Between-hospital and physician variation in outcomes and costs in high- and low-complex surgery: A nationwide multi-level analysis

N. Salet1§, V.A. Stangenberger23§, R. H. Bremmer3, F. Eijkenaar4

Corresponding author:
N. Salet M.D.
salet@eshpm.eur.nl

Author affiliations
1. Erasmus School of Health Policy & Management, Erasmus University, Rotterdam, Zuid-Holland, The Netherlands
2. Amsterdam University Medical Center, University of Amsterdam, Noord-Holland, The Netherlands

Ethical statements
An anonymous database was built from existing reimbursement data that had been amassed by hospitals under the Dutch Healthcare Act (Nederlandse Gezondheidswet). Since this study was based on legally obtained, existing, anonymous data, no additional informed consent was required. All methods were carried out in full accordance with privacy regulations and guidelines.

Author Contributions:
Concept and design: Salet, Stangenberger, Eijkenaar

Acquisition of data: Salet, Stangenberger, Bremmer

Analysis and interpretation of data: Salet, Stangenberger, Bremmer, Eijkenaar

Drafting of the manuscript: Salet, Stangenberger

Critical revision of the paper for important intellectual content: Salet, Stangenberger, Bremmer, Eijkenaar

Statistical analysis: Salet, Stangenberger

Provision of study materials or patients: Salet, Bremmer

Obtaining funding: Eijkenaar

Administrative, technical, or logistic support: Bremmer

Supervision: Eijkenaar

Conflict of interest statement
The authors reported no conflicts of interest.

Funding/Support
None reported

Acknowledgements
The authors gratefully acknowledge the valuable feedback of Erik Schut and Freek Sorgdrager on an earlier draft, as well as the comments of participants of the Erasmus Health Systems and Insurance seminar (October 2021).
Between-hospital and -physician variation in outcomes and costs in high- and low-complex surgery:
A nationwide multi-level analysis

N. Salet\textsuperscript{1§}, V.A. Stangenberger\textsuperscript{23§}, R. H. Bremmer\textsuperscript{3}, F. Eijkenaar\textsuperscript{5}

Corresponding author:
N. Salet
salet@eshpm.eur.nl

Author affiliations
1. Erasmus School of Health Policy & Management, Erasmus University, Rotterdam, Zuid-Holland, The Netherlands
2. Amsterdam University Medical Center, University of Amsterdam, Noord-Holland, The Netherlands

Key words: Performance measurement, hospital, physician, variation, outcome, costs
Abstract

Objectives
Clinicians and policymakers are increasingly exploring strategies to reduce unwarranted variation in outcomes and costs. Adequately accounting for case-mix and better insight into the level(s) at which variation exists is crucial for such strategies. This nationwide study investigates variation in surgical outcomes and costs at the level of hospitals and individual physicians, and evaluates whether these can be reliably compared on performance.

Methods
Variation was analysed using 92,330 patient records collected from 62 Dutch hospitals who underwent surgery for colorectal cancer (n=6,640), urinary bladder cancer (n=14,030), myocardial infarction (n=31,870) or knee osteoarthritis (n=39,790) in the period 2018-2019. Multilevel regression modelling with and without case-mix adjustment was used to partition variation in between-hospital and between-physician components for in-hospital mortality, ICU admission, length of stay, 30-day readmission, 30-day reintervention, and in-hospital costs. Reliability was calculated for each treatment-outcome combination at both levels.

Results
Across outcomes, hospital-level variation relative to total variation ranged between ≤1% and 15%, and given the high caseloads this typically yielded high reliability (>0.9). In contrast, physician-level variation components were typically ≤1%, with limited opportunities to make reliable comparisons. The impact of case-mix adjustment was limited, but nonnegligible.

Conclusions
It is not typically possible to make reliable comparisons between physicians due to limited partitioned variation and low caseloads. For hospitals, however, the opposite often holds. Although variation-reduction efforts directed at hospitals are thus more likely to be successful, this should be approached cautiously, partly because level-specific variation and the impact of case-mix vary considerably across treatments and outcomes.
Highlights

It is often difficult to pinpoint the sources of variation in outcomes and costs. For effective variation reduction strategies, more insight is required into at what level(s) (e.g., hospitals, professionals, patients) variation exists.

Partitioned physician-level variation was typically low and is generally exceeded by hospital-level variation. Additionally, case-mix corrected comparisons on outcomes and costs between individual physicians were unreliable, whereas the opposite often appears to hold for hospitals.

Variation-reduction strategies should be designed and applied with caution and with consideration of the potentially large differences in both level-specific attributed variation and reliability that may exist across treatments and outcomes.
**Introduction**

Clinicians, policymakers and care purchasers are increasingly exploring strategies to identify and mitigate unwarranted variation between healthcare providers in terms of outcomes and costs. In this context, unwarranted variation is defined as variation that may have harmful consequences for patients and may reflect waste. Increasingly, this area has also been the subject of research, with recent examples including studies on mortality rates among COVID-19 patients and readmission and mortality rates in surgical oncology. There are various strategies for mitigating variation, ranging from involving providers in a dialogue about possible causes and solutions and providing structured feedback based on benchmarking results, to more radical strategies such as public reporting and the introduction of value-based payment programmes. Unfortunately, however, the complex nature of healthcare delivery means that it is often difficult to pinpoint potential sources of variation and design effective interventions to reduce (potentially) unwarranted variation. As a result, unwarranted variation often persists despite significant efforts to eliminate it.

To support the design of (more) effective efforts to reduce variation, insight is required into the drivers of variation as well as the level(s) of healthcare delivery (e.g., hospitals, professionals, patients) that the variation can be attributed to. Presently, there are numerous examples of interventions imposed by policymakers following apparent differences in outcomes between providers (e.g., after clinical audits). However, without information on the level to which variation can likely be attributed such interventions are difficult to target and thereby more likely to be ineffective. Identifying level-specific variation is therefore a crucial first step in informing effective policy intervention as well as follow-up research on the drivers of unwarranted variation. Such analyses could help provide insight into how estimated level-specific variation relates to different patient populations and outcomes as well as whether physicians and hospitals can be reliably compared on the outcome(s) in question. Additionally, such insight might assist in preventing misdirected interventions, such as interventions aiming at individual physicians when variation mainly appears to exist at the hospital level.

