmark performance curve, however, Surgeon C performed better than expected, given his low experience, displaying superior operative efficiency. Surgeon D, relative to the benchmark performance curve, was found to display worse than expected operative efficiency.

Control charts adjusted only for patient risk depicted Surgeon A as improving. However, after adjusting for surgical experience, Surgeon A's operative efficiency appeared to worsen relative to the population mean (Figure 3). Similarly, Surgeon B was found to be within control limits for most of his operations when considering patient-risk adjustment only; however, surgical experience adjustment showed all but one of his data-points to lie outside of the upper control limit, indicating consistently slower operative efficiency. Experience adjustment transposed Surgeon C's performance curve from variable about the population mean to clearly 'high' operative efficiency, and inverted the interpretation of Surgeon D's operative efficiency from 'high' to 'poor'. The agreement between the patient-risk and fully-adjusted control charts in detecting operative time variations over surgeons career was very low (Table II, $\kappa = 0.29$ [95% CI, 0.11-0.47], indicating the significant effect of adjusting operative time for surgical experience.

Adjustment for surgeon experience, in addition to patient risk, significantly altered the interpretation of operative efficiency, and enhanced the accuracy of assessing a surgeon's improvement or diminishment in efficiency over time.

DISCUSSION

Our study presents a novel methodology for the adjustment and monitoring of surgical metrics, specifically operative efficiency. Surgery is a technical specialty, improving with volume and experience. Consideration of surgeon-specific factors may seem intuitive, but remains poorly investigated. Our findings quantitatively demonstrated that such adjustment significantly altered the interpretation of operative efficiency monitoring – at times resulting in an inversion of perceived trends in efficiency relative to consideration of patient-adjustment alone. Although we investigated operative time, this methodology can be applied to any surgical outcome.

Fully-adjusted control charts were shown to offer an accurate and perceptive means of interpreting trends in a surgeon's efficiency, identifying outlying or anomalous units and providing early warning of divergence from the cohort mean. We believe such factors may prove particularly advantageous in the context of surgeon monitoring and performance tracking.

Limitations

These implications must be considered in the context of this study's limitations. First, this investigation was performed at a single academic medical center, retrospectively, which may limit the representativeness of our sample. It should be noted, however, that the retrospective nature of our investigation removed any Hawthorne bias with regards to performance and therefore may provide a truer depiction of procedural dynamics than could have been ascertained through prospective methods. Second, our focus on operative time did not implicitly incorporate considerations related to patient outcomes. Studies, however, have indicated that faster completion of the TKR procedure is associated with better outcomes [29]. Indeed, in a variety of work both within surgery and outside, time of task completion has been used as a robust indicator of learning and outcome [17, 30, 31, 32, 33, 45]. Further, longer operative times and increased use of the operating theatre, as discussed below, expose patients to greater risks of surgical site infection, whilst also entailing larger financial costs and reduced efficiency, that amidst rising costs in health care, are of clear importance. Third, our investigation utilised years of training as a proxy for surgical experience, rather than number of cases performed. This limitation is a reflection of the fact that a number of surgeons included in the dataset had been in clinical practice prior to the implementation of our surgical tracking application. Years of training, however, has been utilised as an acceptable substitute for surgical experience in prior published studies [17, 18, 19, 39]. Finally, although we adjusted control charts for patient characteristics and surgeon individual experience, there may be other factors, such as non-technical skills and teamwork which need to be accounted for to enable better interpretation of performance.

Policy Implications

Monitoring operative times with the aim of improving operative efficiency has strong financial implications. Theatres, excluding day surgery, have been shown to account for approximately 6% of NHS Trust budgets, equating to running-costs per theatre per year of £1.5 million [34, 35]. In the U.S., the cost of an operating theatre has been estimated at approximately $130 per minute [36]. Strategies that can improve operative efficiency and reduce operative costs are therefore of importance.

The impact of surgeon-specific adjustment itself, also has implications both within and outside the sphere of individual performance-tracking. In the context of training, it is arguably inequitable to compare the performance of less and more experienced trainees; indeed it would be inappropriate to expect a surgeon who has just started their training to perform as well as a surgeon who is about to complete their training, rather, trainees must be compared to cohorts of the same level of experience. Thus, use of surgeon-specific, experience-adjusted charts will permit the performance of young trainees to be accurately and equitably monitored relative to a relevant benchmark, removing bias. This could give rise to appraisals based upon performance rather than career chronology or volume of cases alone, potentially ensuring progression only upon acquisition of sufficient expertise. The tools outlined in this study could furthermore be used to establish minimal competency requirements for operators and permit important contributors to training to be quantitatively identified. In the context of experienced surgeons, fully-adjusted control charts provide a sensitive and timely means of identifying deviations from expected benchmarks, permitting prompt investigation or intervention to improve the respective surgeon's performance. Recent work has also shown that performance may decline as surgeons approach seniority [37]; patient- and surgeon- adjusted control charts have the capacity to identify this deterioration and supplement the implementation of continuing education programs. In the broader context of outcomes research, any studies investigating the impact of an intervention on performance could gain interpretational benefit from surgeon-factor adjustment. Where groups of surgeons or departments are being compared, it is intuitive that adjustment for the respective experiences, or 'surgeon-mix' of these groups are adjusted for, in parallel with patient-mix adjustment, to improve the transparency of results.

We present this research as a proof-of-concept that i) patient- and surgeon-adjusted control charts can be utilised to inform ongoing professional development and feedback for individual surgeons, and ii) surgeon-specific adjustment is necessary for correct assessment of operative efficiency and performance outcomes; failure to do so exposes metrics to statistically significant misinterpretation. We believe this should be considered in developments regarding surgical monitoring, to permit equitable and accurate performance assessment, addressing concerns of patients, surgeons and policy-makers alike.

REFERENCES

1 Godlee F. Measure your team's performance, and publish the results [Editor's Choice]. BMJ 2012;345:e4590.

2 Davis K, Schoen C, Guterman S, et al. Slowing the Growth of U.S. Health Care Expenditures: What Are the Options? New York: The Commonwealth Fund. 2007

3 Commonwealth Fund Commission on a High Performance Health System. Why Not the Best? Results from a National Scorecard on U.S. Health System Performance, 2008. New York: The Commonwealth Fund. 2008.

4 Dimick JB, Weeks WB, Karia R, et al. Who pays for poor surgical quality? Building a business case for quality improvement. J Am Coll Surg. 2006;202:933.

5 Duclos A, Carty MJ. Value of health care delivery. JAMA 2011;306(3):267.

6 Tavare A. Where are we with transparency over performance of doctors and institutions? BMJ. 2012 Jul 3;345:e4464.

7 Ray S, Simpson I. Professional societies can lead the way on transparency but will need support. BMJ. 2012 Jul 31;345:e5075.

8 Hill M. NHS medical director wants surgeon league tables. BBC News [online]. December 5 2012. Available from: http://www.bbc.co.uk/news/health-20584897 [Accessed 4th March 2013].

9 Holzhey DM, Jacobs S, Walther T, Mochalski M, Mohr FW, Falk V. Cumulative sum failure analysis for eight surgeons performing minimally invasive direct coronary artery bypass. J Thorac Cardiovasc Surg. 2007

10 Kusamura S, Baratti D, Deraco M. Multidimensional analysis of the learning curve for cytoreductive surgery and hyperthermic intraperitoneal chemotherapy in peritoneal surface malignancies. Ann Surg. 2012 Feb;255(2):348-56.

11 Tekkis PP, McCulloch P, Steger AC, Benjamin IS, Poloniecki JD. Mortality control charts for comparing performance of surgical units: validation study using hospital mortality data. BMJ. 2003 Apr 12;326(7393):786-8.

12 Duclos A, Voirin N, Touzet S, Soardo P, Schott AM, Colin C, Peix JL, Lifante JC. Crude versus case-mix-adjusted control charts for safety monitoring in thyroid surgery. Qual Saf Health Care. 2010 Dec;19(6):e17.

13 Cook JA, Ramsay CR, Fayers P. Statistical evaluation of learning curve effects in surgical trials. Clinical Trials 2004;1:421-7.

14 High-Volume versus Low-Volume for Esophageal Resections for Cancer: The Essential Role of Case-Mix Adjustments based on Clinical Data. Michael W. Wouters, Bas P. Wijnhoven, Henrieke E. Karim-Kos, Harriet G. Blaauwgeers, Laurents P. Stassen, Willem-Hans Steup, Huug W. Tilanus, Rob A. Tollenaar Ann Surg Oncol. 2008 January; 15(1): 80–87.

15 Carty MJ, Chan R, Huckman R, Snow D, Orgill DP. A detailed analysis of the reduction mammaplasty learning curve: a statistical process model for approaching surgical performance improvement. Plast Reconstr Surg. 2009 Sep;124(3):706-14.

16 Xu R, Carty MJ, Orgill DP, Lipsitz SR, Duclos A. The Teaming Curve: A Longitudinal Study of the Influence of Surgical Team Familiarity on Operative Time. Ann Surg. 2013 Feb 12.

17 Elbardissi AW, Duclos A, Rawn JD, Orgill DP, Carty MJ. Cumulative team experience matters more than individual surgeon experience in cardiac surgery. J Thorac Cardiovasc Surg. 2012 Oct 17.

18 Bewick DM. Controlling variation in health care: a consultation from Walter Shewhart. Med Care. 1991 Dec;29(12):1212-25.

19 Tennant R, Mohammed MA, Coleman JJ, Matin U. Monitoring patients using control charts: a systematic review. Int J Qual Health Care. 2007 Aug;19(4):187-94.

20 Thor J, Lundberg J, Ask J, Olsson J, Carli C, Härenstam KP, Brommels M. Application of statistical process control in healthcare improvement: systematic review. Qual Saf Health Care. 2007 Oct;16(5):387-99.

21 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 Mar;99(3):324-35.

22 Hastie T, Tibshirani R, Friedman J. (2009) The elements of statistical learning. New York, NY: Springer.

23 Liang KY, Zeger SL. Longitudinal Data Analysis Using Generalized Linear Models. Biometrika 1986;73:13-22.

24 Ramsay CR, Grant AM, Wallace SA, Garthwaite PH, Monk AF, et al. (2001) Statistical assessment of the learning curves of health technologies. Health Technol Assess 5(12): 1-79.

25 Montgomery DC. Statistical Quality Control: A Modern Introduction. Hoboken, NJ: Wiley, 2008.

26 Benneyan JC, Lloyd RC, Plsek PE. Statistical process control as a tool for research and healthcare improvement. Qual Saf Health Care 2003;12(6): 458-64.

27 Duclos A, Voirin N. The p-control chart: a tool for care improvement. Int J Qual Health Care. 2010;22(5):402-407.

28 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968 Oct;70(4):213-20.

29 Peersman G, Laskin R, Davis J, Peterson MG, Richart T. Prolonged operative time correlates with increased infection rate after total knee arthroplasty. HSS J. 2006 Feb;2(1):70-2.

