Health Policy 122 (2018) 660–666

Contents lists available at ScienceDirect

Health Policy

journal homepage: www.elsevier.com/locate/healthpol

Should interventions to reduce variation in care quality target doctors or hospitals?

Nils Gutacker[a,*], Karen Bloor[b], Chris Bojke[c,1], Kieran Walshe[d]

[a] Centre for Health Economics, University of York, Heslington, York, YO10 5DD, United Kingdom
[b] Department of Health Sciences, University of York, United Kingdom
[c] Centre for Health Economics, University of York, United Kingdom
[d] Alliance Manchester Business School, University of Manchester, United Kingdom

ARTICLE INFO

Article history:
Received 27 February 2017
Received in revised form 16 October 2017
Accepted 6 April 2018

Keywords:
Performance measurement
Hospital
Doctor
Reliability
Multilevel modelling

ABSTRACT

Interventions to reduce variation in care quality are increasingly targeted at both individual doctors and the organisations in which they work. Concerns remain about the scope and consequences for such performance management, the relative contribution of individuals and organisations to observed variation, and whether performance can be measured reliably.

This study explores these issues in the context of the English National Health Service by analysing comprehensive administrative data for all patients treated for four clinical conditions (acute myocardial infarction, hip fracture, pneumonia, ischemic stroke) and two surgical procedures (coronary artery bypass, hip replacement) during April 2010–February 2013. Performance indicators are defined as 30-day mortality, 28-day emergency readmission and inpatient length of stay. Three-level hierarchical generalised linear mixed models are estimated to attribute variation in case-mix adjusted indicators to individual doctors and hospital organisations.

Except for length of stay after hip replacement, no more than 11% of variation in case-mix adjusted performance indicators can be attributed to doctors and organisations with the rest reflecting random chance and unobserved patient factors. Doctor variation exceeds hospital variation by a factor of 1.2 or more. However, identifying poor performance amongst doctors is hampered by insufficient numbers of cases per doctor to reliably estimate their individual performances. Policy makers and regulators should therefore be cautious when targeting individual doctors in performance improvement initiatives.

1. Introduction

Large variations in the quality of health care have been reported over many years, and in many countries [1,2]. Policymakers and professional bodies have responded to such variations with a variety of mechanisms including measurement ('profiling'), monitoring, public reporting, regulation and incentives (financial and non-financial) [3,4]. These interventions have mostly been focused on organisational performance, particularly at the level of the hospital or clinical specialty, with the implicit assumption that the variation results from factors that can be influenced or affected by organisations and those who lead them.

Increasingly, interventions to improve care quality and reduce variations operate not just at organisational level but at the level of individual doctors. For example, a number of initiatives have been introduced with the aim of improving hospital specialists' mortality rates through measurement, public reporting and feedback, most notably in cardiac surgery in the UK and US [5,6]. In the National Health Service (NHS) in England, this has been extended to routine publication of outcome data for consultants (fully-trained hospital specialists) working in 13 specialities [7,8].

Despite substantial investments in these mechanisms intended to drive improvements in the quality of care, considerable uncertainty exists about whether individual consultants or the organisations in which they work are more important as drivers of variation in the quality of health care. The utility of information derived from administrative data for individual or organisational performance management purposes, and the potential for unin-

[*] Corresponding author.
E-mail addresses: nils.gutacker@york.ac.uk (N. Gutacker), karen.bloor@york.ac.uk (K. Bloor), c.bojke@leeds.ac.uk (C. Bojke), kieran.walshe@manchester.ac.uk (K. Walshe).
[1] Present address: Leeds Institute of Health Sciences, University of Leeds, United Kingdom.

https://doi.org/10.1016/j.healthpol.2018.04.004
0168-8510/

tended consequences, remain poorly understood. For example, there is only limited UK evidence on the degree of performance variation among doctors for outcomes other than mortality [9]. In addition, the assessment of performance of individual consultants raises a statistical concern: estimates of their performance are more vulnerable to chance events than those of hospitals because they are based on smaller patient populations [10]. A number of studies have suggested that using indicators at individual level may result in often unreliable estimates of true performance [11–15]. Unreliable estimates may result in incorrect decisions about doctors' performance with potentially adverse consequences for individual careers, the welfare of patients, and the credibility of the measurement process.

This paper explores these issues in the context of the English NHS, extending a previous analysis of mortality variation in England [14] and also focusing on two performance indicators (PIs) not previously analysed: emergency re-admission within 28 days of discharge and inpatient length of stay. The analysis seeks to answer two questions. First, how much variation in observed PIs can be attributed to individual hospital consultants and how does this compare with that attributable to the organisations in which they work? Second, are performance estimates for individual consultants sufficiently reliable to be useful estimates of their true performance?