Generally speaking, there are four possible causes or ‘sources’ of observed variation in patient outcomes and costs. First, there is variation caused by factors specific to provider organisations (e.g., hospitals), which include some that are measured and others that are latent and arise from a common distribution. Similarly, factors that are specific to individual healthcare professionals (e.g., physicians) working in these organisations are a second possible driver of variation. A third possible driver of variation are factors specific to the patient populations served by these providers (i.e., case mix), and which are measured. Insofar these characteristics affect outcomes/costs, this should be accounted for
in between-provider comparisons, for example by adjusting for them using a regression model\textsuperscript{11}. However, whether such case-mix adjustment can be done successfully depends heavily on the availability of data on relevant patient characteristics across all provider entities compared and the quality of those data. The final source of variation covers everything else, including chance variation as well as unobserved patient-level, physician-level, and organization-level characteristics. The latter two factors include differences in clinical or financial performance between physicians and between the provider organizations that they work in\textsuperscript{12}. In order to make meaningful between-provider comparisons that are informative for policy, factors that providers cannot influence should be adequately accounted for, but this is often difficult in practice\textsuperscript{13}. In addition, since variation can exist at different levels, to be successful variation-reduction efforts should be directed at the right level(s)\textsuperscript{14}.

Since the role of chance and case-mix factors (either observed or unobserved) in comparing provider performance is likely to vary across different outcomes and treatments, such comparisons are ideally conducted separately for each treatment-outcome combination. In addition, to avoid a one-sided perspective of (patterns in) estimated provider performance, these comparisons should ideally include an analysis of multiple clinical and financial outcomes. Although between-provider variation has been the focus of plenty of research, previous studies have typically analysed variation for a single treatment-outcome combination, and sometimes without adjusting for chance and/or case-mix\textsuperscript{15}. In addition, many of these studies use data from only one provider or payer, do not include analyses of variation in costs, and do not provide information to help determine whether the variation identified is clinically and/or financially relevant\textsuperscript{15,16,17}. Furthermore, research in this field is often limited to analysing variation at a single level (often the hospital level). As a result, insights on variation at lower levels, specifically individual physicians, remains largely absent\textsuperscript{18}.

Using nationwide patient-level data on nearly every hospital in the Netherlands, this study aims to contribute to addressing these gaps by analysing variation in five clinical outcomes and costs for four high-volume surgical treatments. Using multilevel regression modelling and adjusting for chance variation and case-mix, we partition variation into between-hospital and between-physician components. Using these components, we calculate reliability coefficients (i.e., signal-to-noise ratios) to assess whether hospitals and physicians can be reliably compared on the outcomes analysed. Based on the results and insofar possible given our observational data, we formulate implications for future variation-reduction strategies.
Methods

Data and study population
We used patient-level data that are routinely collected from Dutch hospitals’ information systems. The data were retrieved from a benchmark database belonging to LOGEX b.v. (Amsterdam, Netherlands) which contains data on care activities and other administrative information from nearly all (81%) Dutch hospitals. We analysed variation in outcomes and costs for patients who underwent surgery for colorectal cancer (CRC), urinary bladder cancer (UBC), acute myocardial infarction (AMI), or knee osteoarthritis (KOA) in the period 1st January 2018- 31st December 2019. The corresponding surgical treatments were laparoscopic colonic resection (LAP), transurethral resection (TUR), acute percutaneous coronary intervention (PCI) and total knee arthroplasty (TKA), respectively. We selected these treatments because of their wide range of medical complexity and relatively high caseloads. These factors were expected to contribute to more reliable and clinically relevant estimates. In total, 92,330 patient records were included (Figure 1). Patients were randomly distributed among physicians within a hospital, based on availability. Included physicians were the ones who were responsible for admitting patients (except when patients were transferred from a different department within a hospital), surgery, and discharge.

Outcome variables
Patient outcomes can broadly be divided into three tiers, each representing different treatment stages. Tier 1 outcomes refer to achieved/retained health status, tier 2 outcomes represent time to recovery and treatment disutility, and tier 3 outcomes represent the sustainability of health or iatrogenic effects. Given our data, we aimed to capture each of these dimensions when selecting our outcomes. Specifically, for each of the four treatments, we analysed five (proxy) outcomes due to their importance in relation to healthcare delivery: in-hospital mortality (tier 1), intensive care unit (ICU) admission, length of stay (LoS, tier 2), 30-day readmission (tier 3) and 30-day reintervention (tier 3). We only analysed variation in outcomes with at least 100 events; this was not the case for three treatment-outcome combinations (i.e., 30-day reintervention in CRC patients and in-hospital mortality and 30-day reintervention in KOA patients), which were therefore excluded from the analysis.

Because analysing patient outcomes without considering the financial side will inevitably produce an incomplete picture and cost evaluation is highly relevant in affordable and accessible healthcare, we also analysed variation in total in-hospital costs. All costs (i.e., surgical, diagnostic, clinic, and outpatient) incurred in the hospital in relation to the relevant treatment were included. Standardized unit prices for each care activity were used to determine the total in-hospital cost of treatment, which
was defined as the sum of all care activities relating to treatment with regard to the reference unit price. The same reference unit prices were used for all the hospitals included.

**Case-mix variables**

One of the goals of this study was to investigate the influence of observed differences in case-mix on between-provider variation. In the regression analyses we were able to adjust for the following patient characteristics: age (in years), sex, socio-economic status (based on average income of the neighbourhood where patients live, in three categories from high (SES1) to low (SES3)), and the Elixhauser Comorbidity Score (ECS). The ECS is a score that summarizes disease burden for each patient, rather than solely including a collection of dichotomous variables (i.e. yes/no comorbidity). This score was calculated by summing the points for all the comorbidities that a patient had in the Elixhauser Index. Thus, every patient was given one composite score, reflecting the degree and severity of comorbidity. In addition, (very) low-volume providers were excluded from the analysis because these may distort results. We also tested whether correcting for whether a patient was treated in an academic hospital (academic hospitals generally treat more complex patients than non-academic hospitals) improved model fit according to Akaike’s Information Criterion (AIC). Since this was not the case for any of the treatment-outcome combinations in the study, we decided to leave this correction out.