30 Pisano GP, Bohmer MJ, Edmondson AC. 2001. Organizational differences in rates of learning: Evidence from the adoption of minimally invasive cardiac surgery. Management Science, 47, 752-768.

31 Edmondson A, Winslow A, Bohmer R, Pisano G. 2003. Learning how and learning what: Effects of tacit and codified knowledge on performance improvement following technology adoption. Decision Sciences, 34, 197-223.

32 Epple D, Argote L, Devadas R. 1991. Organizational learning curves: A method for investigating intra-plant transfer of knowledge acquired through learning by doing. Organization Science, 2, 58-70.

33 Argote L, Epple D. Learning curves in manufacturing. Science 1990; 247, 920-924.

45. Peersman G, Laskin R, Davis J, Peterson MG, Richart T. Prolonged operative time correlates with increased infection rate after total knee arthroplasty. HSS J. 2006 Feb;2(1):70-2.

34 http://www.isdscotland.org/Health-Topics/Finance/Costs/Detailed-Tables/Theatres.asp [Accessed August 30, 2013]

35 West Hertfordshire Hiospitals NHS Trust. Accessed from: http://www.westhertshospitals.nhs.uk/foi_publication_scheme/disclosure_log/2010/december/documents/170%20-%20140111.pdf [Accessed August 30, 2013]

36 Shippert RD. A study of time-dependent operating room fees and how to save $100,000 by using time-saving products. *American Journal of Cosmetic Surgery*. 2005;22(1):25–34.

37 Duclos A, Peix JL, Colin C, Kraimps JL, Menegaux F, Pattou F, Sebag F, Touzet S, Bourdy S, Voirin N, Lifante JC; CATHY Study Group. Influence of experience on performance of individual surgeons in thyroid surgery: prospective cross sectional multicentre study. BMJ. 2012 Jan 10;344:d8041.

FIGURE LEGENDS

Figure 1. Performance curve for individual surgeon TKR operative efficiency. The graph illustrates how operative time within the cohort changed with surgeon experience.

Figure 2. Individual performance curves for surgeons A-D**.** The graph illustrates the patient-risk adjusted operative times of the 4 surgeons selected to test the control charts, with respect to the expected 'benchmark' performance curve.

Figure 3. Patient-risk vs. fully- adjusted control charts for individual surgeons. For each surgeon a patient-risk adjusted chart, and fully-adjusted (patient-risk and surgeon-experience adjusted) chart is displayed. The horizontal axes indicate the experience of the surgeon in years and the blue curve his/her adjusted performance over time. The central black dotted line represents the expected operative time over the course of surgeon's career. The upper red and lower green lines illustrate poor and high performance limits, set at two standard deviations (dotted warning limits) and three standard deviations (continuous control limits) around the central line. Poor and high performers are defined as those breaching the upper and lower limits, respectively. Average performers are those with operative time around the central line, without crossing the limits.

Figure 1. Improvement curve for individual surgeons

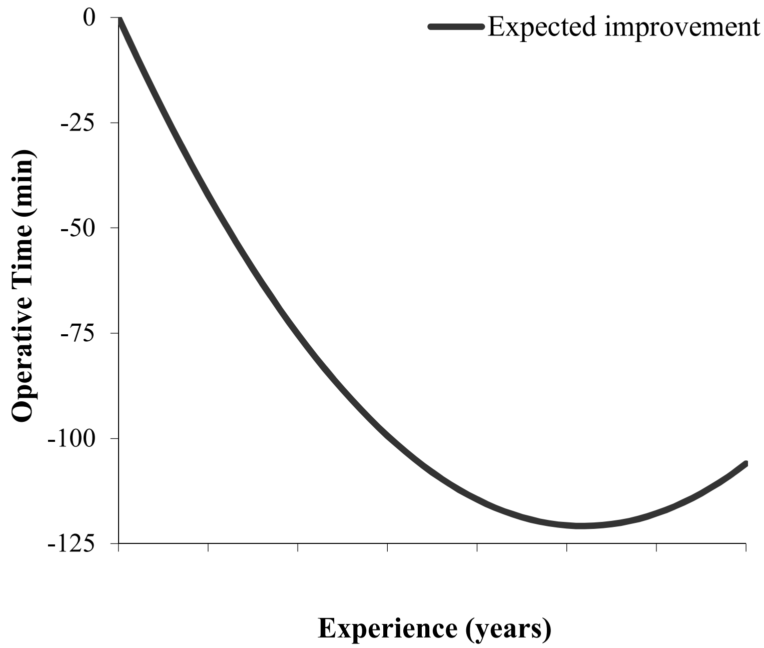


Figure 2. Individual performance for surgeons A-D

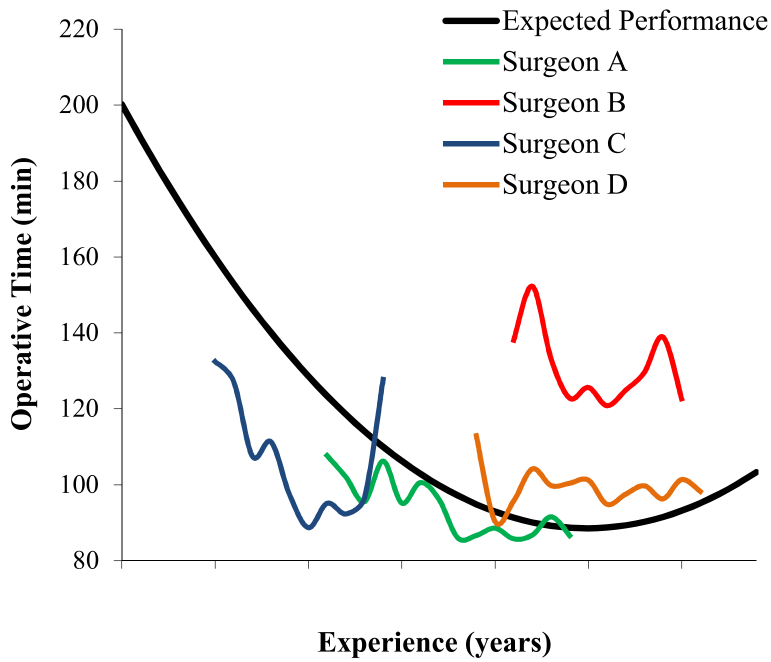


Figure 3: Patient-risk vs. fully- adjusted control charts for surgeons A-D

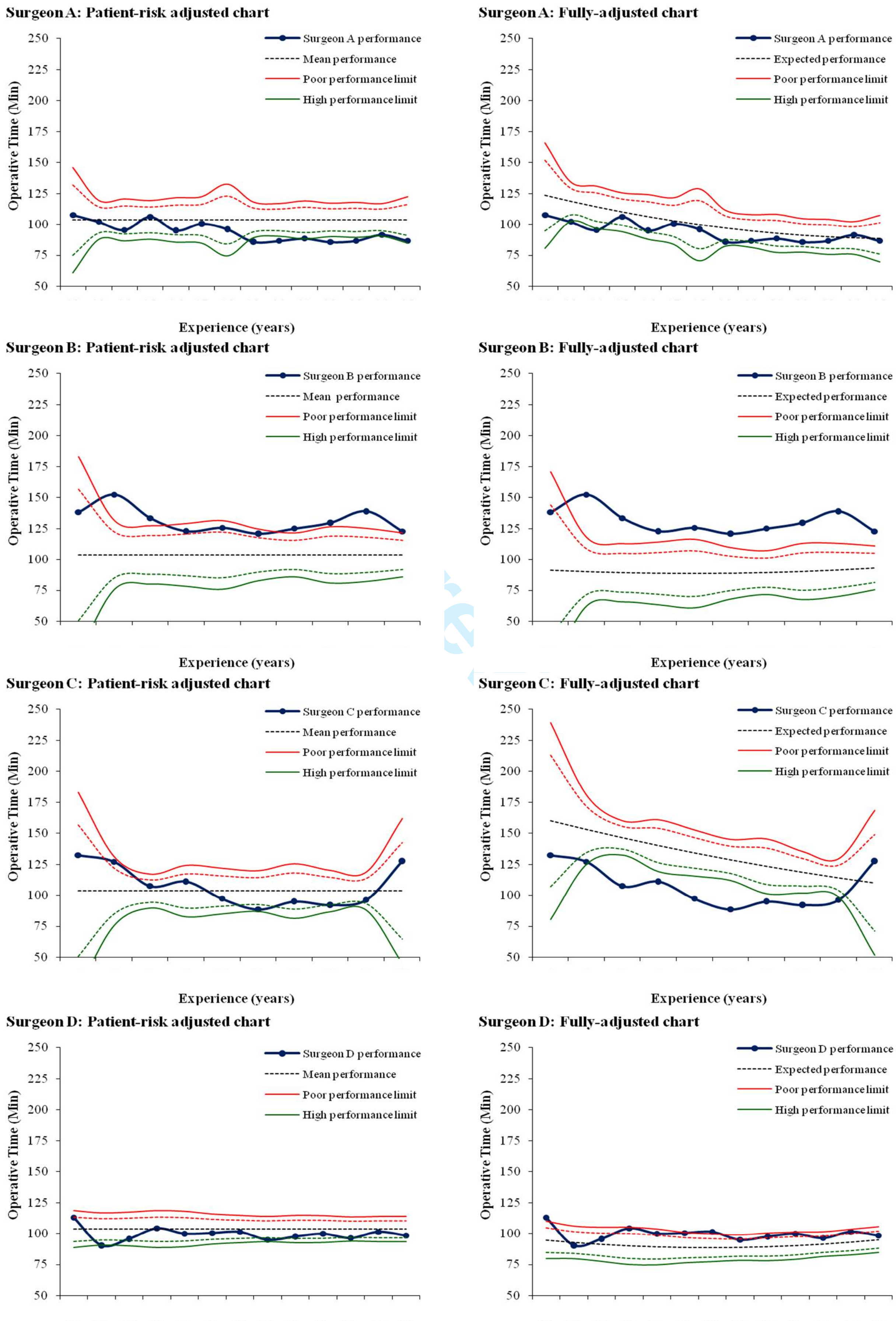


Table 1. Overview of study participants

Attending surgeon (N=17)	
Surgeon experience, years, Median *(Min-Max)*	17 *(1-35)*
Surgeon volume of cases, Median *(Min-Max)*	176 *(10-1,871)*
Surgical cases (N=5,313)	
Patient female gender, No. *(%)*	3,558 *(67.0)*
Patient age, years, Mean *(SD)*	66.2 *(11.3)*
Patient with comorbidity, No. *(%)*	3,388 *(63.8)*
No. of comorbidities, Median *(Min-Max)*	1 *(0-6)*
Coronary artery disease, No. *(%)*	1,074 *(20.2)*
Chronic obstructive pulmonary disease, No. *(%)*	320 *(6.0)*
Diabetes mellitus, No. *(%)*	858 *(16.1)*
Hypertension, No. *(%)*	2,196 *(41.3)*
Obesity, No. *(%)*	1,242 *(23.4)*
Tobacco, No. *(%)*	814 *(15.3)*
Operative time, minutes, Mean *(SD)*	109.2 *(30.3)*

Table 2. Agreement between patient-risk adjusted and fully-adjusted charts in detecting indicator variations.