2. Methods

2.1. Study population

We used data from Hospital Episode Statistics (HES) on all NHS-funded inpatient care provided in hospitals in England between April 2010 and February 2013. We focused on six conditions/procedures that were selected because they are based on validated indicators used internationally [16,17], they cover a range of clinical areas and are either part of the consultant-level reporting initiative in England [8] or constitute a substantial proportion of NHS activity: emergency admissions for treatment of acute myocardial infarction (AMI), acute ischemic stroke (AIS), pneumonia and hip fracture; and elective admissions for unilateral primary (i.e. non-revision) hip replacement and isolated coronary artery bypass graft (CABG) surgery. These groups were constructed following US Agency for Healthcare Research and Quality's inpatient quality indicator (IQI) definitions (IQI#12, #14, #15, #17, #19, #20), which were recently amended for use with English NHS data as part of a European study of health care variations [16]. A full list of relevant ICD-10 diagnosis codes and OPCS-4 procedure codes are reported in the Appendix. Patients were excluded if they were younger than 18 years at the time of admission (<40 years for CABG surgery; <65 for hip fracture) or were living outside of England.

HES records inpatient activity at the level of Finished Consultant Episodes (FCEs), which we linked to create continuous inpatient spells that cover the entire period from admission to discharge (including transfers between hospitals). Data were extracted on all inpatient activity 365 days before index admission and 28 days after discharge (up to 31st March 2013). Record linkage was based on unique NHS identification numbers. Admission spells were assigned to the first consultant responsible for treatment after the index admission. Consultants who provided care in different hospital organisations were treated as separate units of observation. This issue was most prevalent in elective hip replacement surgery, where consultants often work both in NHS hospital trusts and privately operated Independent Sector Treatment Centres. Consultants were identified through their unique General Medical Council (GMC) code. These codes were validated against the GMC database of registered specialists and the Electronic Staff Record system and invalid records were excluded from analysis. Consultants (and their patients) were excluded if they treated less than 30 cases over the three-year period. Similarly, hospitals were excluded if they treated less than 90 cases over this period.

2.2. Performance indicators

We investigated variation in important clinical outcomes and process of care measures that are commonly used as PIs. The clinical outcomes were 28-day all-cause emergency readmission and 30-day all-cause mortality, which was derived from Office for National Statistics date of death data. Length of inpatient stay, measured as the number of overnight stays, was used to approximate the effectiveness of discharge management processes. To reduce the influence of potential miscoding values exceeding the 99th percentile of the distribution of length of stay were replaced with the 99th percentile.

2.3. Case-mix adjustment

All PIs were adjusted for age (5-year bands with separate categories for <25 and ≥85; except in the analysis of mortality in which lowest category is <60), sex, age-sex interactions and year of hospitalisation. Severity adjustment was limited to information contained in administrative records and included an indicator for any hospital emergency admission in the previous year, as well as the number of Elixhauser co-morbid conditions (grouped as 0, 1, 2–3, 4 +) recorded in secondary diagnosis fields in the index admission or admissions in the previous year. Patients' socio-economic status was approximated by the proportion of residents at small area level (Lower Super Output Area; approximately average population of 1500 inhabitants) claiming means-tested social security benefits (divided into five quintile groups) [18].

2.4. Statistical analysis

Three-level hierarchical generalised linear mixed models were fitted to identify variation in PIs due to provider case-mix, additional systematic variation associated with consultants and hospital organisations, and random chance variation [19,20]. Patient episodes are nested in consultants, which are themselves nested in hospitals. Emergency readmissions and mortality were modelled using logistic regression. Length of stay was modelled as count data using Poisson models with an additional over-dispersion parameter. Separate models were estimated for each patient group and PI. Data were pooled across three years to reflect common practice in performance assessment schemes [14].

The fixed part of the model captures variation in PIs associated with observable differences in provider case-mix (see section 2.3). The model error term captures the variation in the PI that is not explained by observed patient characteristics and is further partitioned into separate components varying at patient level (i.e. unmeasured patient characteristics or random noise with variance σ^2), consultant level (τ^2) and hospital level (ω^2). From that we calculated variance partition coefficients (VPC) at the response scale by means of simulation [21,22]. Each VPC measures the proportion of unexplained variation in PIs associated with the respective level of the hierarchy. For example, the VPC at consultant level is defined as

$$VPC_{consultant} = \frac{\tau^2}{\sigma^2 + \tau^2 + \omega^2}$$

and similarly for other levels. By design, all VPCs must sum to unity. Higher values of VPC therefore indicate a larger influence on PIs relative to other levels.