**Statistical analysis**

Variation in outcomes and costs between hospitals and physicians was analysed using multivariable multilevel logistic (in-hospital mortality, ICU admission (all-cause), 30-day readmission (all-cause), 30-day reintervention (all-cause)) and linear (LoS, costs) regression analysis. Separate models were estimated for each treatment-outcome combination, both without and with adjustment for case-mix. Specifically, for each combination we first fitted an ‘empty’ model. This model only contained random intercepts for hospital and responsible physician, providing an insight into the basic partitioning of variance (i.e., hospital-level, physician-level, and residual) while statistically accounting for random variation by ‘shrinking’ the effects for hospitals and physicians with fewer observations. This model was then supplemented with (the fixed effects of) the case-mix variables described above. For consistency across models and because in a previous study most of these case-mix variables were found to be prognostic factors for most of these outcomes and treatments, we included the same case-mix variables in all adjusted ‘full’ models.
In this study, we were particularly interested in the percentage of variation that could be attributed to the level of physicians and hospitals. Therefore, for each model (i.e., ‘empty’ and ‘full’ for all treatment-outcome combinations), variance partitioning coefficients (VPCs, also known as intraclass correlation coefficients) were calculated by dividing the estimated level-specific variance by the total variance\(^{20}\). These VPCs indicate the proportion of total variation that can be attributed to a specific level. For example, the hospital-level VPC for outcome \(o\) and treatment \(t\) was calculated as follows:

\[
VPC_{h,o,t} = \frac{\psi^2}{\varepsilon^2 + \psi^2 + \phi^2},
\]

where \(\psi^2\) = estimated hospital-level variance, \(\phi^2\) = estimated physician-level variance, and \(\varepsilon^2\) = unobserved residual patient-level variation and measurement error, where total \(VPC_{h,o,t} \leq 1\) (note that from this formula it follows that for a specific combination of treatment and outcome, the hospital- and physician-level VPCs do not sum to one). The VPC estimates implicitly correct for statistical uncertainty (or random variation) due to low caseloads of physicians and hospitals by the ‘shrinking’ in random-effect modelling. Specifically, estimated provider effects on outcomes based on relatively small caseloads are ‘pulled’ more towards the mean than estimated provider effects based on larger caseloads. We also analysed the impact of case-mix adjustment by tracking the changes in AIC-values and VPCs (as well as calculating reliability coefficients, see below), and we compared VPCs across outcomes and across treatments.

In general, reliability (also referred to as rankability, signal-to-noise ratio, or statistical uncertainty in this context\(^{31}\)) reflects the reproducibility or consistency of a measure across repeated measurements\(^{32}\). In this study, reliability is a function of the (adjusted) estimated VPC for a specific level and the number of observations at that level: the higher the VPC and caseload \(N\) (i.e., caseload per hospital/physician), the higher the reliability. More specifically, reliability is calculated as \(R = \frac{N \times \text{VPC}}{1 + [N-1] \times \text{VPC}}\) with \(0 \leq R \leq 1\)\(^{33,34}\) (see also Technical Appendix). An R-value close to 1 at a certain level suggests little statistical uncertainty and thus that the variation is likely to reflect ‘true’ variation at that level rather than variation due to chance. Consistent with previous work\(^{33}\), R-values of \(\geq 0.9\) were interpreted as excellent, 0.89-0.80 as good, 0.79-0.70 as moderate, and <0.7 as low (implying limited usefulness in practice). In practice and regarding potential for improvement, excellent reliability in combination with significant and unbiased estimates of level-specific variation is desirable to justify external, ‘high stakes’ variation-reduction strategies such as public reporting or application in value-based payment models. Depending on the situation, reliability coefficients between 0.7 and 0.9 might be deemed enough to warrant ‘low stakes’ strategies such as dialogue and feedback\(^{35,36,37}\).

Using the formula for \(R\) shown above, for each treatment-outcome combination and given the estimated VPCs, we also calculated the caseload required to reach a reliability of 0.7 and 0.9. We then
compared these caseload requirements with the actual caseloads as observed in the data. Finally, we created caterpillar plots to illustrate the relationship between VPCs and R-values across all outcomes and treatments. These plots rank providers from low to high estimated performance, with 95% confidence intervals (CI). All analyses were conducted in R, version-4.0.2.

Results

Descriptive statistics
In total, 92,330 records of patients treated in 62 Dutch hospitals in the period 2018-2019 were included (Figure 1). All these patients received at least one of the four selected surgical interventions: LAP-CRC (n= 6,640), TUR-UBC (n= 14,030), PCI-AMI (n= 31,870), and TKA-KOA (n= 39,790). The mean age of the cohort was 68.2 years and 55.9% were male. Patients with CRC and UBC were more likely to suffer from more severe comorbidity than AMI and KOA patients, translating into higher ECSs: 5.1 and 5.9 for CRC and UBC patients versus 1.3 for AMI and KOA patients (Table 1).

In-hospital mortality was highest in patients who underwent PCI (2.4%) followed by patients who underwent LAP-CRC (1.6%), TUR-UBC (1.1%) and TKA-KOA (0.1%) (Table 2). ICU admission rates were lowest among TKA-KOA patients (1.0%) and highest in LAP-CRC patients (10.2%). The median LoS after surgery was highest in LAP-CRC patients (4 days). By contrast, readmission (11.6%) and reintervention (3.8%) rates were highest in TUR-UBC patients. LAP-CRC was the most expensive treatment with a median total cost of €14,404, followed by TKA-KOA (€10,056), TUR-UBC (€8,241) and PCI-AMI (€5,058) (Table 2).