		Fully-adjusted chart*					
		<LCL	LCL-LWL	LWL-UWL	UWL-UCL	>UCL	Total
Patient-risk adjusted chart*	<LCL	0	1	4	0	0	5
	LCL-LWL	2	0	5	0	0	7
	LWL-UWL	7	0	10	6	2	25
	UWL-UCL	0	1	0	0	3	4
	>UCL	0	0	0	0	6	6
	Total	9	2	19	6	11	47

* Each unit in the table represents the position of a data point on a control chart, according to 5 ordinal levels based on Warning Limits (2SD) and Control Limits (3SD), as follows: <LCL (below the lower control limit), LCL-LWL (between the lower control and warning limits), LWL-UWL (between the lower and upper warning limits), UWL-UCL (between the upper warning and control limits), >UCL (above the upper control limit).

PATIENT- AND SURGEON-ADJUSTED CONTROL CHARTS FOR MONITORING INDIVIDUAL PERFORMANCE

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-004046.R1
Article Type:	Research
Date Submitted by the Author:	07-Dec-2013

Complete List of Authors:	Maruthappu, Mahiben; University of Oxford, Carty, Matthew; Harvard Medical School, Lipsitz, Stuart; Harvard Medical School, Wright, John; Harvard Medical School, Orgill, Dennis; Harvard Medical School, ; Brigham and Women's Hospital, Division of Plastic Surgery Duclos, Antoine; Université de Lyon,

Primary Subject Heading:	Surgery
Secondary Subject Heading:	Public health
Keywords:	Orthopaedic & trauma surgery < SURGERY, SURGERY, PUBLIC HEALTH

PATIENT- AND SURGEON-ADJUSTED CONTROL CHARTS FOR MONITORING PERFORMANCE

Mahiben Maruthappu, M.A., B.M. B.Ch.[1*]

Matthew J. Carty, M.D.[2*]

Stuart R. Lipsitz, Sc.D., Ph.D.[3]

John Wright, M.D.[4]

Dennis Orgill, M.D., Ph.D.[5]

Antoine Duclos, M.D., Ph.D.[6]

[1]Joan & Richard Doll Scholar, Green Templeton College, University of Oxford, UK

[2]Assistant Professor of Surgery, Harvard Medical School, USA

[3]Professor of Medicine, Harvard Medical School, USA

[4]Assistant Professor of Surgery, Harvard Medical School, USA

[5]Professor of Surgery, Harvard Medical School, USA; Vice-Chairman of Surgery, Brigham and Women's Hospital, USA

[6]Associate Professor of Public Health, Université de Lyon, France

* Both authors equally contributed to the development of this manuscript

Please address correspondence to:

Mahiben Maruthappu

Green Templeton College

Oxfordshire

OX2 6HG

maruthappu@post.harvard.edu

ABSTRACT

Objectives: To determine whether an innovative graphical tool for accurate measurement of individual surgeon performance metrics, adjusted for both surgeon-specific and patient-specific factors, significantly alters interpretation of performance data.

Design: Retrospective analysis of all total knee replacements (TKR) conducted at the host institution between 1996 and 2009. The database was randomly divided into training and testing datasets. Using multivariate generalised estimating equation (GEEs) regression models, the training dataset enabled generation of patient-risk and surgeon-experience adjustment factors. To simulate prospective monitoring of individual surgeon outcomes, the testing dataset was mapped on control charts. Weighted κ statistics were calculated to measure the agreement between patient risk adjusted and fully-adjusted control charts.

Setting: Tertiary care academic hospital.

Participants: All patients undergoing TKR at the host institution 1996-2009.

Main outcome measure: Operative efficiency.

Results: 5,313 procedures were analysed. Adjusted control charts were generated using a training dataset comprised of 3,756 procedures performed by 13 surgeons. Operative time gradually declined by 121 minutes with 25 years of experience ($P<0.0001$). Charts were tested by monitoring 4 other surgeons, performing an average of 389 procedures each. Adjustment for surgeon experience significantly altered the interpretation of operative efficiency ($\kappa = 0.29$ [95% CI, 0.11-0.47], and enhanced assessment of a surgeon's improvement or diminishment in efficiency over time. Specifically, experience adjustment inverted the interpretation of surgeon efficiency from above average to below average, or from improving to declining performance.

Conclusions: Adjustment for surgeon experience is necessary for accurate interpretation of metrics over the course of a surgeon's career. Patient- and surgeon-adjusted control charts provide an accurate method of monitoring individual operative efficiency.

Main text word count: 2533

Key Words: performance curve, surgical performance improvement, performance monitoring, control chart, total knee replacement, operative efficiency

Article focus

- In the UK, surgeons have recently been required to publish their individual performance data. In turn, monitoring of performance data is known to improve outcomes.
- Adjustment of metrics for patient case-mix significantly alters their interpretation. Surgical procedures have learning curves, owing to the technical nature of the specialty, however the effects of this on outcomes are currently not adjusted for.
- Our aim was to develop an innovative graphical tool for accurate measurement and monitoring of individual surgeon performance metrics, adjusted for both surgeon-specific and patient-specific factors.

Key messages

- Patient- and surgeon- adjusted control charts can be used to accurately assess trends in performance metrics, with the aim of informing professional development.
- Surgeon-specific adjustment is necessary for correct assessment of operative efficiency and performance metrics; failure to do so exposes metrics to statistically significant misinterpretation.

Strengths and limitations of this study

- Use of high-granularity data on over 5000 procedures, across 14 years, performed by 17 surgeons; robust statistical demonstration of the effects of adjusting for surgeon-specific factors.
- Single centre, retrospective, and focused on operative time, which although has clear relevance to operative efficiency and financial costs, is not a clear patient-centred outcome.

INTRODUCTION

Increasing patient demands, costs and emphasis on safety have led to marked interest in performance tracking of individual healthcare providers [1, 2, 3, 4, 5]. While the adoption of techniques to monitor surgeons has lagged behind that of providers in other areas of medicine, we are now witnessing the insinuation of such tools into the surgical sphere. In the United Kingdom, surgeons have recently been required to publish their individual performance data. There has also been a growing interest, amongst both patients and health authorities, to track surgeons' individual performance. Two expressed concerns have been that surgeons will be averse to operating on high-risk patients (those demonstrating unfavorable case-mix), and less likely to take on junior trainees (those with unfavorable experience) during procedures for fear of poor performance results [1, 6, 7, 8]. Without addressing such concerns, it is possible that publication of performance data may result in unwanted changes in practice, and the generation of inaccurate, inequitable data.

Several methods of individual surgeon performance monitoring have been proposed [9, 10, 11], frequently adjusting for patient-specific characteristics, or case-mix, including demographics and medical co-morbidities. The influence of such characteristics upon surgical outcomes has been well-explored and acknowledged [12, 13, 14]. Very recently, the role of surgeon-specific factors such as operative experience and surgical team familiarity, with respect to outcomes, has also been elucidated [15, 16, 17].

The relative importance of adjusting for both patient- and surgeon-specific factors when assessing operative performance has historically been stymied by difficulties in generating research databases of sufficient granularity and robustness to carry out detailed statistical analyses. Such limitations have been ameliorated by the development of large depositories of electronic medical data. Through the parsing of such data, we are now fortunate to have the opportunity to perform inquiries once thought impossible.

While it is important that surgeons are monitored, and clear that measurements result in improvements [1, 6], it is paramount that the 'measuring stick' with which performance is evaluated, is accurate. To address the aforementioned confounders of case-mix and surgeon-experience, we explored the use of patient- and surgeon-adjusted control charts to permit accurate performance tracking of individual surgeons.

Control charts, a tool initially devised in the manufacturing industry, permit iterative improvements in quality by statistical process control [18]. They comprise of mapping a process metric or outcome on a chart with a pre-determined benchmark. Upper and lower limits are placed on the chart, typically at two or three standard deviations from the benchmark value; exceeding these limits indicates an anomalous or unlikely event that is signaled, investigated and when necessary acted upon to improve the charted process. In health care they have demonstrated benefit by permitting identification of causes of variation and enabling safety monitoring [19]. Control charts have been demonstrated to result in improved health outcomes, efficiency and safety [20, 21].

We believe that when appropriately adjusted, control charts may offer similar benefits in the sphere of surgical efficiency and performance. Specifically, our aim was to determine whether such a tool, when adjusted for both surgeon- and patient-specific factors, would significantly alter interpretation of performance data. We present the results of this research endeavor as a proof-of-concept of the potential value of adjusted control charts over more traditional methodologies in performance monitoring.

METHODS

Study design and population

Following institutional review board approval (protocol 2006p000586), we conducted a longitudinal analysis of 5,313 knee replacement procedures performed by 17 surgeons at a single academic tertiary care center (Brigham & Women's Hospital, Boston, MA, USA) between 1996 and 2009. In order to develop the performance chart independently from the data to be monitored, the database was randomly split into training and testing datasets according to surgeon identity [22]. The training dataset included 3,756 procedures performed by 13 surgeons and was used to define the baseline parameters of the control charts, including upper and lower limits for the charts, as well as models for case-mix adjustment, and experience-adjustment. The testing dataset included 1,557 procedures performed by the 4 other surgeons; operative times of procedures by each of these surgeons was mapped on control charts developed using the training dataset, to simulate performance monitoring.

Outcome measures and data collection

Data were culled from a combination of electronic medical records, an electronic operative time tracking application, and physician employee databases. Operative time, the primary outcome measurement, was measured in minutes and defined as the time elapsed from skin incision to skin closure. Operative time was used as a proxy for operative efficiency. For each procedure, the length of experience of each participating surgeon was calculated. The operative experience of the attending surgeon was calculated as the difference between the date of the procedure and that of the surgeon's completion of training.

Performance curve modeling and case-mix adjustment

We used the training dataset to determine the adjusted performance curve of surgeons during their career based on a multivariate generalised estimating equation (GEE) regression model, also taking into consideration the clustering of patients by surgeon [23]. Operative time was the outcome of interest, while surgeon experience was the predictor and patient case mix (patient age, sex, smoking status and the presence of comorbidities – type II diabetes mellitus, coronary artery disease and chronic obstructive pulmonary disorder) was considered as a covariate in the final model. We included all available co-morbidities from our dataset in the case-mix adjustment model. Because the operative time curve may not necessarily be a linear function of surgeon experience, a number of possible shapes of performance curves were tested. In order to obtain the best fitting shape, surgeon experience was entered as both a linear term and a quadratic term in the final model [24]. An adjusted performance curve was drawn versus the number of years since surgeon graduation. Also, the reduction in operative time independently associated with the attending surgeon's experience was plotted. Model estimates were obtained using the GENMOD procedure in SAS™ 9.2 (SAS Institute Inc., Cary, NC, USA); all tests were 2-tailed, and *p*-values <0.05 were considered significant.

