

# Physician-level variation in clinical outcomes and resource use in inpatient general internal medicine: an observational study

Amol A Verma <sup>1,2,3,4</sup> Yishan Guo,<sup>2</sup> Hae Young Jung,<sup>2</sup> Andreas Laupacis,<sup>1,2,3</sup> Muhammad Mamdani,<sup>1,2,3</sup> Allan S Detsky <sup>1,4,5,6</sup>, Adina Weirnerman,<sup>1,7</sup> Terence Tang,<sup>1,8</sup> Shail Rawal <sup>1,6</sup>, Lauren Lapointe-Shaw,<sup>1,6</sup> Janice L Kwan,<sup>1,5</sup> Fahad Razak<sup>1,2,3,4</sup>

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/bmjqs-2019-010425>).

For numbered affiliations see end of article.

## Correspondence to

Dr Amol A Verma, LKS-CHART, St Michael's Hospital, Toronto, ON M5B 1W8, Canada; [amol.verma@mail.utoronto.ca](mailto:amol.verma@mail.utoronto.ca)

Received 29 September 2019  
Revised 25 February 2020  
Accepted 4 March 2020  
Published Online First  
27 March 2020



© Author(s) (or their employer(s)) 2021. No commercial re-use. See rights and permissions. Published by BMJ.

**To cite:** Verma AA, Guo Y, Jung HY, et al. *BMJ Qual Saf* 2021;**30**:123–132.

## ABSTRACT

**Background** Variations in inpatient medical care are typically attributed to system, hospital or patient factors. Little is known about variations at the physician level within hospitals. We described the physician-level variation in clinical outcomes and resource use in general internal medicine (GIM).

**Methods** This was an observational study of all emergency admissions to GIM at seven hospitals in Ontario, Canada, over a 5-year period between 2010 and 2015. Physician-level variations in inpatient mortality, hospital length of stay, 30-day readmission and use of 'advanced imaging' (CT, MRI or ultrasound scans) were measured. Physicians were categorised into quartiles within each hospital for each outcome and then quartiles were pooled across all hospitals (eg, physicians in the highest quartile at each hospital were grouped together). We report absolute differences between physicians in the highest and lowest quartiles after matching admissions based on propensity scores to account for patient-level variation.

**Results** The sample included 103 085 admissions to 135 attending physicians. After propensity score matching, the difference between physicians in the highest and lowest quartiles for in-hospital mortality was 2.4% (95% CI 0.6% to 4.3%,  $p < 0.01$ ); for readmission was 3.3% (95% CI 0.7% to 5.9%,  $p < 0.01$ ); for advanced imaging was 0.32 tests per admission (95% CI 0.12 to 0.52,  $p < 0.01$ ); and for hospital length of stay was 1.2 additional days per admission (95% CI 0.5 to 1.9,  $p < 0.01$ ). Physician-level differences in length of stay and imaging use were consistent across numerous sensitivity analyses and stable over time. Differences in mortality and readmission were consistent across most sensitivity analyses but were not stable over time and estimates were limited by sample size.

**Conclusions** Patient outcomes and resource use in inpatient medical care varied substantially across physicians in this study. Physician-level variations in length of stay and imaging use were unlikely to be explained by patient factors whereas differences in mortality and readmission should be interpreted with caution and could be explained by unmeasured confounders. Physician-level variations may represent practice differences that highlight quality improvement opportunities.

## INTRODUCTION

Variations in inpatient medical care are well documented and typically attributed to system, hospital or patient factors.<sup>1–6</sup> This approach is reflected in the system-level focus that predominates in the quality improvement and patient safety literature.<sup>7</sup> Measurement efforts have also focused on the outcomes of care delivered by individual surgeons<sup>8–9</sup> and have described notable surgeon-level variations.<sup>10</sup> Recently, Tsugawa and colleagues reported marked variations in physician spending in inpatient medical care in US hospitals.<sup>11</sup> Higher spending was not associated with better clinical outcomes. Broad physician attributes, such as experience,<sup>12–13</sup> sex,<sup>14</sup> training<sup>15</sup> and specialty,<sup>16</sup> may have modest and mixed associations with variations in patient outcomes. However, the magnitude of physician-level variations in inpatient medical care, and the scope of variations across a range of process and outcome measures, have not been described. Research in this area is based on large administrative data sets,<sup>11–15 17 18</sup> which lack clinical data, or smaller clinical cohorts,<sup>19</sup> which are limited by sample size. Thus, it remains unclear whether physician-level variations in outcomes and resource use may be explained by patient-level differences.