Unadjusted between-provider variation
Substantial unadjusted variation in terms of interquartile ranges (IQR) existed at both hospital and physician levels, although there were large differences across treatments and outcomes (Table 2). For example, variation in absolute terms was generally low in KOA patients (except perhaps for 30-day readmission). Overall, unadjusted variation was largest in ICU admission and 30-day readmission, at both levels (e.g., 6% and 7% for ICU admission for AMI patients). For costs, too, considerable absolute variation at both levels can be observed (e.g., for LAP-CRC the IQR is €2,660 at the hospital and €3,049 at the physician level, respectively).
Impact of case-mix adjustment

Although adjustment for observed differences in case-mix generally improved the model fit (especially for ICU admission, LoS and costs) based on a comparison of AIC-values for the ‘empty’ models (adjusted only for random variation and volume) to those of the ‘full’ models (adjusted also for case-mix) (Appendix Table A1), the impact on the estimated VPCs was limited overall. Changes in VPCs ranged from -0.01 to +0.01 (Appendix Table A2). Consequently, case-mix adjustment left the patterns in VPCs across outcomes, treatments, and levels (i.e., hospital-level variation relative to physician-level variation) largely unaffected. Not surprisingly, the impact on reliability estimates was also limited, with some exceptions (Appendix Table A2). The mean change in reliability when moving from the empty to the full model was +0.03 for hospital-level variation (range -0.02 to +0.58, the latter being a clear outlier) and +0.01 for physician-level variation (range -0.06 to +0.08). Below, we will discuss the results from the ‘full’ case-mix adjusted models.

Adjusted VPCs by outcome

There were considerable differences across outcomes and treatments in the share of total variance that could be partitioned to the two levels (Table 3). Relative to residual patient-level variation and measurement error, low amounts of variation could be attributed to either level, but especially to the physician level. Estimated hospital-level VPCs were generally ≤0.15, with some exceptions. At the physician level, however, most VPCs were estimated at ≤0.01. Hospital-level VPCs often (but not always) exceed physician-level VPCs; across outcomes, the mean hospital-level VPC exceeds the mean physician-level VPC by a factor of 1.5 for LAP-CRC (0.03 vs. 0.02), 7 for TUR-UBC (0.07 vs. 0.01), 11 for PCI-AMI (0.11 vs. 0.01) and 16 for TKA-KOA (0.16 vs. 0.01). The following sections present these results in more detail, in which ‘VPC_h’ represents the proportion of variance partitioned to the hospital level and ‘VPC_p’ the proportion of variance partitioned to the physician level.

In-hospital mortality

For both LAP-CRC and TUR-UBC patients, the estimated VPC for in-hospital mortality was higher at the physician level (0.05 and 0.06, respectively) than at the hospital level (<0.01 and 0.01, respectively). This was not the case in PCI-AMI patients, among whom nearly all the variance partitioned (VPC_p=0.01 and VPC_h=0.29) existed at the hospital level. This outcome was not analysed for TKA-KOA patients because the number of events was too low.
ICU admission
In the analysed data, very little variation (i.e., 0.01 or less) could be attributed to the physician level for this outcome, for all treatments. Hospital-level VPCs ranged from 0.09 for LAP-CRC to 0.15 for TKA-KOA, by contrast.

LoS
Similarly, regardless of the treatment, no more than 1% of total variation (maximum VPC= 0.01) in LoS could be partitioned to the physician level. Apart from TKA-KOA (VPC_=0.18), estimated hospital-level VPCs did not exceed 0.02.

30-day readmission
Again, the proportion of variation that could be partitioned to the physician level was low for each of the four treatments. Specifically, 1% of variation in 30-day readmission rates for LAP-CRC could be attributed to either of the two levels. In TUR-UBC patients, this was 3%, which existed almost exclusively at the hospital level. For PCI-AMI patients, the estimated VPC was 0.01 for physician level and 0.06 for hospital level. For TKA-KOA, these figures were 0.02 and 0.03, respectively.

30-day reintervention
Nearly all variation in 30-day reintervention rates was attributed to the hospital level, with estimated physician-level VPCs again not exceeding 0.01. For TUR-UBC and PCI-AMI patients, hospital-level VPCs were 0.11 and 0.09, respectively. The estimates for LAP-CRC and TKA-KOA patients were excluded because there were fewer than 100 events.

Total in-hospital costs
For costs, the amount of variation that could be partitioned to the hospital level was generally higher than for health-related outcomes. In line with the previous outcomes, however, physician-level VPCs were low, typically <0.01. The fraction of total variation attributed to the hospital level was 0.05, 0.10, 0.13 and 0.36 for LAP-CRC, TUR-UBC, PCI-AMI and TKA-KOA patients, respectively.

Provider rankings, reliability, and volume requirements
Estimated reliability coefficients for hospital-level variation often exceeded 0.90, although there were (large) differences across treatments and outcomes. By contrast, only two physician-level reliability estimates exceeded 0.70 (with none reaching 0.90). At the physician level, reliability ranged between 0.50 and 0.70 for nearly half the treatment-outcome combinations analysed.
The relationship between VPCs and reliability coefficients can be illustrated by ranking providers according to their estimated performance. These rankings can be clustered into four broad categories: 1) high VPC but low reliability due to low patient numbers; 2) low VPC but high reliability due to high patient numbers; 3) low VPC and low reliability; and 4) high VPC and high reliability. Outcomes in the fourth category are likely to be the most informative for practice. To illustrate this relationship and the extent to which providers can reliably be compared on the outcomes analysed, we created caterpillar plots using the estimated hospital- and physician-specific effects, ranked from poor to good (Figure 2). In Figure 2, the physician-level rankings in panels A and C reflect reliability estimates that do not exceed 0.7 (R=0.67 and 0.60, respectively), with physician effect estimates that cannot be accurately distinguished from each other (all the confidence intervals overlap). In contrast, the hospital-level rankings in panels B and D are both characterized by high estimated reliability coefficients (i.e., R=0.97 and 0.96), but with relatively high and low estimated VPCs (i.e., 0.13 and 0.02), respectively.

Overall, Figure 2 shows that in the case of small patient numbers, reliable between-provider comparison is not generally possible, regardless of VPC. This is particularly true for individual physicians (see also Table 1). To illustrate this further, Appendix Table A3 shows physician-level caseloads that, given the estimated VPCs, would need to reach a reliability of 0.70 and 0.90. Comparison with the actual caseloads (over a two-year period) at this level reveals that it is generally difficult to reliably distinguish physicians on these outcomes and treatments. This is not only due to the small caseloads, but also because of the low estimated VPCs at this level. This is different at the hospital level, where caseload requirements are typically met.