Design and comparison of patient-risk and fully-adjusted control charts

Operative times from procedures in the testing dataset were adjusted using models derived from the training dataset. Operative times were adjusted for patient-risk alone, or for patient-risk and surgeon experience together (fully-adjusted). For every surgeon, adjusted outcomes at a given year of experience were calculated as the ratio between the observed and the expected operative time multiplied by the overall mean operative time. Once adjusted, operative time was then plotted on a Shewhart control chart to simulate prospective outcome monitoring for every individual surgeon over the course of his/her career [25]. Each data point depicted the surgeon operative time per year since graduation. The central line of the patient-risk adjusted chart was constant and was determined based on the overall mean operative time, while the central line value of the fully-adjusted chart varied and depicted the adjusted performance curve of surgeons as a function of their previous experience that was previously generated from the training dataset. Control and warning limits were set at 3 SD and 2 SD around the central line, respectively, to indicate whether a particular surgeon's performance differed significantly from this goal. The detection of a significantly high or poor performance was defined as a single point outside the control limits, respectively, or two out of three successive points between a warning limit and a control limit on the same side of the central line [26]. Underperforming surgeons were positioned above the upper limits (i.e. longer operative time), while surgeons with unusually good results were below the lower limits (shorter operative time) [27].

The agreement between patient-risk and fully-adjusted charts in detecting indicator variations was measured using the weighted Cohen k statistic [28]. The positions of the data points for surgeon individual performance were compared in terms of 5 ordinal levels based on warning and control limits.

RESULTS

A total of 5,313 total knee replacement procedures performed by 17 surgeons were analysed. Median surgical experience was 17 years, ranging from 1 to 35 years in practice since graduation. The mean operative time was 109 minutes (standard deviation 30.3 minutes). A substantial decline in risk-adjusted operative time was observed over the course of surgeon's career, resulting in a concave shaped performance curve ($P<0.0001$, Figure 1; to maintain anonymity, the number of years of experience has been removed from the x-axis of all figures). Table 1 summarises the surgeon and patient characteristics of the training and testing datasets; Table 2 displays the number of procedures performed by each surgeon.

Figure 2 demonstrates the patient-risk adjusted operative times of 4 surgeons for TKR with respect to the expected 'benchmark' performance curve over time, with slower operative times being placed above the benchmark, indicating reduced operative efficiency, and faster operative times being placed below the benchmark, indicating improved operative efficiency. Inspection of each surgeon's performance curve revealed that surgeon A displayed better operative efficiency than surgeon B, with lower operative times. Furthermore, surgeon A's operative efficiency, unlike surgeon B's, was better than expected, demonstrating operative times below the benchmark performance curve. Surgeons C and D demonstrated similar operative times. With respect to the benchmark performance curve, however, Surgeon C performed better than expected, given his low experience, displaying superior operative efficiency. Surgeon D, relative to the benchmark performance curve, was found to display worse than expected operative efficiency.

Control charts adjusted only for patient risk depicted Surgeon A as improving. However, after adjusting for surgical experience, Surgeon A's operative efficiency appeared to worsen relative to the population mean (Figure 3). Similarly, Surgeon B was found to be within control limits for most of his operations when considering patient-risk adjustment only; however, surgical experience adjustment showed all but one of his data-points to lie outside of the upper control limit, indicating consistently slower operative efficiency. Experience adjustment transposed Surgeon C's performance curve from variable about the population mean to clearly 'high' operative efficiency, and inverted the interpretation of Surgeon D's operative efficiency from 'high' to 'poor'. The agreement between the patient-risk and fully-adjusted control charts in detecting operative time variations over surgeons career was very low (Table 3, $\kappa = 0.29$ [95% CI, 0.11-0.47], indicating the significant effect of adjusting operative time for surgical experience.

Adjustment for surgeon experience, in addition to patient risk, significantly altered the interpretation of operative efficiency, and enhanced the accuracy of assessing a surgeon's improvement or diminishment in efficiency over time.

DISCUSSION

Our study presents a novel methodology for the adjustment and monitoring of surgical metrics, specifically operative efficiency. Surgery is a technical specialty, improving with volume and experience. Consideration of surgeon-specific factors may seem intuitive, but remains poorly investigated. Our findings quantitatively demonstrated that such adjustment significantly altered the interpretation of operative efficiency monitoring – at times resulting in an inversion of perceived trends in efficiency relative to consideration of patient-adjustment alone. Although we investigated operative time, this methodology can be applied to any surgical outcome.

Fully-adjusted control charts were shown to offer an accurate and perceptive means of interpreting trends in a surgeon's efficiency, identifying outlying or anomalous units and providing early warning of divergence from the cohort mean. We believe such factors may prove particularly advantageous in the context of surgeon monitoring and performance tracking.

Limitations

These implications must be considered in the context of this study's limitations. First, this investigation was performed at a single academic medical center, retrospectively, which may limit the representativeness of our sample. It should be noted, however, that the retrospective nature of our investigation removed any Hawthorne bias with regards to performance and therefore may provide a truer depiction of procedural dynamics than could have been ascertained through prospective methods. Importantly, our database covered a substantial time period and it is possible that operative technique and technology may have changed during this. Second, our focus on operative time did not implicitly incorporate considerations related to patient outcomes. Studies, however, have indicated that faster completion of the TKR procedure is associated with better outcomes [29]. Indeed, in a variety of work both within surgery and outside, time of task completion has been used as a robust indicator of learning and outcome [17, 30, 31, 32, 33, 34]. Further, longer operative times and increased use of the operating theatre, as discussed below, expose patients to greater risks of surgical site infection, whilst also entailing larger financial costs and reduced efficiency, that amidst rising costs in health care, are of clear importance. Third, our investigation utilised years of training as a proxy for surgical experience, rather than number of cases performed. This limitation is a reflection of the fact that a number of surgeons included in the dataset had been in clinical practice prior to the implementation of our surgical tracking application. Years of training, however, has been utilised as an acceptable substitute for surgical experience in prior published studies [15, 16, 17]. Finally, although we adjusted control charts for patient characteristics and surgeon individual experience, there may be other factors, such as non-technical skills, teamwork and resident involvement which need to be accounted for to enable better interpretation of performance.

Policy Implications

Monitoring operative times with the aim of improving operative efficiency has strong financial implications. Theatres, excluding day surgery, have been shown to account for approximately 6% of NHS Trust budgets, equating to running-costs per theatre per year of £1.5 million [35, 36]. In the U.S., the cost of an operating theatre has been estimated at approximately $130 per minute [37]. Strategies that can improve operative efficiency and reduce operative costs are therefore of importance.

The impact of surgeon-specific adjustment itself, also has implications both within and outside the sphere of individual performance-tracking. In the context of training, it is arguably inequitable to compare the performance of less and more experienced trainees; indeed it would be inappropriate to expect a surgeon who has just started their training to perform as well as a surgeon who is about to complete their training, rather, trainees must be compared to cohorts of the same level of experience. Thus, use of surgeon-specific, experience-adjusted charts will permit the performance of young trainees to be accurately and equitably monitored relative to a relevant benchmark, removing bias. This could give rise to appraisals based upon performance rather than career chronology or volume of cases alone, potentially ensuring progression only upon acquisition of sufficient expertise. The tools outlined in this study could furthermore be used to establish minimal competency requirements for operators and permit important contributors to training to be quantitatively identified. In the context of experienced surgeons, fully-adjusted control charts provide a sensitive and timely means of identifying deviations from expected benchmarks, permitting prompt investigation or intervention to improve the respective surgeon's performance. Recent work has also shown that performance may decline as surgeons approach seniority [38]; patient- and surgeon- adjusted control charts have the capacity to identify this deterioration and supplement the implementation of continuing education programs. In the broader context of outcomes research, any studies investigating the impact of an intervention on performance could gain interpretational benefit from surgeon-factor adjustment. Where groups of surgeons or departments are being compared, it is intuitive that adjustment for the respective experiences, or 'surgeon-mix' of these groups are adjusted for, in parallel with patient-mix adjustment, to improve the transparency of results.

We present this research as a proof-of-concept that i) patient- and surgeon-adjusted control charts can be utilised to inform ongoing professional development and feedback for individual surgeons, and ii) surgeon-specific adjustment is necessary for correct assessment of operative efficiency and performance outcomes; failure to do so exposes metrics to statistically significant misinterpretation. We believe this should be considered in developments regarding surgical monitoring, to permit equitable and accurate performance assessment, addressing concerns of patients, surgeons and policy-makers alike.

Contributorship Statement All authors fulfill the ICMJE criteria for authorship.

Competing interests None

Data Sharing Statement Further information on data used in the study is available on request.

REFERENCES

1 Godlee F. Measure your team's performance, and publish the results [Editor's Choice]. BMJ 2012;345:e4590.

2 Davis K, Schoen C, Guterman S, et al. Slowing the Growth of U.S. Health Care Expenditures: What Are the Options? New York: The Commonwealth Fund. 2007

3 Commonwealth Fund Commission on a High Performance Health System. Why Not the Best? Results from a National Scorecard on U.S. Health System Performance, 2008. New York: The Commonwealth Fund. 2008.

4 Dimick JB, Weeks WB, Karia R, et al. Who pays for poor surgical quality? Building a business case for quality improvement. J Am Coll Surg. 2006;202:933.

5 Duclos A, Carty MJ. Value of health care delivery. JAMA 2011;306(3):267.

6 Tavare A. Where are we with transparency over performance of doctors and institutions? BMJ. 2012 Jul 3;345:e4464.

7 Ray S, Simpson I. Professional societies can lead the way on transparency but will need support. BMJ. 2012 Jul 31;345:e5075.

8 Hill M. NHS medical director wants surgeon league tables. BBC News [online]. December 5 2012. Available from: http://www.bbc.co.uk/news/health-20584897 [Accessed 4th March 2013].

9 Holzhey DM, Jacobs S, Walther T, Mochalski M, Mohr FW, Falk V. Cumulative sum failure analysis for eight surgeons performing minimally invasive direct coronary artery bypass. J Thorac Cardiovasc Surg. 2007

10 Kusamura S, Baratti D, Deraco M. Multidimensional analysis of the learning curve for cytoreductive surgery and hyperthermic intraperitoneal chemotherapy in peritoneal surface malignancies. Ann Surg. 2012 Feb;255(2):348-56.

11 Tekkis PP, McCulloch P, Steger AC, Benjamin IS, Poloniecki JD. Mortality control charts for comparing performance of surgical units: validation study using hospital mortality data. BMJ. 2003 Apr 12;326(7393):786-8.