In most performance assessment schemes, case-mix adjusted performance estimates are obtained by means of indirect standardisation. There is a risk that any performance estimates for individual consultants conflate true variation across consultants with random noise. The reliability (R) of performance estimates for individual consultants is a function of their case-load N and the $VPC_{consultant}$. It is calculated as

$$R_{consultant} = \frac{N \times VPC}{1 + [N - 1] \times VPC}$$

with $0 < R \leq 1$ [10]. Higher values of R indicate that estimates of individual consultants' performance are less subject to unrelated variation and are thus more suitable for performance assessment purposes. Values of ≥0.7 are often required for low-stakes applications such as confidential reports to clinicians with limited risk of punitive actions [15]. Conversely, values of ≥0.9 have been suggested for high-stakes applications such as public reporting of performance or pay-for-performance schemes. The minimum level of activity required for a given level of reliability R can be obtained by solving the above equation for N. We calculated minimum activity thresholds and the proportion of consultants fulfilling these thresholds to achieve reliability of 0.7 and 0.9, respectively.

All statistical analyses were performed in Stata 13 (StataCorp LP, College Station, TX, USA) and MLwiN 2.36 (Centre for Multilevel Modelling, University of Bristol, UK).

3. Results

A total of 1,211,983 patients were included in the initial sample. Of these, 172,826 (14.3%) patients did not fulfil the inclusion criteria, leaving 1,039,157 patients for further analysis (Table 1). These patients received care from 7197 consultants (6731 unique GMC codes) in 240 hospitals. The number of patients per consultant varied substantially within and across conditions. The lowest case-load was observed for consultants treating AIS patients (median = 55; IQR = 38–149) and the highest consultant case-load was for CABG surgery (median = 104; IQR = 72–158).

Patients in our sample were on average 73 years old and approximately half were male. The overall 28-day emergency readmission and 30-day mortality rates were 12.0% and 11.0% respectively, and patients stayed in hospital for 12.5 nights on average. There was marked variation in patient characteristics and PIs across conditions. Patients admitted for planned care were on average younger (68 vs. 75 years), stayed shorter in hospital (5.8 vs. 14.1 nights) and were at lower risk of readmission (6.2% vs. 13.4%) and mortality (0.2% vs 13.5%).

3.1. Variation across hospitals and consultants

All coefficients on case-mix variables show the expected sign and internally consistent ranking of magnitudes. The McKelvey-Zavoina Pseudo R^2 statistics [10,23,24] measure the proportion of variance in PIs explained by observed patient characteristics and range from 16.7% to 26.9% for mortality, 2.7% to 4.4% for emergency readmission, and 5.7% to 22.8% for the number of inpatient days. More detail on regression coefficients and explained variance are provided in the Appendix.

Our primary interest is in the proportion of variation not explained by case-mix and how this relates to consultants and hospital organisations. Fig. 1 shows the estimated VPCs at consultant and hospital level (stacked) for each of the three PIs and by condition. Approximately 0.6% to 4.1% of unexplained variation in the case-mix adjusted probability of readmission can be attributed to hospitals and consultants. The remainder reflects random variation at patient level that is not associated with observed patient characteristics. VPCs for mortality are of similar magnitude and range from 0.3% to 2.0%. Conversely, hospitals and consultants have a relatively larger influence over patients' length of stay. Between 1.9% and 22.6% of unexplained variation in length of stay is associated with either consultant or hospital. Note that the noticeably larger variation in length of stay after planned hip replacement surgery may reflect differences in the performance of public and private hospitals [25]; with hip replacement being the only condition studied for which this distinction is relevant.

The proportion of unexplained variation at consultant level exceeds that at hospital level by a factor of 1.2 or more, except for emergency readmission after AMI. It was not possible to differentiate consultant and hospital variation for mortality after planned hip replacement surgery as part of the estimation procedure, and the presented number should therefore be interpreted as a composite.

3.2. Reliability of consultant and hospital performance estimates

Table 2 shows the reliability of consultant performance estimates for the three PIs, the level of activity required to achieve reliability of at least 0.7 and 0.9, and the proportion of consultants that fulfil this requirement. The reliability of consultants' emergency readmission rates as indicators of their performance ranges from 0.19 to 0.71. The required 3-year activity to achieve a reliability of ≥0.7 lies between 92–563 admissions for the six conditions studied. Very few consultants achieve such case-loads. By extension, even fewer consultants reach case-loads required for a reliability of ≥0.9. A noteworthy exception is hip replacement surgery, where more than half of consultants treat a sufficient number of patients to obtain reliable estimates at r≥0.7.