Discussion

Main findings and implications

In this study, we analysed variation in five clinical outcomes and costs for four high-volume surgical treatments, at the hospital and physician level. Four key findings stand out from our analysis. First, although variation that could be attributed to either level was often substantial in absolute terms, this proportion was generally limited relative to residual variation at the patient level, which typically comprised 85% or more of total variation. This finding is consistent with previous work.1,16,38 Although case-mix adjustment reduced residual (patient-level) variation, variation remained largest at this level which means that it could not be explained by observed physician, hospital, or case-mix variables. It is possible that variation at this level would be reduced further if other potentially relevant case-mix
variables (which ideally include outcome-specific prognostic factors) would be added to the models, although based on prior work we believe variation is likely to remain largest at this level.

Second, variation partitioned to the level of individual physicians was typically low (except perhaps for in-hospital mortality in LAP-CRC and TUR-UBC patients). Although this suggests limited between-physician variation, this might also reflect the difficulty to characterize physician effects due to limited available information at this level (i.e., low caseloads), resulting in low VPC and reliability estimates. The finding that physician-level variation appears relatively high for in-hospital mortality in LAP-CRC and TUR-UBC patients might related to the technical complexity or duration of these surgical treatments (more complex procedures might show larger variation in outcome and vice versa\(^{39}\)). Additionally, the relatively high VPC estimates for in-hospital mortality for elective cancer operations on the physician level relative to the hospital level might be explained by the difference between elective and emergency care, which in future work could (partly) be accounted for by adding the exact time of procedure to the models.

Although meaningful variation might exist at the physician level (even though VPCs were typically low), low caseloads generally make between-physician comparisons highly unreliable, despite the use of nationwide data on high-volume treatments over two years. Possible options for increasing the effective caseload might be to use composite outcome measures (which are useful particularly if scores on the constituent outcomes are strongly correlated\(^{40,41}\)) or to include data over longer periods. The downside of both these approaches, however, is that they would reduce the actionability of the results because these would be on a higher level of aggregation and/or less likely to represent the current population and treatment standards. Additionally, when physician-level variation in outcome is low, there seems to be little room for improving quality through eliminating variation at that level regardless of caseload considerations.

All in all, notwithstanding the limitations of our data (see below), variation-reduction strategies aimed at individual physicians do not seem justified, at least not for the outcomes and treatments analysed here.

Third, variation partitioned to the hospital-level typically exceeded physician-level variation. Combined with the inherently higher caseloads at the hospital level, this often seems to allow for reliable comparison between hospitals, for instance in terms of distinguishing between hospitals with a high, average, or low ranking. Variation partitioned to this level was particularly large for several outcome-treatment combinations (which could therefore be appropriate targets for further analysis), including ICU admission, 30-day reintervention, and costs for TUR-UBC patients; in-hospital mortality
and ICU admission for PCI-AMI patients; and ICU admission, LoS and costs in TKA-KOA patients. In total, estimated reliability coefficients exceeded 0.90 for 14 treatment-outcome-treatment combinations (most in PCI-AMI and TKA-KOA patients, and to a lesser extent in TUR-UBC and LAP-CRC patients). However, even in these cases caution is advised when designing and applying variation-reduction strategies. One reason is that, as also found in our study for several treatment-outcome combinations, high caseloads can yield high reliability even when estimated level-specific variation is limited relative to total variation (i.e., small between-provider differences can be accurately identified). It is therefore important to consider between-provider variation both in relative (i.e., in terms of VPCs) and absolute terms (e.g., a low VPC may still be meaningful if absolute variation is high overall). Particularly for ICU admissions and in-hospital costs, estimated hospital-level VPCs and reliability coefficients were relatively high for all four treatments, suggesting that these may be particularly suitable targets for further analysis to inform variation-reduction efforts. Another reason to be cautious is that although a high reliability implies that it seems possible to reliably distinguish poor-performing providers from high-performing providers (and from the average), it will not necessarily be possible to distinguish these outlier providers from providers with slightly higher or lower scores. This is also illustrated by panel B in Figure 2, which shows overlapping confidence intervals for approximately 80% of the hospitals despite a reliability coefficient of 0.97.

Finally, our results show that there are large differences in estimated provider-level variation across treatment-outcome combinations, as well as across different outcomes for the same treatment and across treatments for the same outcome. This underlines the importance of analysing variation separately for each relevant treatment-outcome combination. In addition, although the impact of case-mix adjustment on the estimated VPCs was limited overall, it was nonnegligible and quite substantial in some cases. Hence, we believe case-mix adjustment should be routinely applied in between-provider comparisons on outcome.

Comparison with previous research
When we compare our results with those of a recent nationwide observational study on multilevel provider variation in outcomes in the context of the English National Health Service, some similarities and differences are worth discussing. Consistent with our findings, the NHS study concluded that it was often impossible to reliably distinguish individual physicians on outcome. Most variation was attributed to unobserved factors, with estimated physician- and hospital-level variance components mainly ranging between <0.01 and 0.11, which is broadly similar to what we have found.
However, contrary to our results and those of prior research into practice variation, in the NHS study physician-level variation generally exceeded hospital-level variation, including for treatment-outcome combinations that were also analysed in our study (i.e., mortality, LoS and readmission in AMI-patients). Possible explanations for this include the analysis of different outcomes and different treatments (although there was some overlap), the use of older data (i.e., physician-level variation might have declined over time) and/or international variation in physician performance (e.g., due to different clinical experience and/or standards of care).

In a literature review published in 2010 that included 39 studies on multilevel variation in quality and outcomes of care, the overall proportion of variation that could be attributed to the hospital or physician level was found to be low; combined with low caseloads this resulted in low reliability coefficients and thus a limited ability to detect meaningful variation in performance. In contrast, in our analysis volume-requirements for reaching high reliability at the hospital-level were often met, which, in addition to the differing study contexts, may be related to our analysis being limited to high-volume treatments and the use of nationwide data.

Overall, the differences in results across studies conducted in different settings as well as across treatments and outcomes within settings underlines the limited generalizability of findings on between-provider variation in outcomes and costs, and thereby the importance of tailored variation-reduction efforts that are based on context-specific analyses of variation.