Much of clinical care is driven by decisions made by individual physicians and their patients. A growing number of quality improvement initiatives, such as audit-and-feedback reports,<sup>20</sup> target individual physicians. Understanding the degree of physician-level variations in care is important for understanding the

potential effectiveness of physician-targeted quality improvement interventions. General internal medicine (GIM) inpatient care provides a unique opportunity to examine physician-level variation because admissions are non-elective and patients are 'quasirandomly' allocated to physicians based on work schedules.<sup>11</sup> Thus, physician-level variation within hospitals may reflect practice differences rather than differences in patient characteristics. Moreover, GIM patients account for nearly 40% of emergency department admissions to hospital.<sup>21</sup>

The purpose of this study was first to measure physician-level variations in resource use and selected patient outcomes in GIM at seven hospitals over 5 years. Second, we used a granular clinical data set to explore whether physician-level variations may reflect real practice differences or are more likely explained by differences related to patient characteristics or case mix. We focused on within-hospital variations to isolate physician-level variation from system or hospital factors.

## METHODS

### Design, setting and participants

This observational study included GIM inpatients at seven large hospital sites participating in the General Medicine Inpatient Initiative (GEMINI) in Toronto and Mississauga, two adjacent cities in Ontario, Canada.<sup>21</sup> Participating hospitals included five academic institutions and two large community-based hospitals affiliated with the University of Toronto. All participating healthcare organisations are independent, publicly funded and provide tertiary and/or quaternary care. Public insurance covers the costs of hospital care and physician services and thus patients come from all socioeconomic backgrounds. Residents and medical students from the University of Toronto rotate through all the participating hospitals.

GIM services at GEMINI hospitals have been previously described in detail.<sup>21</sup> Inpatient GIM care is provided in a hospitalist model, predominantly by clinical teaching teams in the academic centres and by non-teaching teams in the community hospitals. The attending physicians are internists (93%) and family physicians (7%) who typically cover the inpatient service in 'blocks' lasting between 1 and 4 consecutive weeks (most commonly 2 weeks). GIM admissions are almost entirely non-elective and through the emergency department. There are typically at least four physicians attending on GIM in parallel at each hospital at any given time. Patients are assigned to attending physicians 'quasirandomly' by the on-call internal medicine resident or staff physician at the time of admission in the emergency department. As a result, over large samples, patient characteristics should be balanced across physicians within a hospital. Thus, physician-level differences in clinical outcomes and resource use within a hospital can be attributed to differences in

clinical practice rather than differences in patient characteristics or case mix. This assumption has informed several large analyses examining emergency physician and hospitalist practice patterns in the USA,<sup>11 13 14 22</sup> but it has not been tested outside this setting or using clinical data to enhance risk adjustment.

### Inclusions and exclusions

We included all GIM hospitalisations between 1 April 2010 and 31 March 2015, defined as patients who were either admitted to or discharged from the GIM service. GIM services were distinguished from other hospital services based on data extracted from hospital information systems.<sup>21</sup> In order to maximise comparability between physicians within a hospital, we only included GIM visits for whom the 'most responsible physician' was an internist (including internal medicine subspecialists) who attends on the inpatient GIM service. A study investigator at each hospital identified the physicians who attended on the inpatient GIM service during the study period. We excluded family physician hospitalists because their training, case mix and practice patterns may differ systematically from internists. We excluded GIM patients who were admitted from any route other than the emergency department (n=4213, 3.4%), to avoid elective admissions and interhospital transfers that might not have been assigned to an attending physician quasirandomly. Finally, we excluded admissions with hospital length of stay greater than 30 days (n=5423, 4.4%), because it is difficult to attribute their care to a single physician. Hospital admissions were attributed to the 'most responsible physician' as per the Canadian Institute for Health Information (CIHI) Discharge Abstract Database, defined as the physician who is 'responsible for the care and treatment of the patient for the majority of the visit to the health care facility'.<sup>23</sup> The most responsible physician is assigned retrospectively by the hospital (typically by a trained chart abstractor) after discharge and if a definition has been used extensively in health services research.<sup>12 24–26</sup> We included only physicians responsible for at least 100 admissions over the study period to avoid unstable estimates related to small sample size. Two of the hospitals participating in GEMINI exist within a larger healthcare organisation. Transfers between hospitals within the organisation occurred in a small number (3.8%) of that organisation's hospitalisations. In these cases, the hospitalisation was attributed to the admitting hospital and was only included in the physician-level analysis if the most responsible physician was located at the admitting hospital.