12 Duclos A, Voirin N, Touzet S, et al. Crude versus case-mix-adjusted control charts for safety monitoring in thyroid surgery. Qual Saf Health Care. 2010 Dec;19(6):e17.

13 Cook JA, Ramsay CR, Fayers P. Statistical evaluation of learning curve effects in surgical trials. Clinical Trials 2004;1:421-7.

14 High-Volume versus Low-Volume for Esophageal Resections for Cancer: The Essential Role of Case-Mix Adjustments based on Clinical Data. Michael W. Wouters, Bas P. Wijnhoven, Henrieke E. Karim-Kos, Harriet G. Blaauwgeers, Laurents P. Stassen, Willem-Hans Steup, Huug W. Tilanus, Rob A. Tollenaar Ann Surg Oncol. 2008 January; 15(1): 80–87.

15 Carty MJ, Chan R, Huckman R, et al.. A detailed analysis of the reduction mammaplasty learning curve: a statistical process model for approaching surgical performance improvement. Plast Reconstr Surg. 2009 Sep;124(3):706-14.

16 Xu R, Carty MJ, Orgill DP, et al. The Teaming Curve: A Longitudinal Study of the Influence of Surgical Team Familiarity on Operative Time. Ann Surg. 2013 Feb 12.

17 Elbardissi AW, Duclos A, Rawn JD, et al. Cumulative team experience matters more than individual surgeon experience in cardiac surgery. J Thorac Cardiovasc Surg. 2012 Oct 17.

18 Bewick DM. Controlling variation in health care: a consultation from Walter Shewhart. Med Care. 1991 Dec;29(12):1212-25.

19 Tennant R, Mohammed MA, Coleman JJ, et al. Monitoring patients using control charts: a systematic review. Int J Qual Health Care. 2007 Aug;19(4):187-94.

20 Thor J, Lundberg J, Ask J, et al. Application of statistical process control in healthcare improvement: systematic review. Qual Saf Health Care. 2007 Oct;16(5):387-99.

21 Nicolay CR, Purkayastha S, Greenhalgh Aet al. Systematic review of the application of quality improvement methodologies from the manufacturing industry to surgical healthcare. Br J Surg. 2012 Mar;99(3):324-35.

22 Hastie T, Tibshirani R, Friedman J. (2009) The elements of statistical learning. New York, NY: Springer.

23 Liang KY, Zeger SL. Longitudinal Data Analysis Using Generalized Linear Models. Biometrika 1986;73:13-22.

24 Ramsay CR, Grant AM, Wallace SA, , et al. (2001) Statistical assessment of the learning curves of health technologies. Health Technol Assess 5(12): 1-79.

25 Montgomery DC. Statistical Quality Control: A Modern Introduction. Hoboken, NJ: Wiley, 2008.

26 Benneyan JC, Lloyd RC, Plsek PE. Statistical process control as a tool for research and healthcare improvement. Qual Saf Health Care 2003;12(6): 458-64.

27 Duclos A, Voirin N. The p-control chart: a tool for care improvement. Int J Qual Health Care. 2010;22(5):402-407.

28 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968 Oct;70(4):213-20.

29 Peersman G, Laskin R, Davis J, et al. Prolonged operative time correlates with increased infection rate after total knee arthroplasty. HSS J. 2006 Feb;2(1):70-2.

30 Pisano GP, Bohmer MJ, Edmondson AC. 2001. Organizational differences in rates of learning: Evidence from the adoption of minimally invasive cardiac surgery. Management Science, 47, 752-768.

31 Edmondson A, Winslow A, Bohmer R, et al. 2003. Learning how and learning what: Effects of tacit and codified knowledge on performance improvement following technology adoption. Decision Sciences, 34, 197-223.

32 Epple D, Argote L, Devadas R. 1991. Organizational learning curves: A method for investigating intra-plant transfer of knowledge acquired through learning by doing. Organization Science, 2, 58-70.

33 Argote L, Epple D. Learning curves in manufacturing. Science 1990; 247, 920-924.

34 Peersman G, Laskin R, Davis J, et al. Prolonged operative time correlates with increased infection rate after total knee arthroplasty. HSS J. 2006 Feb;2(1):70-2.

35 http://www.isdscotland.org/Health-Topics/Finance/Costs/Detailed-Tables/Theatres.asp [Accessed August 30, 2013]

36 West Hertfordshire Hiospitals NHS Trust. Accessed from: http://www.westhertshospitals.nhs.uk/foi_publication_scheme/disclosure_log/2010/december/documents/170%20-%20140111.pdf [Accessed August 30, 2013]

37 Shippert RD. A study of time-dependent operating room fees and how to save $100,000 by using time-saving products. *American Journal of Cosmetic Surgery*. 2005;22(1):25–34.

38 Duclos A, Peix JL, Colin C, et al; CATHY Study Group. Influence of experience on performance of individual surgeons in thyroid surgery: prospective cross sectional multicentre study. BMJ. 2012 Jan 10;344:d8041.

PATIENT- AND SURGEON-ADJUSTED CONTROL CHARTS FOR MONITORING PERFORMANCE

Mahiben Maruthappu, M.A., B.M. B.Ch.[1*]

Matthew J. Carty, M.D.[2*]

Stuart R. Lipsitz, Sc.D., Ph.D.[3]

John Wright, M.D.[4]

Dennis Orgill, M.D., Ph.D.[5]

Antoine Duclos, M.D., Ph.D.[6]

[1]Joan & Richard Doll Scholar, Green Templeton College, University of Oxford, UK

[2]Assistant Professor of Surgery, Harvard Medical School, USA

[3]Professor of Medicine, Harvard Medical School, USA

[4]Assistant Professor of Surgery, Harvard Medical School, USA

[5]Professor of Surgery, Harvard Medical School, USA; Vice-Chairman of Surgery, Brigham and Women's Hospital, USA

[6]Associate Professor of Public Health, Université de Lyon, France

* Both authors equally contributed to the development of this manuscript

Please address correspondence to:

Mahiben Maruthappu

Green Templeton College

Oxfordshire

OX2 6HG

maruthappu@post.harvard.edu

Financial Disclosure and Products Statement

None of the participating authors has a conflicting financial interest related to the work detailed in this manuscript, nor do any of the authors maintain a financial stake in any product, device or drug cited in this report. Ethical and institutional board approval was granted prior to conducting this study. Information on data used in this report is available upon request.

ABSTRACT

Objectives: To determine whether an innovative graphical tool for accurate measurement of individual surgeon performance metrics, adjusted for both surgeon-specific and patient-specific factors, significantly alters interpretation of performance data.

Design: Retrospective analysis of all total knee replacements (TKR) conducted at the host institution between 1996 and 2009. The database was randomly divided into training and testing datasets. Using multivariate generalised estimating equation (GEEs) regression models, the training dataset enabled generation of patient-risk and surgeon-experience adjustment factors. To simulate prospective monitoring of individual surgeon outcomes, the testing dataset was mapped on control charts. Weighted κ statistics were calculated to measure the agreement between patient risk adjusted and fully-adjusted control charts.

Setting: Tertiary care academic hospital.

Participants: All patients undergoing TKR at the host institution 1996-2009.

Main outcome measure: Operative efficiency.

Results: 5,313 procedures were analysed. Adjusted control charts were generated using a training dataset comprised of 3,756 procedures performed by 13 surgeons. Operative time gradually declined by 121 minutes with 25 years of experience ($P<0.0001$). Charts were tested by monitoring 4 other surgeons, performing an average of 389 procedures each. Adjustment for surgeon experience significantly altered the interpretation of operative efficiency ($\kappa = 0.29$ [95% CI, 0.11-0.47], and enhanced assessment of a surgeon's improvement or diminishment in efficiency over time. Specifically, experience adjustment inverted the interpretation of surgeon efficiency from above average to below average, or from improving to declining performance.

Conclusions: Adjustment for surgeon experience is necessary for accurate interpretation of metrics over the course of a surgeon's career. Patient- and surgeon-adjusted control charts provide an accurate method of monitoring individual operative efficiency.

Main text word count: 2533

Key Words: performance curve, surgical performance improvement, performance monitoring, control chart, total knee replacement, operative efficiency

Article focus

- In the UK, surgeons have recently been required to publish their individual performance data. In turn, monitoring of performance data is known to improve outcomes.
- Adjustment of metrics for patient case-mix significantly alters their interpretation. Surgical procedures have learning curves, owing to the technical nature of the specialty, however the effects of this on outcomes are currently not adjusted for.
- Our aim was to develop an innovative graphical tool for accurate measurement and monitoring of individual surgeon performance metrics, adjusted for both surgeon-specific and patient-specific factors.

Key messages

- Patient- and surgeon- adjusted control charts can be used to accurately assess trends in performance metrics, with the aim of informing professional development.
- Surgeon-specific adjustment is necessary for correct assessment of operative efficiency and performance metrics; failure to do so exposes metrics to statistically significant misinterpretation.

Strengths and limitations of this study

- Use of high-granularity data on over 5000 procedures, across 14 years, performed by 17 surgeons; robust statistical demonstration of the effects of adjusting for surgeon-specific factors.
- Single centre, retrospective, and focused on operative time, which although has clear relevance to operative efficiency and financial costs, is not a clear patient-centred outcome.

6

INTRODUCTION

Increasing patient demands, costs and emphasis on safety have led to marked interest in performance tracking of individual healthcare providers [1, 2, 3, 4, 5]. While the adoption of techniques to monitor surgeons has lagged behind that of providers in other areas of medicine, we are now witnessing the insinuation of such tools into the surgical sphere. In the United Kingdom, surgeons have recently been required to publish their individual performance data. There has also been a growing interest, amongst both patients and health authorities, to track surgeons' individual performance. Two expressed concerns have been that surgeons will be averse to operating on high-risk patients (those demonstrating unfavorable case-mix), and less likely to take on junior trainees (those with unfavorable experience) during procedures for fear of poor performance results [1, 6, 7, 8]. Without addressing such concerns, it is possible that publication of performance data may result in unwanted changes in practice, and the generation of inaccurate, inequitable data.

Several methods of individual surgeon performance monitoring have been proposed [9, 10, 11], frequently adjusting for patient-specific characteristics, or case-mix, including demographics and medical co-morbidities. The influence of such characteristics upon surgical outcomes has been well-explored and acknowledged [12, 13, 14]. Very recently, the role of surgeon-specific factors such as operative experience and surgical team familiarity, with respect to outcomes, has also been elucidated [15, 16, 17].

The relative importance of adjusting for both patient- and surgeon-specific factors when assessing operative performance has historically been stymied by difficulties in generating research databases of sufficient granularity and robustness to carry out detailed statistical analyses. Such limitations have been ameliorated by the development of large depositories of electronic medical data. Through the parsing of such data, we are now fortunate to have the opportunity to perform inquiries once thought impossible.

While it is important that surgeons are monitored, and clear that measurements result in improvements [1, 6], it is paramount that the 'measuring stick' with which performance is evaluated, is accurate. To address the aforementioned confounders of case-mix and surgeon-experience, we explored the use of patient- and surgeon-adjusted control charts to permit accurate performance tracking of individual surgeons.