Literature on (unwarranted) variation in healthcare delivery dates back half a century. Gradually, as research methods matured, between-provider variation was identified at different levels, albeit with often low reliably due to small caseloads. Literature on variation in process and outcome measures of quality of care was last summarized in 2010. Because the demand for transparency in quality and costs has increased substantially and given developments in data collection and computing power over the past decade, a new systematic review of studies examining multi-level variation may provide important new insights. In addition, as also underscored by the limitations described below, future research should focus on methods to disentangle warranted from unwarranted variation, addressing unobserved confounding to enable causal interpretations of findings, as well as on providing insight into how much variation (both in absolute terms and relative to other levels) would be enough to warrant intervention and how much such intervention would reduce disease burden and/or costs.
Strengths and limitations
An important strength of our study is the use of nationwide data on high-volume surgical treatments. The use of multilevel regression modelling allowed us to gain insight in the partitioning of variation not only at the level of hospitals, but also at the level of physicians. In addition, we analysed variation in multiple diverse and clinically relevant outcome measures as well as costs, while adjusting for observed differences in case-mix among providers.

Several limitations should also be mentioned. First and foremost, unobserved confounding especially at the patient level may have introduced biased estimates of variance components, precluding causal interpretations of our findings. Unfortunately, our data did not allow for the application of methods to formally address such selection bias (e.g., instrumental variable analysis). The same holds for adjusting for other potentially relevant (outcome-specific) prognostic factors. We believe these to be important topics for follow-up research.

Second, in this study we focused on variation ‘in general’ as a fundamental first step. That is, we could not explicitly distinguish between warranted and unwarranted variation, which is naturally important for improving care in practice. It is likely that not all between-provider variation in outcome is unwarranted. For example, healthcare professionals may have valid reasons to opt for longer length of stay, for example in cases of clear expected patient benefit. In this respect, more in-depth (mixed methods) research into the exact sources of physician-level variation (e.g., difficulty/complexity of surgical procedure, professional uncertainty, practice style, teamwork, and/or strategic behaviour due to financial incentives) is important to design effective strategies and further bolster the actionability of benchmarking results. Relatedly, as variation might be linked to the specific treatment rather than the clinical condition, it would be interesting for future research to compare treatments with the same chirurgical intervention (e.g., laparoscopic resection) for different clinical conditions (e.g., appendicitis, cholecystitis).

A third and related limitation is that except for the distinction between general and academic hospitals, no information was available on specific hospitals characteristics. In the Dutch institutional context, general/academic hospitals are quite homogeneous because all hospitals must be non-profit entities by law, typically offer a broad range of hospital services, and are almost all located in urbanized areas (due to the high population density) and serve the general public. Nevertheless, specific (institutional) characteristics of hospitals might impact outcome variation between hospitals. Fourth, the fact that 16 hospitals were unavailable in our data might have introduced some selection bias. However, as these hospitals are located across the country and have similar accessibility and
volumes (based on revenue) compared to the hospitals we do have data on, we believe the risk of selection bias to be low.

Fifth, we cannot fully preclude the possibility that physicians incidentally perform procedures at multiple hospitals. Although this too might have introduced bias, the impact on our results is expected to be low because the number of physicians for whom this is the case is likely to be small.

Finally, our conclusions only directly apply to the specific treatments and outcomes analysed in the Dutch hospital sector. For rare but potentially devastating outcomes (e.g., mortality in TKA-KOA) it is for example statistically close to impossible to reliably compare providers in an informative manner, even though care might be suboptimal.

**Conclusion**

Across the outcomes and surgical treatments analysed, it was not typically possible to make reliable comparisons between individual physicians due to the limited share of variation attributed to the physician level and low caseloads. On the other hand, it often did seem possible to reliably distinguish hospitals on outcome and costs due to the larger partitioned variation and larger numbers of patients. Nevertheless, even though variation-reduction strategies are therefore expected to yield more meaningful results when aimed at hospitals rather than individual physicians, such strategies should still be designed and applied with caution, with careful consideration of the limitations of the data used and the potentially large differences in variation and reliability across treatments and outcomes.
Acknowledgements
The authors gratefully acknowledge the valuable feedback of Erik Schut and Freek Sorgdrager on earlier drafts, as well as the comments of participants of the Erasmus Health Systems and Insurance seminar (October 2021).

Data availability
This study brought together existing data obtained upon request and subject to license restrictions from several different sources. The database is not publicly available due to the (commercially, politically, ethically) sensitive nature of the data. No source consented to their data being retained or shared. Permission was acquired from a third party for use of the data in this study and following the publication of this paper.

Ethical statements
An anonymous database was built from existing reimbursement data that had been amassed by hospitals under the Dutch Healthcare Act (Nederlandse Gezondheidswet). Since this study was based on legally obtained, existing, anonymous data, no additional informed consent was required. All methods were carried out in full accordance with privacy regulations and guidelines.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions
NS and VS designed the study, drafted the manuscript, and had a leading role in all other aspects of the study. FE provided the original idea of the study and FE and RB contributed to shaping the analysis. FE and RB performed critical revision of the manuscript. All authors read and approved the final manuscript.

Conflict of interest statement
The authors declare no competing interests, financial or otherwise.
References


45. Jong, J. D. de (Judith D. Explaining medical practice variation: social organization and
institutional mechanisms = Het verklaren van variatie in medisch handelen: sociale organisatie en institutionele mechanismen. NIVEL (2008).