### Data sources

Data collection for GEMINI has previously been described in detail<sup>21</sup> and included linking hospital administrative data with electronic clinical data at the level of individual hospital admissions. Demographic

and clinical data, including diagnoses coded using the *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Canada* (ICD-10-CA), were collected from hospital administrative databases as reported to CIHI. Laboratory and radiology data were extracted from hospital information systems.

### Outcome measures

We report four measures of clinical outcomes and resource use: inpatient mortality, hospital length of stay, 30-day readmission at any participating hospital and use of ‘advanced imaging’, defined as the number of diagnostic CT, MRI or ultrasound scans per admission. These were selected to represent important aspects of clinical care but are not intended to be direct indicators of quality or appropriateness. For example, in-hospital mortality is sometimes expected because delivering end-of-life care is an important part of GIM as many patients have life-limiting illnesses. Thus, we do not intend to suggest that a lower mortality rate or shorter hospital length of stay necessarily indicates higher quality care.

### Patient characteristics

The following patient characteristics were used in multivariable adjustment and matching: age, sex, Charlson Comorbidity Index score,<sup>27</sup> day of admission (categorised as weekday vs weekend), time of admission (categorised as daytime between 08:00 and 17:00 or night-time otherwise), fiscal year of admission, admission hospital, primary discharge diagnosis, admission to GIM within the previous 30 days and the Laboratory-based Acute Physiology Score (LAPS)<sup>28</sup> which ranges from 0 to 256 (higher scores reflect a greater risk of inpatient and 30-day mortality). The combination of age, sex, LAPS and Charlson Comorbidity Index score has been validated as a predictor of inpatient mortality in Ontario, Canada (reported c-statistic 0.89).<sup>29</sup> These covariates were included to adjust for patient-level differences in mortality risk. As we previously described,<sup>30</sup> discharge diagnoses were categorised based on ICD-10-CA codes into clinically meaningful and mutually exclusive groups, using the Clinical Classifications Software tool V.2018.1.<sup>31</sup>

### Statistical analysis

For our first research question, to measure physician-level variation, we report the unadjusted absolute differences between the physicians with the highest and lowest values on each outcome measure within each hospital. To test the statistical significance of physician-level differences at each hospital, we set the three lowest physicians as the reference group and performed pairwise comparisons between each physician and the reference group. To quantify physician-level variation across hospitals, we categorised physicians into quartiles within each hospital for each of

the four outcome measures. We then pooled the quartiles across all hospitals and compared the highest and lowest quartiles, testing for statistical significance with the Kruskal-Wallis test for continuous outcomes and  $\chi^2$  test for binary outcomes.

For our second research question, to examine whether differences in patient characteristics contributed to the observed physician-level variations, we performed seven analyses. First, we report patient baseline characteristics for all admissions across physician quartiles to determine whether they were balanced by the quasirandom allocation of patients to physicians. For each patient characteristic, we calculated the standardised difference between each quartile (eg, we calculated the standardised difference between quartiles 1 and 2, 1 and 3, 1 and 4, 2 and 3, 2 and 4, and so on). Standardised difference was computed as the mean difference between quartiles divided by the SD across hospitalisations in both quartiles. We report the largest standardised difference across all the pairwise comparisons for each patient characteristic, referring to this as the ‘maximum standardized difference’. Standardised differences greater than 0.1 are considered meaningful markers of imbalance.<sup>32</sup>

Second, we compared physicians in the highest and lowest quartiles of each outcome before and after restricting the comparison to a matched sample of hospital admissions to reduce bias that might arise from non-random assignment of patients to physicians attending in GIM. We matched admissions 1:1 based on patient age (categorised into 5-year intervals), sex, hospital site, fiscal year of admission and propensity score. The propensity score was calculated using logistic regression models predicting the probability of being in the highest physician quartile based on the baseline patient characteristics reported above. We compared estimates of physician-level differences before and after propensity score matching. We hypothesised that if patients were quasirandomly allocated to physicians, the estimates of physician-level differences would be similar before and after matching, suggesting that patient-level differences contributed little to the observed physician-level differences.

Third, because there is a large diversity in the conditions cared for in GIM,<sup>30</sup> we compared admissions with only the 10 most common primary discharge diagnoses to reduce potential confounding related to case mix differences among uncommon diagnoses. We again compared the highest and lowest physician quartiles after performing the same matching algorithm based on patient age, sex, hospital, fiscal year of admission and propensity score in this restricted sample.