Estimates of reliability and required case-load for 30-day mortality follow the same pattern.

The reliability of consultant performance estimates for length of stay is significantly higher. At median activity level, the reliability is estimated to range from 0.46 to 0.93. For each of the six conditions studied, at least 25% of consultants treat enough patients to achieve a reliability of at least 0.7. In some cases, such as cardiac surgeons performing CABG surgery, this is true for more than 90% of consultants. Between 0.4% and 70% of consultants achieve a reliability of 0.9 or more.

Table 3 reports the same information for hospital performance estimates. As hospital organisations are usually held accountable for all variation that is not attributable to case-mix and random noise, including variation that derives from consultants working for them, the reported estimates are based on the pooled VPC, calculated as $VPC_{Consultant} + VPC_{Hospital}$.

Unsurprisingly, performance estimates at hospital level are significantly more reliable than those calculated for consultants due to the substantially larger case-loads and the increased VPC. The reliability of performance estimates for a hospital of median activity levels exceeded 0.85 for all indicators and conditions. A large share of hospitals fulfils volume requirements to achieve reliability of 0.9, ranging from 29% of hospitals for emergency readmission after AMI to 100% for length of stay after bypass surgery.

4. Discussion

This study demonstrated that for the performance indicators and conditions chosen, the amount of case-mix adjusted variation that is attributable to consultants generally exceeds that which is attributable to organisations, although both are substantially outweighed by random variation at patient level that is not explained by the observed patient characteristics. In addition, we found that a large proportion of consultants do not treat a sufficient volume of patients for performance estimates based on these measures to represent reliably their underlying performance.

Table 1
Descriptive statistics of patient sample (April 2010–February 2013).

	AMI		CABG		Hip fracture		Hip replacement		Pneumonia		Stroke		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Patient level														
28-day emergency readmission (yes/no)	0.15	0.36	0.12	0.32	0.12	0.32	0.05	0.22	0.15	0.35	0.10	0.30	0.12	0.33
30-day mortality (yes/no)	0.07	0.26	0.01	0.09	0.06	0.23	0.00	0.03	0.19	0.40	0.11	0.32	0.11	0.31
Length of stay (in days)	7.74	9.09	9.08	6.67	23.51	21.34	5.32	3.99	10.62	13.07	19.65	26.17	12.52	16.83
Patient age (in years)	69.75	14.14	66.07	9.35	81.00	11.42	67.96	11.51	73.80	16.63	75.37	13.20	73.42	14.81
Male (yes/no)	0.65	0.48	0.83	0.37	0.27	0.44	0.40	0.49	0.51	0.50	0.49	0.50	0.48	0.50
Elixhauser: 0 comorbidities	0.15	0.36	0.06	0.23	0.15	0.36	0.32	0.47	0.10	0.30	0.11	0.31	0.15	0.36
Elixhauser: 1 comorbidity	0.21	0.41	0.11	0.32	0.22	0.41	0.28	0.45	0.14	0.34	0.20	0.40	0.19	0.39
Elixhauser: 2–3 comorbidities	0.33	0.47	0.35	0.48	0.35	0.48	0.29	0.45	0.29	0.45	0.37	0.48	0.32	0.46
Elixhauser: 4+ comorbidities	0.31	0.46	0.48	0.50	0.28	0.45	0.11	0.31	0.48	0.50	0.32	0.47	0.34	0.47
Emergency admission in last year (yes/no)	0.26	0.44	0.38	0.49	0.33	0.47	0.09	0.29	0.51	0.50	0.31	0.46	0.35	0.48
Number of patients	138,044		24,505		156,145		170,678		405,671		144,114		1,039,157	
Consultant level														
Number of consultants	1,746		212		1,735		1,325		3,760		1,214		9,992	
Case-load: Median	56		104		86		95		83		55		78	
Case-load: 25th percentile	39		72		60		56		52		38		47	
Case-load: 75th percentile	94		158		112		167		131		149		125	
Hospital level														
Number of hospitals	148		30		148		229		152		144		851	
Case-load: Median	787		734		1,000		649		2,471		946		946	
Case-load: 25th percentile	505		616		705.5		224		1,794		632.5		570	
Case-load: 75th percentile	1,214.5		953		1,337.5		985		3,350		1,348.5		1,571	