Control charts, a tool initially devised in the manufacturing industry, permit iterative improvements in quality by statistical process control [18]. They comprise of mapping a process metric or outcome on a chart with a pre-determined benchmark. Upper and lower limits are placed on the chart, typically at two or three standard deviations from the benchmark value; exceeding these limits indicates an anomalous or unlikely event that is signaled, investigated and when necessary acted upon to improve the charted process. In health care they have demonstrated benefit by permitting identification of causes of variation and enabling safety monitoring [19]. Control charts have been demonstrated to result in improved health outcomes, efficiency and safety [20, 21].

We believe that when appropriately adjusted, control charts may offer similar benefits in the sphere of surgical efficiency and performance. Specifically, our aim was to determine whether such a tool, when adjusted for both surgeon- and patient-specific factors, would significantly alter interpretation of performance data. We present the results of this research endeavor as a proof-of-concept of the potential value of adjusted control charts over more traditional methodologies in performance monitoring.

METHODS

Study design and population

Following institutional review board approval (protocol 2006p000586), we conducted a longitudinal analysis of 5,313 knee replacement procedures performed by 17 surgeons at a single academic tertiary care center (Brigham & Women's Hospital, Boston, MA, USA) between 1996 and 2009. In order to develop the performance chart independently from the data to be monitored, the database was randomly split into training and testing datasets according to surgeon identity [22]. The training dataset included 3,756 procedures performed by 13 surgeons and was used to define the baseline parameters of the control charts, including upper and lower limits for the charts, as well as models for case-mix adjustment, and experience-adjustment. The testing dataset included 1,557 procedures performed by the 4 other surgeons; operative times of procedures by each of these surgeons was mapped on control charts developed using the training dataset, to simulate performance monitoring.

Outcome measures and data collection

Data were culled from a combination of electronic medical records, an electronic operative time tracking application, and physician employee databases. Operative time, the primary outcome measurement, was measured in minutes and defined as the time elapsed from skin incision to skin closure. Operative time was used as a proxy for operative efficiency. For each procedure, the length of experience of each participating surgeon was calculated. The operative experience of the attending surgeon was calculated as the difference between the date of the procedure and that of the surgeon's completion of training.

Performance curve modeling and case-mix adjustment

We used the training dataset to determine the adjusted performance curve of surgeons during their career based on a multivariate generalised estimating equation (GEE) regression model, also taking into consideration the clustering of patients by surgeon [23]. Operative time was the outcome of interest, while surgeon experience was the predictor and patient case mix (patient age, sex, smoking status and the presence of comorbidities – type II diabetes mellitus, coronary artery disease and chronic obstructive pulmonary disorder) was considered as a covariate in the final model. We included all available co-morbidities from our dataset in the case-mix adjustment model. Because the operative time curve may not necessarily be a linear function of surgeon experience, a number of possible shapes of performance curves were tested. In order to obtain the best fitting shape, surgeon experience was entered as both a linear term and a quadratic term in the final model [24]. An adjusted performance curve was drawn versus the number of years since surgeon graduation. Also, the reduction in operative time independently associated with the attending surgeon's experience was plotted. Model estimates were obtained using the

GENMOD procedure in SAS™ 9.2 (SAS Institute Inc., Cary, NC, USA); all tests were 2-tailed, and *p*-values <0.05 were considered significant.

Design and comparison of patient-risk and fully-adjusted control charts

Operative times from procedures in the testing dataset were adjusted using models derived from the training dataset. Operative times were adjusted for patient-risk alone, or for patient-risk and surgeon experience together (fully-adjusted). For every surgeon, adjusted outcomes at a given year of experience were calculated as the ratio between the observed and the expected operative time multiplied by the overall mean operative time. Once adjusted, operative time was then plotted on a Shewhart control chart to simulate prospective outcome monitoring for every individual surgeon over the course of his/her career [25]. Each data point depicted the surgeon operative time per year since graduation. The central line of the patient-risk adjusted chart was constant and was determined based on the overall mean operative time, while the central line value of the fully-adjusted chart varied and depicted the adjusted performance curve of surgeons as a function of their previous experience that was previously generated from the training dataset. Control and warning limits were set at 3 SD and 2 SD around the central line, respectively, to indicate whether a particular surgeon's performance differed significantly from this goal. The detection of a significantly high or poor performance was defined as a single point outside the control limits, respectively, or two out of three successive points between a warning limit and a control limit on the same side of the central line [26]. Underperforming surgeons were positioned above the upper limits (i.e. longer operative time), while surgeons with unusually good results were below the lower limits (shorter operative time) [27].

The agreement between patient-risk and fully-adjusted charts in detecting indicator variations was measured using the weighted Cohen k statistic [28]. The positions of the data points for surgeon individual performance were compared in terms of 5 ordinal levels based on warning and control limits.

RESULTS

A total of 5,313 total knee replacement procedures performed by 17 surgeons were analysed. Median surgical experience was 17 years, ranging from 1 to 35 years in practice since graduation. The mean operative time was 109 minutes (standard deviation 30.3 minutes). A substantial decline in risk-adjusted operative time was observed over the course of surgeon's career, resulting in a concave shaped performance curve (P<0.0001, Figure 1; to maintain anonymity, the number of years of experience has been removed from the x-axis of all figures). Table 1 summarises the surgeon and patient characteristics of the training and testing datasets; Table 2 displays the number of procedures performed by each surgeon.

Figure 2 demonstrates the patient-risk adjusted operative times of 4 surgeons for TKR with respect to the expected 'benchmark' performance curve over time, with slower operative times being placed above the benchmark, indicating reduced operative efficiency, and faster operative times being placed below the benchmark, indicating improved operative efficiency. Inspection of each surgeon's performance curve revealed that surgeon A displayed better operative efficiency than surgeon B, with lower operative times. Furthermore, surgeon A's operative efficiency, unlike surgeon B's, was better than expected, demonstrating operative times below the benchmark performance curve. Surgeons C and D demonstrated similar operative times. With respect to the benchmark performance curve, however, Surgeon C performed better than expected, given his low experience, displaying superior operative efficiency. Surgeon D, relative to the benchmark performance curve, was found to display worse than expected operative efficiency.

Control charts adjusted only for patient risk depicted Surgeon A as improving. However, after adjusting for surgical experience, Surgeon A's operative efficiency appeared to worsen relative to the population mean (Figure 3). Similarly, Surgeon B was found to be within control limits for most of his operations when considering patient-risk adjustment only; however, surgical experience adjustment showed all but one of his data-points to lie outside of the upper control limit, indicating consistently slower operative efficiency. Experience adjustment transposed Surgeon C's performance curve from variable about the population mean to clearly 'high' operative efficiency, and inverted the interpretation of Surgeon D's operative efficiency from 'high' to 'poor'. The agreement between the patient-risk and fully-adjusted control charts in detecting operative time variations over surgeons career was very low (Table 3, $\kappa = 0.29$ [95% CI, 0.11-0.47], indicating the significant effect of adjusting operative time for surgical experience.

Adjustment for surgeon experience, in addition to patient risk, significantly altered the interpretation of operative efficiency, and enhanced the accuracy of assessing a surgeon's improvement or diminishment in efficiency over time.

DISCUSSION

Our study presents a novel methodology for the adjustment and monitoring of surgical metrics, specifically operative efficiency. Surgery is a technical specialty, improving with volume and experience. Consideration of surgeon-specific factors may seem intuitive, but remains poorly investigated. Our findings quantitatively demonstrated that such adjustment significantly altered the interpretation of operative efficiency monitoring – at times resulting in an inversion of perceived trends in efficiency relative to consideration of patient-adjustment alone. Although we investigated operative time, this methodology can be applied to any surgical outcome.

Fully-adjusted control charts were shown to offer an accurate and perceptive means of interpreting trends in a surgeon's efficiency, identifying outlying or anomalous units and providing early warning of divergence from the cohort mean. We believe such factors may prove particularly advantageous in the context of surgeon monitoring and performance tracking.

Limitations

These implications must be considered in the context of this study's limitations. First, this investigation was performed at a single academic medical center, retrospectively, which may limit the representativeness of our sample. It should be noted, however, that the retrospective nature of our investigation removed any Hawthorne bias with regards to performance and therefore may provide a truer depiction of procedural dynamics than could have been ascertained through prospective methods. Importantly, our database covered a substantial time period and it is possible that operative technique and technology may have changed during this. Second, our focus on operative time did not implicitly incorporate considerations related to patient outcomes. Studies, however, have indicated that faster completion of the TKR procedure is associated with better outcomes [29]. Indeed, in a variety of work both within surgery and outside, time of task completion has been used as a robust indicator of learning and outcome [17, 30, 31, 32, 33, 34]. Further, longer operative times and increased use of the operating theatre, as discussed below, expose patients to greater risks of surgical site infection, whilst also entailing larger financial costs and reduced efficiency, that amidst rising costs in health care, are of clear importance. Third, our investigation utilised years of training as a proxy for surgical experience, rather than number of cases performed. This limitation is a reflection of the fact that a number of surgeons included in the dataset had been in clinical practice prior to the implementation of our surgical tracking application. Years of training, however, has been utilised as an acceptable substitute for surgical experience in prior published studies [15, 16, 17]. Finally, although we adjusted control charts for patient characteristics and surgeon individual experience, there may be other factors, such as non-technical skills, teamwork and resident involvement which need to be accounted for to enable better interpretation of performance.

Policy Implications

Monitoring operative times with the aim of improving operative efficiency has strong financial implications. Theatres, excluding day surgery, have been shown to account for approximately 6% of NHS Trust budgets, equating to running-costs per theatre per year of £1.5 million [35, 36]. In the U.S., the cost of an operating theatre has been estimated at approximately $130 per minute [37]. Strategies that can improve operative efficiency and reduce operative costs are therefore of importance.

The impact of surgeon-specific adjustment itself, also has implications both within and outside the sphere of individual performance-tracking. In the context of training, it is arguably inequitable to compare the performance of less and more experienced trainees; indeed it would be inappropriate to expect a surgeon who has just started their training to perform as well as a surgeon who is about to complete their training, rather, trainees must be compared to cohorts of the same level of experience. Thus, use of surgeon-specific, experience-adjusted charts will permit the performance of young trainees to be accurately and equitably monitored relative to a relevant benchmark, removing bias. This could give rise to appraisals based upon performance rather than career chronology or volume of cases alone, potentially ensuring progression only upon acquisition of sufficient expertise. The tools outlined in this study could furthermore be used to establish minimal competency requirements for operators and permit important contributors to training to be quantitatively identified. In the context of experienced surgeons, fully-adjusted control charts provide a sensitive and timely means of identifying deviations from expected benchmarks, permitting prompt investigation or intervention to improve the respective surgeon's performance. Recent work has also shown that performance may decline as surgeons approach seniority [38]; patient- and surgeon- adjusted control charts have the capacity to identify this deterioration and supplement the implementation of continuing education programs. In the broader context of outcomes research, any studies investigating the impact of an intervention on performance could gain interpretational benefit from surgeon-factor adjustment. Where groups of surgeons or departments are being compared, it is intuitive that adjustment for the respective experiences, or 'surgeon-mix' of these groups are adjusted for, in parallel with patient-mix adjustment, to improve the transparency of results.