Fourth, we fit multivariable linear or logistic regression models to estimate physician-level outcomes after adjusting for the baseline patient characteristics described above. We calculated the Pearson correlation between the adjusted and unadjusted physician-level

outcomes, with the expectation that a high degree of correlation suggests that patient-level factors explain relatively little physician-level variation.

Fifth, we performed a split sample validation to reduce the effects of unmeasured patient factors on physician-level differences and assess the stability in physician-level differences over time. The 5-year study sample was split into the first 2 years and last 3 years. Physicians were assigned to quartiles based on their within-hospital rankings on outcomes in the first part of the sample, and then patient characteristics and outcomes were reported for the same physician quartiles in the second part of the sample. To maximise stability of estimates, we included physicians with at least 100 admissions in each period. We performed a sample size calculation to estimate the sample that would be required to identify meaningful physician-level differences in mortality and readmission (see online supplementary appendix 1 for further details).

Sixth, we restricted the sample to hospitalisations with only a single attending physician to account for differences that might arise from misattribution in the 'most responsible physician' when there are multiple attending physicians involved in a patient's care. We

included only admissions with the same admitting, discharging and most responsible physician.

Seventh, to better understand physician-level differences in mortality, we excluded all patients with a palliative diagnostic code at hospital discharge (ICD-10-CA code Z51.5).

When reporting differences between physician quartiles, 95% CIs and p values were computed using adjusted SEs<sup>33</sup> to account for clustering at the hospital level. All statistical analyses were performed using R V.3.5.0 (R Foundation for Statistical Computing).

## RESULTS

There were 118 165 GIM admissions from the emergency department during the study period. After excluding physicians who were responsible for fewer than 100 admissions and admissions with length of stay greater than 30 days, the final study cohort included 103 085 admissions and 135 physicians (mean 764 admissions per physician, SD 533). The median patient age for admissions was 73 years (IQR 56–84), 50.6% were female and 42.6% had a Charlson Comorbidity Index score of 2 or greater (table 1).

**Table 1** Characteristics of all GIM hospital admissions and categorised across physician quartiles of in-hospital mortality

Characteristic	All hospital admissions*	Physician mortality quartile 1	Physician mortality quartile 2	Physician mortality quartile 3	Physician mortality quartile 4	Maximum standardised difference†
Admissions, n	103 085	19 781	24 298	27 568	19 786	
Mortality, n (%)	4474 (4.3)	761 (3.8)	1069 (4.4)	1416 (5.1)	1228 (6.2)	0.11
Age, median (IQR)	73 (56, 84)	71 (54, 83)	72 (55, 84)	73 (56, 84)	73 (56, 84)	0.07
Female sex, n (%)	52 192 (50.6)	50.6	50.3	51.0	50.2	0.02
High Charlson Comorbidity Index score‡, n (%)	43 910 (42.6)	42.9	41.9	42.1	43.0	0.02
Day of admission (weekday), n (%)	75 527 (73.3)	74.1	72.5	73.3	74.6	0.05
Time of admission (daytime)§, n (%)	22 283 (21.6)	21.2	23.7	19.6	22.4	0.10
Hospitalisation in previous 30 days¶, n (%)	11 091 (11.1)	11.3	11.3	11.2	11.5	0.01
LAPS**, mean (SD)	20.2 (17.5)	18.7 (16.7)	19.3 (16.9)	20.3 (17.1)	20.0 (17.2)	0.10
Discharge diagnoses††, n (%)						
Pneumonia	5602 (5.4)	1086 (5.5)	1380 (5.7)	1531 (5.6)	1095 (5.5)	0.01
Urinary tract infection	5517 (5.4)	1061 (5.4)	1190 (4.9)	1452 (5.3)	1022 (5.2)	0.02
Heart failure	5161 (5.0)	980 (5.0)	1335 (5.5)	1337 (4.8)	1027 (5.2)	0.03
COPD	4634 (4.5)	876 (4.4)	1181 (4.9)	1237 (4.5)	842 (4.3)	0.03
Stroke	3333 (3.2)	612 (3.1)	819 (3.4)	838 (3.0)	643 (3.2)	0.02
Gastrointestinal haemorrhage	3081 (3.0)	499 (2.5)	721 (3.0)	785 (2.8)	508 (2.6)	0.03
Delirium, dementia, cognitive disorder	2940 (2.9)	510 (2.6)	695 (2.9)	763 (2.8)	520 (2.6)	0.02
Fluid and electrolyte disorder	2765 (2.7)	490 (2.5)	661 (2.7)	754 (2.7)	495 (2.5)	0.02
Intestinal infection	2596 (2.4)	505 (2.6)	601 (2.5)	737 (2.7)	511 (2.6)	0.01
Diabetes mellitus with complication	2454 (2.4)	485 (2.5)	577 (2.4)	594 (2.2)	465 (2.4)	0.02