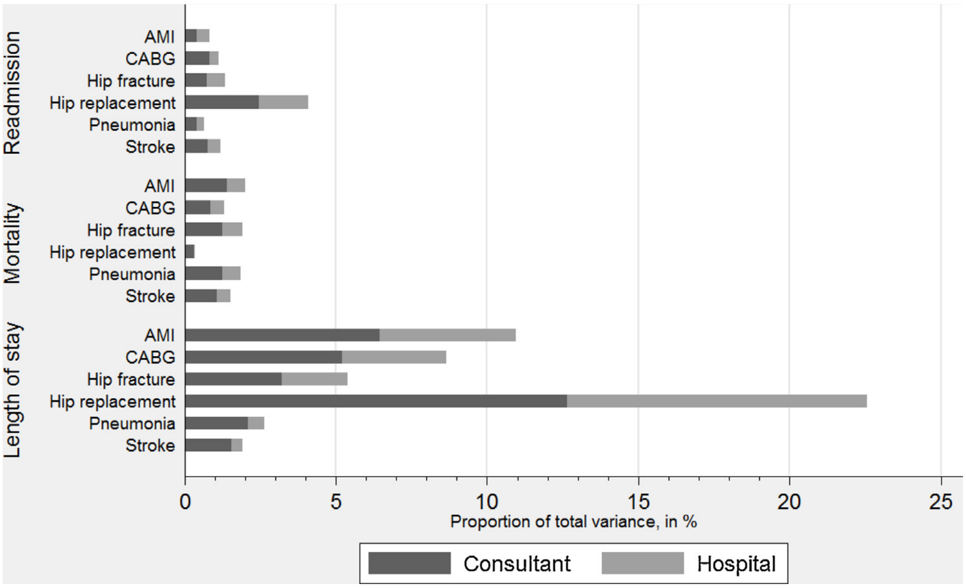

Fig. 1. Proportion of variation attributable to consultants and hospitals; case-mix adjusted.

Commentators have considered the estimated proportion of variability in performance indicators at levels higher than patients (including physicians, groups and organisations) as low or even trivial and have raised concerns about the purpose of performance management [26]. However, we wish to stress that such judgements must consider not only the amount of unwarranted variation but also the value of the performance indicators and the direct and indirect costs of initiatives aimed at eradicating it [27]. For example, assuming an average cost of an emergency readmission in the English NHS of £2,100 [28], we estimate the overall value of improving consultant performance to match that of the current average for our sample alone to be approximately £8.4 million. This ignores any patient health benefits associated with a reduced risk of readmissions. The organisations in which consultants work also play a role in determining outcomes, albeit less than consultants. Hence, the possible benefit of reducing unwarranted variation between consultants and/or organisations is unlikely to be negligible, although this does not necessarily imply that any such effort is a cost-effective use of resources.

As the amount of case-mix adjusted variation between consultants generally exceeds that which occurs between organisations, a focus on individual doctors' performance may be thought justified. In practice, however, there are obstacles to realising the potential benefit of consultant-level performance information. In particular, efforts to identify poorly performing consultants using outcome measures such as readmission and mortality derived from routine data are likely to encounter measurement problems: a large proportion of consultants do not treat a sufficient number

Table 2
Reliability of consultant performance estimates.

Condition	Estimated variance components			VPC	Case-load (median)		Case-load required		% Consultants with sufficient case-load over 35 months	
	σ^2	τ^2	ω^2			R	R=0.7	R=0.9	R=0.7	R=0.9
28-day emergency readmission										
AMI	0.23955	0.00102	0.00095	0.4%	56	0.19	552	2131	0.0%	0.0%
CABG	0.21058	0.00178	0.00065	0.8%	104	0.47	277	1068	0.5%	0.0%
Hip fracture	0.23855	0.00180	0.00144	0.7%	86	0.39	312	1203	0.3%	0.1%
Hip replacement	0.20456	0.00527	0.00348	2.5%	95	0.71	92	355	51.0%	3.8%
Pneumonia	0.24823	0.00103	0.00060	0.4%	83	0.26	563	2171	0.7%	0.0%
Stroke	0.24581	0.00188	0.00107	0.8%	55	0.29	307	1183	8.9%	0.0%
30-day mortality										
AMI	0.143812	0.002077	0.000857	1.4%	56	0.45	163	627	10.0%	0.0%
CABG	0.065208	0.000578	0.000289	0.9%	104	0.48	264	1020	0.5%	0.0%
Hip fracture	0.186773	0.002375	0.001295	1.2%	86	0.52	185	713	2.1%	0.1%
Pneumonia	0.204589	0.002588	0.001311	1.2%	83	0.51	186	716	10.5%	0.4%
Stroke	0.174487	0.001901	0.000806	1.1%	55	0.37	215	830	16.6%	0.2%
Length of stay										
AMI	118.748	8.623	5.987	6.5%	56	0.79	34	130	89.6%	14.5%
CABG	64.090	3.648	2.417	5.2%	104	0.85	43	164	93.4%	19.3%
Hip fracture	2270.457	77.193	51.934	3.2%	86	0.74	70	271	65.6%	0.4%
Hip replacement	30.123	4.921	3.855	12.7%	95	0.93	16	62	100.0%	70.0%
Pneumonia	2354.658	50.766	13.211	2.1%	83	0.64	109	420	35.1%	1.5%
Stroke	28799.816	453.625	105.254	1.5%	55	0.46	149	573	25.0%	1.6%