We present this research as a proof-of-concept that i) patient- and surgeon-adjusted control charts can be utilised to inform ongoing professional development and feedback for individual surgeons, and ii) surgeon-specific adjustment is necessary for correct assessment of operative efficiency and performance outcomes; failure to do so exposes metrics to statistically significant misinterpretation. We believe this should be considered in developments regarding surgical monitoring, to permit equitable and accurate performance assessment, addressing concerns of patients, surgeons and policy-makers alike.

REFERENCES

1 Godlee F. Measure your team's performance, and publish the results [Editor's Choice]. BMJ 2012;345:e4590.

2 Davis K, Schoen C, Guterman S, et al. Slowing the Growth of U.S. Health Care Expenditures: What Are the Options? New York: The Commonwealth Fund. 2007

3 Commonwealth Fund Commission on a High Performance Health System. Why Not the Best? Results from a National Scorecard on U.S. Health System Performance, 2008. New York: The Commonwealth Fund. 2008.

4 Dimick JB, Weeks WB, Karia R, et al. Who pays for poor surgical quality? Building a business case for quality improvement. J Am Coll Surg. 2006;202:933.

5 Duclos A, Carty MJ. Value of health care delivery. JAMA 2011;306(3):267.

6 Tavare A. Where are we with transparency over performance of doctors and institutions? BMJ. 2012 Jul 3;345:e4464.

7 Ray S, Simpson I. Professional societies can lead the way on transparency but will need support. BMJ. 2012 Jul 31;345:e5075.

8 Hill M. NHS medical director wants surgeon league tables. BBC News [online]. December 5 2012. Available from: http://www.bbc.co.uk/news/health-20584897 [Accessed 4th March 2013].

9 Holzhey DM, Jacobs S, Walther T, Mochalski M, Mohr FW, Falk V. Cumulative sum failure analysis for eight surgeons performing minimally invasive direct coronary artery bypass. J Thorac Cardiovasc Surg. 2007

10 Kusamura S, Baratti D, Deraco M. Multidimensional analysis of the learning curve for cytoreductive surgery and hyperthermic intraperitoneal chemotherapy in peritoneal surface malignancies. Ann Surg. 2012 Feb;255(2):348-56.

11 Tekkis PP, McCulloch P, Steger AC, Benjamin IS, Poloniecki JD. Mortality control charts for comparing performance of surgical units: validation study using hospital mortality data. BMJ. 2003 Apr 12;326(7393):786-8.

12 Duclos A, Voirin N, Touzet S, Soardo P, Schott AM, Colin C, Peix JL, Lifante JC. Crude versus case-mix-adjusted control charts for safety monitoring in thyroid surgery. Qual Saf Health Care. 2010 Dec;19(6):e17.

13 Cook JA, Ramsay CR, Fayers P. Statistical evaluation of learning curve effects in surgical trials. Clinical Trials 2004;1:421-7.

14 High-Volume versus Low-Volume for Esophageal Resections for Cancer: The Essential Role of Case-Mix Adjustments based on Clinical Data. Michael W. Wouters, Bas P. Wijnhoven, Henrieke E. Karim-Kos, Harriet G. Blaauwgeers, Laurents P. Stassen, Willem-Hans Steup, Huug W. Tilanus, Rob A. Tollenaar Ann Surg Oncol. 2008 January; 15(1): 80–87.

15 Carty MJ, Chan R, Huckman R, Snow D, Orgill DP. A detailed analysis of the reduction mammaplasty learning curve: a statistical process model for approaching surgical performance improvement. Plast Reconstr Surg. 2009 Sep;124(3):706-14.

16 Xu R, Carty MJ, Orgill DP, Lipsitz SR, Duclos A. The Teaming Curve: A Longitudinal Study of the Influence of Surgical Team Familiarity on Operative Time. Ann Surg. 2013 Feb 12.

17 Elbardissi AW, Duclos A, Rawn JD, Orgill DP, Carty MJ. Cumulative team experience matters more than individual surgeon experience in cardiac surgery. J Thorac Cardiovasc Surg. 2012 Oct 17.

18 Bewick DM. Controlling variation in health care: a consultation from Walter Shewhart. Med Care. 1991 Dec;29(12):1212-25.

19 Tennant R, Mohammed MA, Coleman JJ, Matin U. Monitoring patients using control charts: a systematic review. Int J Qual Health Care. 2007 Aug;19(4):187-94.

20 Thor J, Lundberg J, Ask J, Olsson J, Carli C, Härenstam KP, Brommels M. Application of statistical process control in healthcare improvement: systematic review. Qual Saf Health Care. 2007 Oct;16(5):387-99.

21 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 Mar;99(3):324-35.

22 Hastie T, Tibshirani R, Friedman J. (2009) The elements of statistical learning. New York, NY: Springer.

23 Liang KY, Zeger SL. Longitudinal Data Analysis Using Generalized Linear Models. Biometrika 1986;73:13-22.

24 Ramsay CR, Grant AM, Wallace SA, Garthwaite PH, Monk AF, et al. (2001) Statistical assessment of the learning curves of health technologies. Health Technol Assess 5(12): 1-79.

25 Montgomery DC. Statistical Quality Control: A Modern Introduction. Hoboken, NJ: Wiley, 2008.

26 Benneyan JC, Lloyd RC, Plsek PE. Statistical process control as a tool for research and healthcare improvement. Qual Saf Health Care 2003;12(6): 458-64.

27 Duclos A, Voirin N. The p-control chart: a tool for care improvement. Int J Qual Health Care. 2010;22(5):402-407.

28 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968 Oct;70(4):213-20.

29 Peersman G, Laskin R, Davis J, Peterson MG, Richart T. Prolonged operative time correlates with increased infection rate after total knee arthroplasty. HSS J. 2006 Feb;2(1):70-2.

30 Pisano GP, Bohmer MJ, Edmondson AC. 2001. Organizational differences in rates of learning: Evidence from the adoption of minimally invasive cardiac surgery. Management Science, 47, 752-768.

31 Edmondson A, Winslow A, Bohmer R, Pisano G. 2003. Learning how and learning what: Effects of tacit and codified knowledge on performance improvement following technology adoption. Decision Sciences, 34, 197-223.

32 Epple D, Argote L, Devadas R. 1991. Organizational learning curves: A method for investigating intra-plant transfer of knowledge acquired through learning by doing. Organization Science, 2, 58-70.

33 Argote L, Epple D. Learning curves in manufacturing. Science 1990; 247, 920-924.

34 Peersman G, Laskin R, Davis J, Peterson MG, Richart T. Prolonged operative time correlates with increased infection rate after total knee arthroplasty. HSS J. 2006 Feb;2(1):70-2.

35 http://www.isdscotland.org/Health-Topics/Finance/Costs/Detailed-Tables/Theatres.asp [Accessed August 30, 2013]

36 West Hertfordshire Hiospitals NHS Trust. Accessed from: http://www.westhertshospitals.nhs.uk/foi_publication_scheme/disclosure_log/2010/december/documents/170%20-%20140111.pdf [Accessed August 30, 2013]

37 Shippert RD. A study of time-dependent operating room fees and how to save $100,000 by using time-saving products. *American Journal of Cosmetic Surgery*. 2005;22(1):25–34.

38 Duclos A, Peix JL, Colin C, Kraimps JL, Menegaux F, Pattou F, Sebag F, Touzet S, Bourdy S, Voirin N, Lifante JC; CATHY Study Group. Influence of experience on performance of individual surgeons in thyroid surgery: prospective cross sectional multicentre study. BMJ. 2012 Jan 10;344:d8041.

Figure 1. Improvement curve for individual surgeons

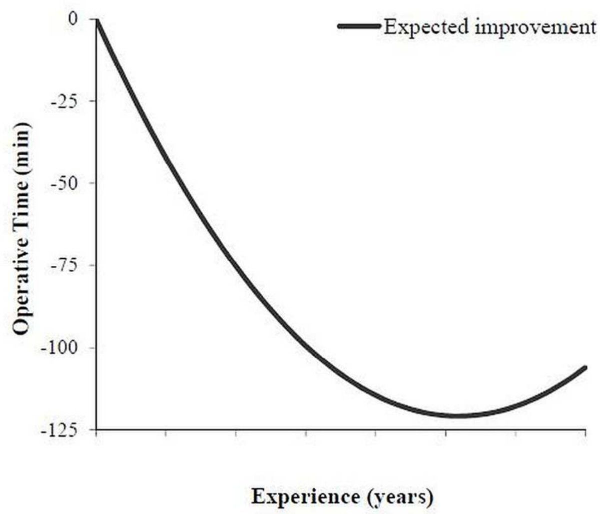


Figure 1. Performance curve for individual surgeon TKR operative efficiency. The graph illustrates how operative time within the cohort changed with surgeon experience.
132x90mm (300 x 300 DPI)

Figure 2. Individual performance for surgeons A-D

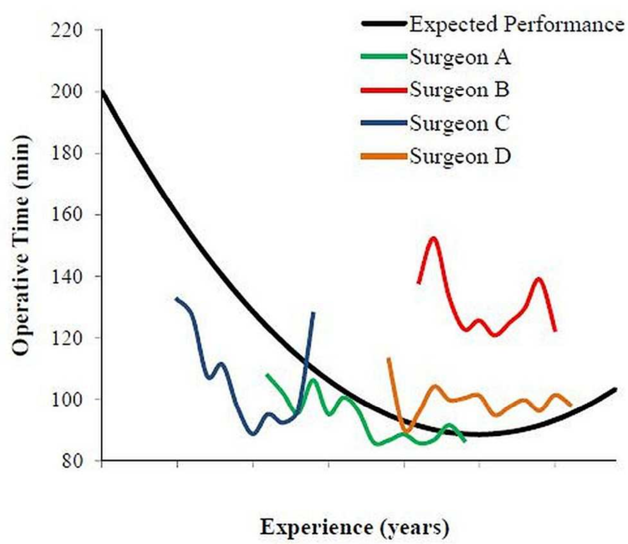


Figure 2. Individual performance curves for surgeons A-D. The graph illustrates the patient-risk adjusted operative times of the 4 surgeons selected to test the control charts, with respect to the expected 'benchmark' performance curve.
123x90mm (300 x 300 DPI)