Physicians were categorised into quartiles within each hospital based on the rate of inpatient mortality and then quartiles were pooled across hospitals

\*The data reported for all hospital admissions reflect the entire study cohort (patients admitted or discharged from GIM), whereas the data reported for the physician mortality quartiles reflect only the admissions used in the calculation of mortality (only patients initially admitted to GIM).

†The maximum standardised difference across physician mortality quartiles was calculated using the two quartiles that had the largest standardised difference for each variable (calculated as the mean difference divided by SD).

‡Charlson Comorbidity Index score of 2 or greater was considered high comorbidity.

§Admissions between 08:00 and 17:00 were considered 'daytime' admissions.

¶To a GIM ward at a participating hospital.

\*\*Laboratory-based Acute Physiology Score (LAPS) is a validated score to predict inpatient mortality, which ranges from 0 to 256 points, with higher scores indicating a higher risk of mortality.

††The primary discharge diagnosis for each admission was categorised using the Clinical Classifications Software based on International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10) codes.

COPD, chronic obstructive pulmonary disease; GIM, general internal medicine.

### Physician-level variations in care and outcomes

There were marked within-hospital variations in inpatient mortality, length of stay, 30-day readmission and use of advanced imaging at all hospitals (figure 1). The absolute difference between the highest and lowest ranked physicians within each hospital in inpatient mortality ranged from 3.1% to 6.9% (figure 1), difference in average length of stay ranged from 1.4 to 3.6 days per admission, difference in readmission rate ranged from 4.4% to 12.7% and difference in use of advanced imaging ranged from 0.3 to 0.9 tests per admission.

The absolute difference between physicians in the highest and lowest quartiles at all hospitals was 2.4% (95% CI 0.7% to 4.1%,  $p=0.007$ ) for in-hospital mortality, 1.2 days (95% CI 0.6 to 1.9 days,  $p<0.001$ ) per admission in length of stay, 3.9% (95% CI 1.2% to 6.7%,  $p=0.005$ ) for readmission and 0.35 advanced imaging tests (95% CI 0.15 to 0.55 tests,  $p<0.001$ ) per admission (table 2).

### Patient-level differences and physician variations

The results of all seven analyses suggest that measured patient characteristics did not account for the observed physician-level variations in hospital length of stay and imaging use. Physician-level differences in mortality and readmission were consistent across all analyses except the split sample validation, which was limited by sample size. Patient characteristics, including predictors of mortality (eg, age, Charlson Comorbidity Index score and LAPS), were generally well balanced across physician quartiles for all outcomes, before and after matching (table 1, online supplementary tables A–K). Compared with unadjusted analyses, the physician-level differences were similar after propensity score matching and after propensity score matching in a sample restricted to only the 10 most common discharge diagnoses (table 2). There was a strong correlation between unadjusted and adjusted physician-level estimates for each outcome measure, with correlation coefficients ranging from 0.81 to 0.96 ( $p<0.001$  for all measurements, table 3).

In the split sample validation, there remained significant physician-level differences in length of stay (difference between highest and lowest quartiles was 0.7 days per admission, 95% CI 0.04 to 1.3 days,  $p=0.04$ ) and imaging use (difference between highest and lowest quartiles was 0.26 tests per admission, 95% CI 0.06 to 0.47 tests,  $p=0.01$ , online supplementary appendix tables M and O). There were no significant differences observed in the second period of the split sample for mortality or readmission (table 2, online supplementary tables L and N), but these findings were limited by sample size as only 80 physicians had enough hospitalisations in both periods to be included (see online supplementary appendix 1 for further details).

When the sample was restricted to admissions with only a single attending physician, 65 982 admissions

were included (64% of the original sample). Patient-level characteristics were well balanced across all quartiles and physician-level differences were similar to the main analysis (table 2, online supplementary tables P–S).