Notes: R = Reliability; VPC = Variance partition coefficient at consultant level. Median case-load is measured over the period April 2010–February 2013. Variation in mortality after hip replacement at consultant level could not be differentiated from that at hospital level and the corresponding statistics are therefore not recorded.

Table 3
Reliability of hospital performance estimates.

Condition	Estimated variance components			VPC*	Case-load (median)		Case-load required		% Consultants with sufficient case-load over 35 months	
	σ^2	τ^2	ω^2			R	R=0.7	R=0.9	R=0.7	R=0.9
28-day emergency readmission										
AMI	0.23955	0.00102	0.00095	0.8%	787	0.87	284	1095	89.9%	29.1%
CABG	0.21058	0.00178	0.00065	1.1%	734	0.89	202	779	100.0%	46.7%
Hip fracture	0.23855	0.00180	0.00144	1.3%	1000	0.93	172	664	96.6%	79.1%
Hip replacement	0.20456	0.00527	0.00348	4.1%	649	0.97	55	210	97.8%	77.3%
Pneumonia	0.24823	0.00103	0.00060	0.7%	2471	0.94	356	1371	98.7%	86.2%
Stroke	0.24581	0.00188	0.00107	1.2%	946	0.92	194	750	91.7%	63.9%
30-day mortality										
AMI	0.143812	0.002077	0.000857	2.0%	787	0.94	114	441	94.6%	80.4%
CABG	0.065208	0.000578	0.000289	1.3%	734	0.91	176	677	100.0%	56.7%
Hip fracture	0.186773	0.002375	0.001295	1.9%	1000	0.95	119	458	98.6%	87.8%
Pneumonia	0.204589	0.002588	0.001311	1.9%	2471	0.98	122	472	99.3%	97.4%
Stroke	0.174487	0.001901	0.000806	1.5%	946	0.94	150	580	92.4%	77.8%
Length of stay										
AMI	118.748	8.623	5.987	11.0%	787	0.99	19	73	100.0%	95.3%
CABG	64.090	3.648	2.417	8.6%	734	0.99	25	95	100.0%	100.0%
Hip fracture	2270.457	77.193	51.934	5.4%	1000	0.98	41	158	99.3%	96.6%
Hip replacement	30.123	4.921	3.855	22.6%	649	0.99	8	31	100.0%	99.6%
Pneumonia	2354.658	50.766	13.211	2.6%	2471	0.99	86	331	100.0%	98.7%
Stroke	28799.816	453.625	105.254	1.9%	946	0.95	120	464	93.1%	80.6%

Notes: R = Reliability; VPC* = Sum of variance partition coefficients at consultant and hospital levels. Median case-load is measured over the period April 2010–February 2013. Variation in mortality after hip replacement at consultant level could not be differentiated from that at hospital level and the corresponding statistics are therefore not recorded.

of patients over a three-year period for these performance estimates to be reliable representations of their individual underlying performances. There are several ways in which the reliability of individual performance estimates can be improved, although each comes with their own problems. Firstly, most consultants provide a variety of treatments for different patient groups and this can be exploited to generate more comprehensive performance profiles on larger, and thus more reliable, patient samples [29]. This, however, requires a more complex case-mix adjustment strategy and may hide differential performances among the components of the composite for individual consultants [30]. An alternative approach is to employ *shrinkage estimators*, which take into account reliability to generate estimates that are less subject to random variation and regression-to-the-mean [31,32]. This means, however, that resulting estimates of consultant performance are overly conservative and biased towards the average [33]. The implication is that poorly performing consultants with smaller caseloads would be less likely to be identified correctly as negative outliers.