Figure 3: Patient-risk vs. fully- adjusted control charts for surgeons A-D

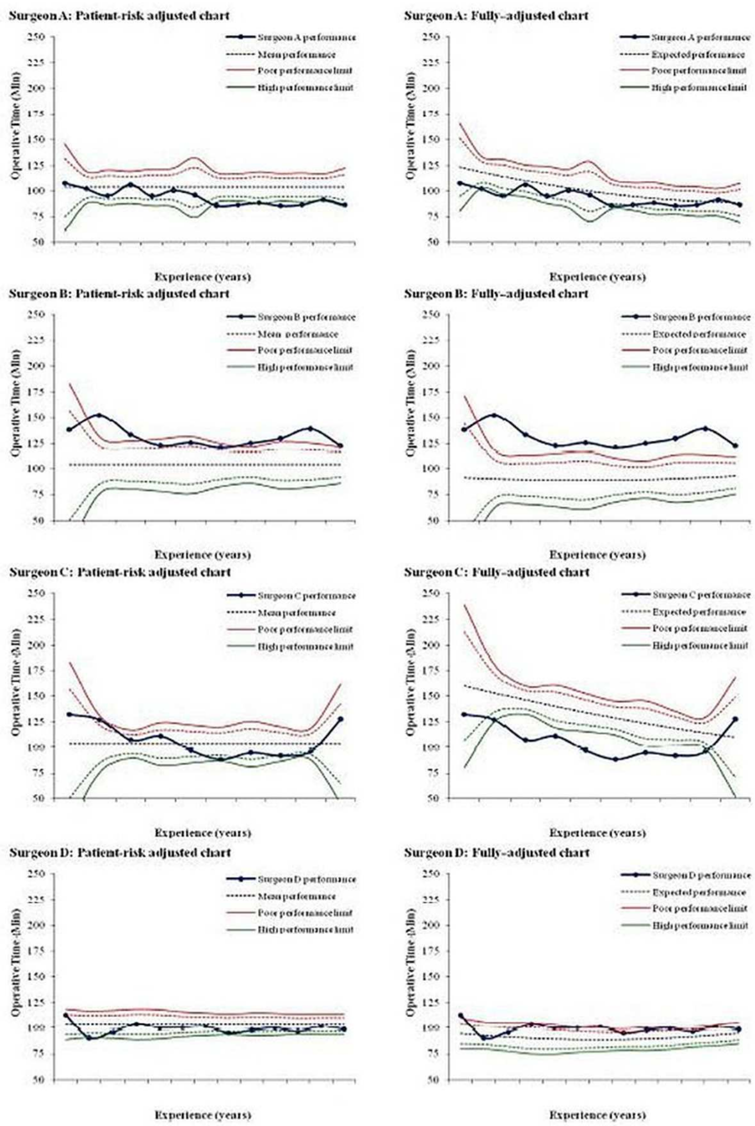


Figure 3. Patient-risk vs. fully- adjusted control charts for individual surgeons. For each surgeon a patient-risk adjusted chart, and fully-adjusted (patient-risk and surgeon-experience adjusted) chart is displayed. The horizontal axes indicate the experience of the surgeon in years and the blue curve his/her adjusted performance over time. The central black dotted line represents the expected operative time over the course of surgeon's career. The upper red and lower green lines illustrate poor and high performance limits, set at two standard deviations (dotted warning limits) and three standard deviations (continuous control limits) around the central line. Poor and high performers are defined as those breaching the upper and lower limits, respectively. Average performers are those with operative time around the central line, without crossing the limits.
66x90mm (300 x 300 DPI)

Table 1. Overview of study participants

		Training dataset	Testing dataset
Attending surgeon	**(N=17)**	**(N=13)**	**(N=4)**
Surgeon experience, years, Median *(Min-Max)*	17 *(1-35)*	15 *(1-35)*	23 *(6-32)*
Surgeon volume of cases, Median *(Min-Max)*	176 *(10-1,871)*	144 *(10-1,871)*	319 *(157-761)*
Surgical cases	**(N=5,313)**	**(N=3,756)**	**(N=1,557)**
Patient female gender, No. *(%)*	3,558 *(67.0)*	2,543 *(67.7)*	1,015 *(65.2)*
Patient age, years, Mean *(SD)*	66.2 *(11.3)*	65.8 *(11.4)*	67.0 *(11.0)*
Patient with comorbidity, No. *(%)*	3,388 *(63.8)*	2,440 *(65.0)*	948 *(60.9)*
No. of comorbidities, Median *(Min-Max)*	1 *(0-6)*	1 *(0-6)*	1 *(0-5)*
Coronary artery disease, No. *(%)*	1,074 *(20.2)*	751 *(20.0)*	323 *(20.7)*
Chronic obstructive pulmonary disease, No. *(%)*	320 *(6.0)*	230 *(6.1)*	90 *(5.8)*
Diabetes mellitus, No. *(%)*	858 *(16.1)*	636 *(16.9)*	222 *(14.3)*
Hypertension, No. *(%)*	2,196 *(41.3)*	1,609 *(42.8)*	587 *(37.7)*
Obesity, No. *(%)*	1,242 *(23.4)*	935 *(24.9)*	307 *(19.7)*
Tobacco, No. *(%)*	814 *(15.3)*	590 *(15.7)*	224 *(14.4)*
Operative time, minutes, Mean *(SD)*	109.2 *(30.3)*	103.5 *(29.8)*	123.0 *(26.9)*

Table 2. Number of total knee replacements performed by each surgeon.

Attending surgeon	Number of cases
1*	11
2*	61
3**	427
4*	63
5**	157
6*	264
7**	212
8*	184
9*	144
10*	367
11*	66
12*	176
13*	10
14*	59
15*	478
16**	761
17*	1871

*Testing dataset, **Training dataset

Table 3. Agreement between patient-risk adjusted and fully-adjusted charts in detecting indicator variations.

		Fully-adjusted chart*					
		<LCL	LCL-LWL	LWL-UWL	UWL-UCL	>UCL	Total
Patient-risk adjusted chart*	<LCL	0	1	4	0	0	5
	LCL-LWL	2	0	5	0	0	7
	LWL-UWL	7	0	10	6	2	25
	UWL-UCL	0	1	0	0	3	4
	>UCL	0	0	0	0	6	6
	Total	9	2	19	6	11	47

* Each unit in the table represents the position of a data point on a control chart, according to 5 ordinal levels based on Warning Limits (2SD) and Control Limits (3SD), as follows: <LCL (below the lower control limit), LCL-LWL (between the lower control and warning limits), LWL-UWL (between the lower and upper warning limits), UWL-UCL (between the upper warning and control limits), >UCL (above the upper control limit).

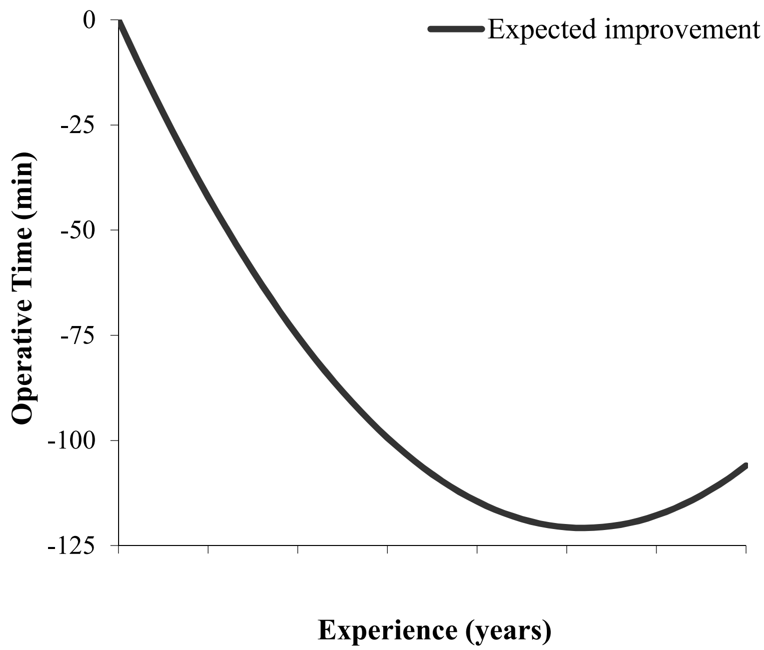
FIGURE LEGENDS

Figure 1. Performance curve for individual surgeon TKR operative efficiency. The graph illustrates how operative time within the cohort changed with surgeon experience.

Figure 2. Individual performance curves for surgeons A-D**.** The graph illustrates the patient-risk adjusted operative times of the 4 surgeons selected to test the control charts, with respect to the expected 'benchmark' performance curve.

Figure 3. Patient-risk vs. fully- adjusted control charts for individual surgeons. For each surgeon a patient-risk adjusted chart, and fully-adjusted (patient-risk and surgeon-experience adjusted) chart is displayed. The horizontal axes indicate the experience of the surgeon in years and the blue curve his/her adjusted performance over time. The central black dotted line represents the expected operative time over the course of surgeon's career. The upper red and lower green lines illustrate poor and high performance limits, set at two standard deviations (dotted warning limits) and three standard deviations (continuous control limits) around the central line. Poor and high performers are defined as those breaching the upper and lower limits, respectively. Average performers are those with operative time around the central line, without crossing the limits.

Figure 1. Improvement curve for individual surgeons



6

Figure 2. Individual performance for surgeons A-D

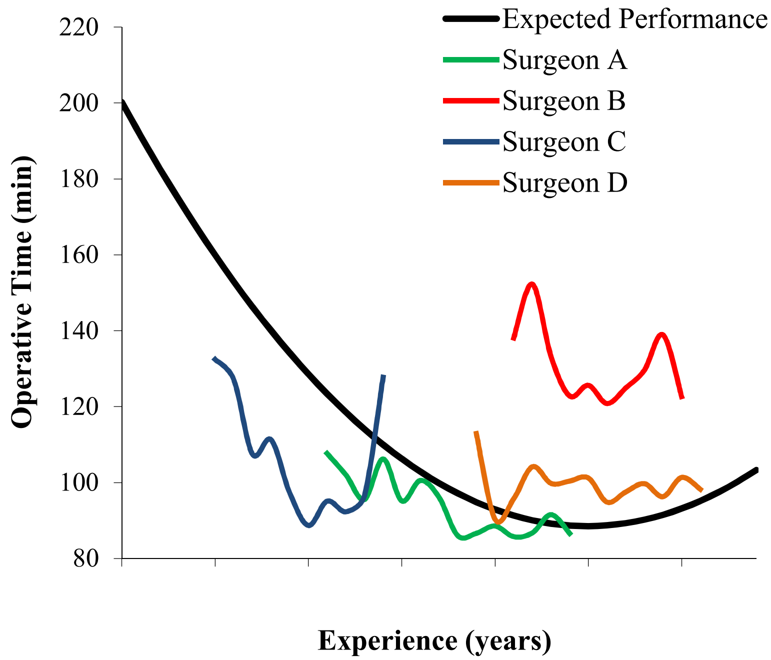


Figure 3: Patient-risk vs. fully- adjusted control charts for surgeons A-D