Figure 1. Flowchart of study population, selection procedure and exclusion criteria. The difference between the total amount of patient records and unique patients was caused by 401 patients who had 2 or more of the included procedures.
<table>
<thead>
<tr>
<th>Number of hospitals and patient records</th>
<th>Laparoscopic resection of colorectal carcinoma (LAP-CRC)</th>
<th>Transurethral resection of urinary bladder carcinoma (TUR-UBC)</th>
<th>Acute percutaneous coronary intervention (PCI-AMI)</th>
<th>Total knee arthroplasty for osteoarthritis (TKA-OA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hospitals, n (%)</td>
<td>49 (79.0)</td>
<td>62 (100)</td>
<td>24 (38.7)</td>
<td>62 (100)</td>
</tr>
<tr>
<td>Number of physicians</td>
<td>181</td>
<td>310</td>
<td>202</td>
<td>531</td>
</tr>
<tr>
<td>Patient volume, n</td>
<td>6,617</td>
<td>14,017</td>
<td>31,864</td>
<td>39,947</td>
</tr>
<tr>
<td>Patient volume per hospital, mean</td>
<td>138</td>
<td>226</td>
<td>1.328</td>
<td>642</td>
</tr>
<tr>
<td>Patient volume per physician, mean</td>
<td>37</td>
<td>45</td>
<td>158</td>
<td>75</td>
</tr>
<tr>
<td>Year 2018, n (%)</td>
<td>3,274 (49.5)</td>
<td>7,618 (54.3)</td>
<td>15,698 (49.3)</td>
<td>19,695 (49.3)</td>
</tr>
<tr>
<td>Year 2019, n (%)</td>
<td>3,343 (50.5)</td>
<td>6,399 (45.7)</td>
<td>16,166 (50.7)</td>
<td>20,252 (50.7)</td>
</tr>
<tr>
<td>Age, mean (SD)</td>
<td>67.1 (13.5)</td>
<td>71.1 (10.9)</td>
<td>65.7 (11.8)</td>
<td>69.3 (9.1)</td>
</tr>
<tr>
<td>Male sex, n (%)</td>
<td>3,276 (49.5)</td>
<td>10,741 (76.6)</td>
<td>22,977 (72.1)</td>
<td>14,669 (36.7)</td>
</tr>
<tr>
<td>Socio-economic status 1 (high), n (%)</td>
<td>2,244 (33.9)</td>
<td>4,321 (30.8)</td>
<td>9,296 (29.2)</td>
<td>11,913 (29.8)</td>
</tr>
<tr>
<td>Socio-economic status 2 (medium), n (%)</td>
<td>2,432 (36.8)</td>
<td>4,977 (35.5)</td>
<td>11,086 (34.8)</td>
<td>15,037 (37.6)</td>
</tr>
<tr>
<td>Socio-economic status 3 (low), n (%)</td>
<td>1,941 (29.3)</td>
<td>4,719 (33.7)</td>
<td>11,482 (36.0)</td>
<td>12,997 (32.5)</td>
</tr>
<tr>
<td>Elixhauser 0, n (%)</td>
<td>722 (10.9)</td>
<td>487 (3.5)</td>
<td>23,614 (74.1)</td>
<td>27,541 (68.9)</td>
</tr>
<tr>
<td>Elixhauser 1, n (%)</td>
<td>3,600 (54.4)</td>
<td>8,089 (57.7)</td>
<td>5,088 (16.0)</td>
<td>8,758 (21.9)</td>
</tr>
<tr>
<td>Elixhauser 2, n (%)</td>
<td>1,352 (20.4)</td>
<td>3,403 (24.3)</td>
<td>1,776 (5.6)</td>
<td>2,599 (6.5)</td>
</tr>
<tr>
<td>Elixhauser 3, n (%)</td>
<td>511 (7.7)</td>
<td>1,343 (9.6)</td>
<td>707 (2.2)</td>
<td>743 (1.9)</td>
</tr>
<tr>
<td>Elixhauser 4, n (%)</td>
<td>265 (4.0)</td>
<td>468 (3.3)</td>
<td>246 (0.8)</td>
<td>202 (0.5)</td>
</tr>
<tr>
<td>Elixhauser 5, n (%)</td>
<td>96 (1.5)</td>
<td>138 (1.0)</td>
<td>289 (0.9)</td>
<td>67 (0.2)</td>
</tr>
<tr>
<td>Elixhauser &gt;5, n (%)</td>
<td>71 (1.1)</td>
<td>89 (0.6)</td>
<td>144 (0.5)</td>
<td>37 (0.1)</td>
</tr>
<tr>
<td>Elixhauser comorbidity score, mean (SD)</td>
<td>1.5 (1.1)</td>
<td>1.6 (1.0)</td>
<td>0.4 (0.9)</td>
<td>0.4 (0.8)</td>
</tr>
</tbody>
</table>

Table 1. Descriptive statistics of study population, by surgical treatment. SD = standard deviation.
<table>
<thead>
<tr>
<th>Outcome and costs summary statistics and unadjusted variation (interquartile ranges, IQR) at hospital and physician level, by surgical treatment</th>
<th>Laparoscopic resection of colorectal carcinoma (LAP-CRC)</th>
<th>Transurethral resection of urinary bladder carcinoma (TUR-UBC)</th>
<th>Acute percutaneous coronary intervention (PCI-AMI)</th>
<th>Total knee arthroplasty for osteoarthritis (TKA-KOA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-hospital mortality rate, overall (%)</td>
<td>1.6</td>
<td>1.1</td>
<td>2.4</td>
<td>*</td>
</tr>
<tr>
<td>ICU admission rate, overall (%)</td>
<td>10.2</td>
<td>4.3</td>
<td>6.4</td>
<td>1.0</td>
</tr>
<tr>
<td>Length of stay in days, overall median [IQR]</td>
<td>4.0 [4]</td>
<td>1.0 [0]</td>
<td>1.0 [3]</td>
<td>2.0 [2]</td>
</tr>
<tr>
<td>Length of stay in days, hospital level median [IQR] (%)</td>
<td>6.0 [2]</td>
<td>1.0 [0]</td>
<td>2.0 [1]</td>
<td>2.0 [1]</td>
</tr>
<tr>
<td>30-day readmission rate, overall (%)</td>
<td>6.5</td>
<td>11.6</td>
<td>7.1</td>
<td>2.5</td>
</tr>
<tr>
<td>30-day reintervention rate, overall (%)</td>
<td>*</td>
<td>3.8</td>
<td>1.6</td>
<td>*</td>
</tr>
<tr>
<td>30-day reintervention rate, hospital level median [IQR] (%)</td>
<td>*</td>
<td>4.0 [4]</td>
<td>1.0 [1]</td>
<td>*</td>
</tr>
<tr>
<td>30-day reintervention rate, physician level median [IQR] (%)</td>
<td>*</td>
<td>4.0 [5]</td>
<td>1.0 [2]</td>
<td>*</td>
</tr>
</tbody>
</table>