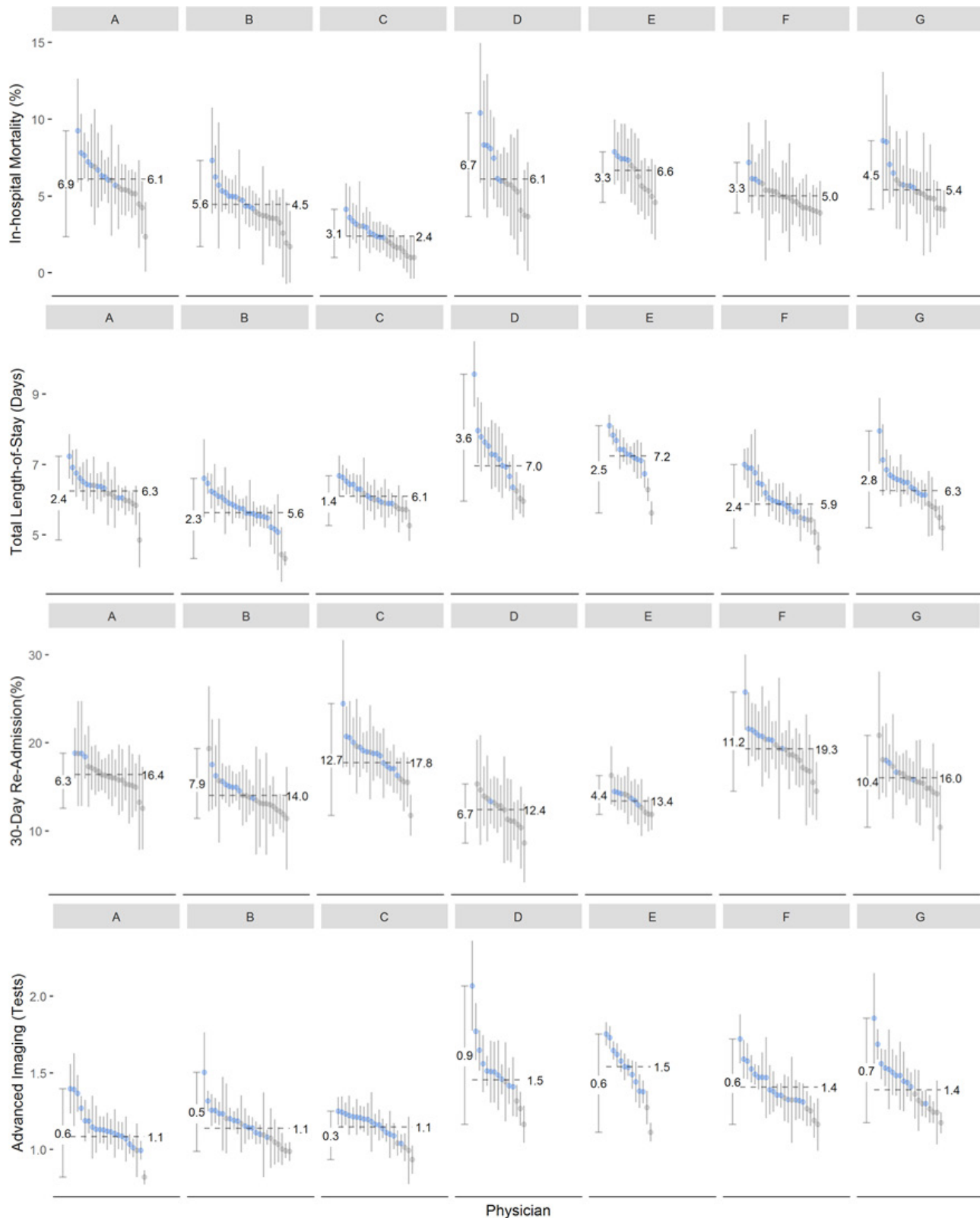
After excluding patients with a ‘palliative’ diagnostic code, the difference in mortality between the highest and lowest physician quartiles was 1.8% (95% CI 0.6% to 3.0%,  $p=0.005$ , online supplementary table T).