These results suggest that policymakers seeking to manage performance and reduce unwarranted variation pursue the right target but do so by the wrong means. While the variation across consultants *overall* is larger than between hospitals, the performance of an *individual* consultant is difficult to establish reliably. This suggests that performance management approaches seeking to leverage routinely collected data on individual consultants' performances risk generating a non-trivial number of false positive warnings, which may undermine trust in the validity and fairness of the assessment. Until methods to increase the reliability of individual consultants' performance estimates have been agreed and implemented, approaches to performance management may be best aimed at the entire population of consultants (e.g. through enhanced professional regulation) rather than a subset identified by unreliable means.

There are a number of limitations to our study. First, in line with current UK health policy we have chosen consultants (fully trained hospital specialists) as the unit of analysis. However, consultants generally lead teams of healthcare professionals and we cannot observe the actions taken by each individual. It may therefore not be the consultant that had a measurable effect on outcomes; although some may argue that, as leaders of these teams, they remain ultimately responsible. Second, as in all observational studies, our results may be subject to unobserved confounding. Most importantly, length of stay and emergency readmission may be determined by local supply factors, such as the availability of primary care services or care home places. This may explain some of the variation observed across hospitals but is unlikely to explain variation between consultants within the same hospital. Thus, consultant-level variance partition coefficients and the reliability of individual performance estimates may be underestimated. Similarly, performance estimates may be biased by unobserved differences in case-mix. If, for example, more severely ill patients are more likely to seek treatment from providers offering reportedly better services then the estimated variation in performance would be biased downwards. This is clearly of less concern for emergency care where patients have limited ability to choose and so may affect estimates differently across conditions. Third, variation among healthcare providers in dichotomous outcomes (mortality, readmission) may be more difficult to estimate than in continuous outcomes (length of stay) for a given sample size. Since the probability of mortality and readmission, rather than the actual event, can never be observed, this constitutes an inherent limitation of these metrics. Fourth, we have focussed on a number of high-volume procedures and conditions that form part of performance assessment initiatives in England or elsewhere and for which validated performance indicators exist. But these conditions necessarily capture only a subset of all inpatient activity in English hospitals and it is, therefore, unclear in how far our results can be generalised to other patient populations. Finally, while our analysis provides estimates of the degree of variation in patient outcomes and length of inpatient stay that is associated with consultants and hospitals it was not designed to identify the influences and decisions that result in this variation. For example, some of the observed variation at hospital level may be due to differences in infrastructure, which may be difficult to resolve in the short run or is outside the control of the organisation entirely.

5. Conclusions

Policy makers, healthcare regulators and professional bodies in the UK and elsewhere are increasingly targeting both organisations and individual hospital consultants through a variety of performance management schemes and mechanisms. Our study shows that consultants vary in terms of their clinical outcomes and resource utilisation, and that in general the proportion of unexplained variation at consultant level exceeds that at hospital level. However, both consultant and hospital factors explain only a small fraction of the variation in risk-adjusted patient outcomes and process measures (length of stay, mortality and readmissions) compared with unmeasured patient characteristics and random noise, which seems to suggest that the potential impact of these performance management schemes aimed at organisations, individual consultants or both is likely to be relatively limited. In addition, relatively small patient samples per consultant make it difficult to form reliable judgements about consultants' individual performance, and suggest that producing and publishing such comparisons may be at best uninformative and at worst misleading.

Conflict of interest

The authors report no conflicts of interest.

Acknowledgements

This paper is an output from independent research commissioned and funded by the Department of Health Policy Research Programme (PR-R9-0114-11002 Evaluating the development of medical revalidation in England and its impact on organisational performance and medical practice). The views expressed in this publication are those of the author(s) and not necessarily those of the Department of Health. HES data are copyright © 2010–2012, re-used with the permission of NHS Digital. All rights reserved.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.healthpol.2018.04.004.