Table 2. Outcome and costs summary statistics and unadjusted variation (interquartile ranges, IQR) at hospital and physician level, by surgical treatment * = indicator was excluded because of too little events (<100).
<table>
<thead>
<tr>
<th>Treatment</th>
<th>Outcome</th>
<th>VPC hospital</th>
<th>R hospital</th>
<th>VPC physician</th>
<th>R physician</th>
<th>Caseload required for R=0.70 (hospital)</th>
<th>Caseload required for R=0.90 (hospital)</th>
<th>Caseload required for R=0.70 (physician)</th>
<th>Caseload required for R=0.90 (physician)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laparoscopic resection of colorectal carcinoma (LAP-CRC)</td>
<td>Mortality</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>0.05</td>
<td>0.67</td>
<td>&gt;100,000</td>
<td>&gt;100,000</td>
<td>43</td>
<td>165</td>
</tr>
<tr>
<td></td>
<td>ICU admission</td>
<td>0.09</td>
<td>0.93</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>23</td>
<td>90</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Length of stay</td>
<td>0.02</td>
<td>0.68</td>
<td>0.00</td>
<td>0.11</td>
<td>149</td>
<td>575</td>
<td>685</td>
<td>2643</td>
</tr>
<tr>
<td></td>
<td>Readmission</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>0.01</td>
<td>0.28</td>
<td>NA</td>
<td>NA</td>
<td>216</td>
<td>833</td>
</tr>
<tr>
<td></td>
<td>Reintervention*</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Total costs</td>
<td>0.04</td>
<td>0.86</td>
<td>0.01</td>
<td>0.28</td>
<td>51</td>
<td>198</td>
<td>220</td>
<td>847</td>
</tr>
<tr>
<td>Transurethral resection of urinary bladder carcinoma (TUR-UBC)</td>
<td>Mortality</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>0.06</td>
<td>0.73</td>
<td>&gt;100,000</td>
<td>&gt;100,000</td>
<td>38</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td>ICU admission</td>
<td>0.12</td>
<td>0.97</td>
<td>&lt;0.01</td>
<td>0.08</td>
<td>18</td>
<td>68</td>
<td>1212</td>
<td>4675</td>
</tr>
<tr>
<td></td>
<td>Length of stay</td>
<td>0.01</td>
<td>0.74</td>
<td>0.01</td>
<td>0.25</td>
<td>190</td>
<td>733</td>
<td>313</td>
<td>1207</td>
</tr>
<tr>
<td></td>
<td>Readmission</td>
<td>0.03</td>
<td>0.88</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>75</td>
<td>288</td>
<td>&gt;100,000</td>
<td>&gt;100,000</td>
</tr>
<tr>
<td></td>
<td>Reintervention</td>
<td>0.11</td>
<td>0.97</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>19</td>
<td>73</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Total costs</td>
<td>0.13</td>
<td>0.97</td>
<td>&lt;0.01</td>
<td>0.08</td>
<td>15</td>
<td>60</td>
<td>1280</td>
<td>4939</td>
</tr>
<tr>
<td>Acute percutaneous coronary intervention (PCI-AMI)</td>
<td>Mortality</td>
<td>0.29</td>
<td>1.00</td>
<td>0.01</td>
<td>0.66</td>
<td>6</td>
<td>22</td>
<td>189</td>
<td>729</td>
</tr>
<tr>
<td></td>
<td>ICU admission</td>
<td>0.13</td>
<td>0.99</td>
<td>0.01</td>
<td>0.51</td>
<td>16</td>
<td>63</td>
<td>361</td>
<td>1391</td>
</tr>
<tr>
<td></td>
<td>Length of stay</td>
<td>0.02</td>
<td>0.97</td>
<td>0.00</td>
<td>0.19</td>
<td>92</td>
<td>357</td>
<td>1532</td>
<td>5909</td>
</tr>
<tr>
<td></td>
<td>Readmission</td>
<td>0.06</td>
<td>0.99</td>
<td>0.01</td>
<td>0.54</td>
<td>36</td>
<td>138</td>
<td>315</td>
<td>1213</td>
</tr>
<tr>
<td></td>
<td>Reintervention</td>
<td>0.09</td>
<td>0.99</td>
<td>0.01</td>
<td>0.70</td>
<td>24</td>
<td>93</td>
<td>156</td>
<td>603</td>
</tr>
<tr>
<td></td>
<td>Total costs</td>
<td>0.09</td>
<td>0.99</td>
<td>0.01</td>
<td>0.48</td>
<td>24</td>
<td>92</td>
<td>395</td>
<td>1524</td>
</tr>
<tr>
<td>Total knee arthroplasty for osteoarthritis (TKA-OA)</td>
<td>Mortality*</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ICU admission</td>
<td>0.15</td>
<td>0.99</td>
<td>0.01</td>
<td>0.41</td>
<td>13</td>
<td>50</td>
<td>242</td>
<td>932</td>
</tr>
<tr>
<td></td>
<td>Length of stay</td>
<td>0.18</td>
<td>0.99</td>
<td>0.01</td>
<td>0.43</td>
<td>10</td>
<td>40</td>
<td>236</td>
<td>910</td>
</tr>
<tr>
<td></td>
<td>Readmission</td>
<td>0.03</td>
<td>0.95</td>
<td>0.02</td>
<td>0.60</td>
<td>86</td>
<td>333</td>
<td>116</td>
<td>448</td>
</tr>
<tr>
<td></td>
<td>Reintervention*</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Total costs</td>
<td>0.35</td>
<td>1.00</td>
<td>0.01</td>
<td>0.30</td>
<td>4</td>
<td>17</td>
<td>404</td>
<td>1558</td>
</tr>
</tbody>
</table>

Table 3. Variance partition coefficients (VPC), corresponding reliability (R), and minimal caseload required to reach R=0.7/R=0.9 per outcome. NA = required caseload could not be calculated when VPC was (very close to) 0. * = indicator was excluded because of too little events (<100).
Figure 2 - Physician- and hospital-specific effects (performance scores) and 95% confidence intervals for 4 outcomes, ranked from good to poor. Red dots represent (physicians working in) general hospitals, whereas green triangles (in panel B and D) represent academic hospitals. Panel A: 30-day in-hospital mortality rate in CRC patients, physician level (VPC=0.06, R=0.67). Panel B: total in-hospital costs in UBC patients, hospital level (VPC=0.13, R=0.97). Panel C: 30-day readmission rate in TKA patients, physician level (VPC=0.02, R=0.60). Panel D: length of stay in days in PCI patients, hospital level (VPC=0.02, R=0.96). VPC = variance partitioning coefficient and R = reliability.