## DISCUSSION

### Principal findings

This multicentre study of 103 085 hospital admissions demonstrates significant physician-level variations in inpatient clinical outcomes and resource use in GIM at all participating hospitals. Using granular clinical data, we found that measured patient differences did not substantially account for the observed physician-level variation. Over large samples, measured patient characteristics and case mix were well balanced across physicians within a hospital. Physician-level variations persisted among hospitalisations with only a single attending physician, which removed misattribution that could result from handovers in care. Physician-level variation in hospital length of stay and advanced imaging use were consistent across multiple sensitivity analyses and stable over time in a split sample validation. Physician-level variations in mortality and readmission rate were consistent across numerous analyses but could not be demonstrated to be stable over time, which is likely due in part to insufficient sample size. Thus, our main findings regarding mortality and readmission should be interpreted with caution as unmeasured confounders could account for the observed variations. Our findings suggest that there are meaningful physician-level variations in care that are not explained by system, hospital or patient factors. Further research is needed to determine whether such variations exist primarily in process measures (such as imaging use) or whether they can also be stably demonstrated for clinical outcomes (such as mortality or readmission).

When comparing physicians in the highest and lowest quartiles in the same hospital on a matched sample of admissions, we found absolute differences of 2.4% in inpatient mortality (representing one additional death per 42 admissions), 3.3% in readmission (one additional readmission per 30 admissions), 1.2 days per admission in length of stay and 0.32 advanced imaging tests per admission (one additional test per three admissions). These effects could not be attributed to differences in measured patient risk. In particular, the physician-level estimates of mortality were not substantially changed after adjusting for age, sex, Charlson Comorbidity Index score and LAPS, the combination of which has been validated as a predictor of mortality risk.<sup>28 29</sup> Significant mortality differences remained after excluding deaths coded as ‘palliative’,



**Figure 1** Physician-level variations in in-hospital mortality, length of stay, 30-day readmission and use of advanced imaging. Physician-level data are presented for each hospital (A–G). The figure reports in-hospital mortality rate (panel A), mean total length of stay (panel B), rate of 30-day readmission (panel C) and use of advanced imaging, defined as number of CT, ultrasound (US) and MRI scans per admission (panel D). Physician-level data are presented for each hospital (A–G). Each dot represents the actual value for one physician and the vertical bars are the 95% CI of the estimate. The dotted horizontal lines are the hospital average, or overall hospital rate. The vertical bars for each hospital are the difference between the physicians with the highest and lowest values. Each physician was compared with the average of the three physicians with the lowest values at each hospital. When the difference was statistically significant ( $p < 0.05$ ), the physician data point is labelled blue. Grey labels represent no statistically significant difference.

**Table 2** Physician-level variations in resource use and outcomes in GIM

Outcome measures	Absolute difference between highest and lowest physician quartiles				
	All admissions (95% CI)	Propensity score matching (95% CI)	Propensity score matching, restricted sample* (95% CI)	Split sample validation†	Single attending physician‡
In-hospital mortality (%)	2.4 (0.7 to 4.1)	2.4 (0.6 to 4.3)	1.7 (0.3 to 3.0)	0.3 (−1.3, 2.0)	2.5 (1.1, 3.9)
Length of stay (mean days per admission)	1.2 (0.6 to 1.9)	1.2 (0.5 to 1.9)	1.2 (0.5 to 2.0)	0.7 (0.04, 1.34)	1.0 (0.1, 1.8)
30-day readmission (%)	3.9 (1.2 to 6.7)	3.3 (0.7 to 5.9)	4.1 (2.3 to 6.0)	0.8 (−2.0, 3.6)	4.1 (0.8, 7.3)
Advanced imaging use (mean tests per admission)	0.35 (0.15 to 0.55)	0.32 (0.12 to 0.52)	0.25 (0.03 to 0.48)	0.26 (0.06, 0.47)	0.30 (0.1, 0.5)

Physicians were categorised into quartiles within each hospital based on each outcome measure. We report the difference between the highest and lowest physician quartiles for each measure among all admissions and after matching. In the matched samples, each admission for physicians in the highest and lowest quartiles was matched 1:1 on age, sex, hospital site, fiscal year of admission and a propensity score that included all baseline characteristics.

\*Only admissions involving the 10 most common discharge diagnoses were included to avoid case mix differences that might be related to rare diagnoses.

†The 5-year sample was split and physicians were categorised into quartiles based on outcomes in the first 2 years and here we report outcomes in the same quartiles in the last 3 years.

‡Only admissions in which there was a single attending physician throughout were included.

GIM, general internal medicine.

suggesting that both unexpected and expected deaths contributed to the observed variation. However, physician-level differences in mortality were not consistent over time and thus should be interpreted with caution. Unmeasured factors, such as patient preferences regarding resuscitation, could account for the observed variation. We believe our findings are likely to be generalisable because the study sample included five hospitals across Toronto, Canada's largest city, and the only two hospitals that provide GIM care in Mississauga, Canada's sixth largest city. The organisation of inpatient GIM care in Ontario and most of Canada is similar to the hospitals included in this study. Given that GIM patients account for nearly 40% of emergency department admissions to hospital,<sup>21</sup> in aggregate, these differences have large system-wide effects.