References

[1] Wennberg J, Gittelsohn A. Small Area Variations in Health Care Delivery: a population-based health information system can guide planning and regulatory decision-making. Science 1973;182(4117):1102–8.
[2] Corallo A, Croxford R, Goodman D, Bryan E, Srivastava D, Stukel T. A systematic review of medical practice variation in OECD countries. Health Policy 2014;114(1):5–14.
[3] Smith PC. Performance measurement in health care: history, challenges and prospects. Public Money & Management 2005;25(4):213–20.
[4] Oliver A. Incentivising improvements in health care delivery. Health Economics Policy and Law 2015;10(3):327–43.
[5] Bridgewater B, et al. Has the publication of cardiac surgery outcome data been associated with changes in practice in northwest England: an analysis of 25,730 patients undergoing CABG surgery under 30 surgeons over eight years. Heart 2007;93(6):744–8.
[6] Hannan EL, et al. The new York state cardiac registries: history, contributions, limitations, and lessons for future efforts to assess and publicly report healthcare outcomes. Journal of the American College of Cardiology 2012;59(25):2309–16.
[7] NHS England. Major breakthrough in NHS Transparency as consultant mortality data goes online for first time; 2013.
[8] HQIP, Clinical Outcomes Publication. 02/10/2017]. Available from: http://www.hqip.org.uk/national-programmes/clinical-outcomes-publication/.
[9] Varagunam M, Hutchings A, Black N. Relationship between patient-reported outcomes of elective surgery and hospital and consultant volume. Medical Care 2015;53(4):310–6.
[10] Snijders TAB, Bosker RJ. Multilevel analysis – an introduction to basic and advanced multilevel modeling, 2 ed. Los Angeles: Sage; 2012.
[11] Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG. The unreliability of individual physician report cards for assessing the costs and quality of care of a chronic disease. JAMA 1999;281(22):2098–105.
[12] Dimick JB, Staiger DO, Birkmeyer JD. Ranking hospitals on surgical mortality: the importance of reliability adjustment. Health Services Research 2010;45(6p1):1614–29.
[13] Adams JL, Mehrotra A, Thomas JW, McGlynn EA. Physician cost profiling — reliability and risk of misclassification New England. Journal of Medicine 2010;362(11):1014–21.
[14] Walker K, et al. Public reporting of surgeon outcomes: low numbers of procedures lead to false complacency. The Lancet 2013;382(9905):1674–7.
[15] Eijkenaar F, van Vliet RCJA. Profiling individual physicians using administrative data from a single insurer: variance components, reliability, and implications for performance improvement efforts. Medical Care 2013;51(8):731–9.
[16] Bernal-Delgado E, et al. ECHO: health-care performance assessment in several European health systems. European Journal of Public Health 2015;25(Suppl. 1):3–7.

[17] Agency for Healthcare Research and Quality. Inpatient quality indicators technical specifications updates – Version v7.0 (ICD 10), September 2017. 2017 13/10/2017; 2017. Available from: http://www.qualityindicators.ahrq.gov/Modules/IQI_TechSpec_ICD10_v70.aspx.

[18] McLennan D, et al. The English Indices of Deprivation 2010. London: Department for Communities and Local Government; 2011.

[19] McCulloch CE. Generalized linear mixed models. Beachwood, Ohio: Institute of Mathematical Statistics; 2003.

[20] Rodriguez G. Multilevel generalized linear models. In: de Leuww J, Meijer E, editors. Handbook of multilevel analysis. New York: Springer; 2008.

[21] Goldstein H, Browne W, Rasbash J. Partitioning variation in multilevel models. Understanding Statistics 2002;1(4):223–31.

[22] Browne WJ, et al. Variance partitioning in multilevel logistic models that exhibit overdispersion. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2005;168(3):599–613.

[23] McKelvey RD, Zavoina W. A statistical model for the analysis of ordinal level dependent variables. The Journal of Mathematical Sociology 1975;4(1):103–20.

[24] Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 2013;4(2):133–42.

[25] Siciliani L, Sivey P, Street A. Differences in length of stay for hip replacement between public hospitals, specialised treatment centres and private providers: selection or efficiency? Health Economics 2013;22(2):234–42.

[26] Fung V, et al. Meaningful variation in performance: a systematic literature review. Medical Care 2010;48(2):140–8.

[27] Meacock R, Kristensen SR, Sutton M. The cost-effectiveness of using financial incentives to improve provider quality: a framework and application. Health Economics 2014;23(1):1–13.

[28] Billings J, et al. Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). BMJ Open 2012;2(4).

[29] Smith KA, et al. Improving the reliability of physician report cards. Medical Care 2013;51(3):266–74.

[30] Gutacker N, et al. Hospital variation in patient-reported outcomes at the level of EQ-5D dimensions: evidence from England. Medical Decision Making 2013;33(6):804–18.

[31] Efron B, Morris C. Stein's estimation rule and its competitors–an empirical bayes approach. Journal of the American Statistical Association 1973;68(341):117–30.

[32] Skrondal A, Rabe-Hesketh S. Prediction in multilevel generalized linear models. Journal of the Royal Statistical Society. Series A (Statistics in Society) 2009;172(3):659–87.

[33] Austin PC, Alter DA, Tu JV. The use of fixed-and random-effects models for classifying hospitals as mortality outliers: a Monte Carlo assessment. Medical Decision Making 2003;23(6):526–39.