### Clinical and policy implications

There is a well-developed body of literature examining clinical practice variation at the regional or hospital level.<sup>34–39</sup> Surgeon-level variations have also been described in several settings.<sup>8–10</sup> Less is known about physician-level variations in inpatient medicine. Studies have focused on comparing physicians across a range of characteristics. In general, there appear to be small or no differences in patient care and outcomes based on physician sex,<sup>14</sup> experience,<sup>12–13</sup> training<sup>15</sup> or type of practice (hospitalist vs primary care).<sup>16</sup> Patient care and outcomes differ between specialists and generalists for particular diseases, such as heart failure,<sup>19–24, 40–41</sup> however important case mix differences limit these comparisons. Because of this focus on specific physician characteristics, the magnitude and scope of physician-level variation in inpatient

**Table 3** Correlation between unadjusted and adjusted physician-level estimates of resource use and clinical outcomes in GIM

Hospital	Length of stay (Corr, p value)	Readmission (Corr, p value)	Mortality (Corr, p value)	Imaging (Corr, p value)
All hospitals	0.96 (<0.001)	0.96 (<0.001)	0.81 (<0.001)	0.96 (<0.001)
A	0.94 (<0.001)	0.98 (<0.001)	0.75 (<0.001)	0.99 (<0.001)
B	0.94 (<0.001)	0.91 (<0.001)	0.86 (<0.001)	0.89 (<0.001)
C	0.87 (<0.001)	0.96 (<0.001)	0.92 (<0.001)	0.92 (<0.001)
D	0.98 (<0.001)	0.90 (<0.001)	0.76 (0.001)	0.96 (<0.001)
E	0.96 (<0.001)	0.93 (<0.001)	0.81 (<0.001)	0.95 (<0.001)
F	0.98 (<0.001)	0.97 (<0.001)	0.77 (<0.001)	0.97 (<0.001)
G	0.98 (<0.001)	0.94 (<0.001)	0.84 (<0.001)	0.98 (<0.001)

The correlation was calculated between unadjusted and adjusted estimates of each measure. The following patient baseline covariates were used in adjustment: age, sex, Charlson Comorbidity Index score, day of admission, time of admission, Clinical Classifications Software (CCS) classification of discharge diagnosis, hospitalisation within 30 previous days, Laboratory-based Acute Physiology Score and study year. To examine the correlation across all hospitals as a single measure, all physicians were included and we adjusted for hospital fixed effects.

Corr, correlation coefficient; GIM, general internal medicine.

medicine has been poorly understood. Our study advances the understanding of quality measurement in inpatient medicine by identifying meaningful and stable physician-level differences in length of stay and imaging use and illuminating the need for further research regarding variations in mortality and readmission. This is particularly salient given the interest in audit and feedback as a means of quality improvement<sup>20</sup> and highlights the importance of standardised multicentre clinical data sets to support quality measurement. In Ontario, Canada, GIM physicians have historically received little systematic feedback about their patterns of clinical practice and thus opportunities for reflection, improvement or standardisation of care based on observed variations have been limited.

Recently, a series of papers using administrative databases in the USA has asserted that quasirandomisation in non-elective care permits physician-level comparisons.<sup>11 13–15 22</sup> To our knowledge, this hypothesis has not been tested outside of the USA or outside of administrative databases. We linked administrative data with detailed clinical data in a Canadian context and examined a more comprehensive set of patient characteristics than has been previously reported. For example, adjusting for laboratory data has been shown to account for substantial residual confounding compared with analyses based purely on administrative data,<sup>42 43</sup> and is an important strength of our study. Physician-level differences were similar across multiple analyses to account for patient characteristics. This detailed exploration suggests that patients were indeed quasirandomly allocated to physicians resulting in balanced characteristics within a hospital and permitting valid comparisons of physician-level practice differences.

The magnitude of physician-level variations observed in this study suggests that further work to understand and address these variations should be a priority for quality improvement. Further research is needed to better understand the drivers of physician-level practice differences in GIM, given that physician age, sex, training or experience explain only small amounts of variation.<sup>12–16</sup> Drivers may include different physician attitudes, skills or approaches to organising their practice. Alternatively, the observed physician-level clustering may reflect variations in the practices of different hospital wards or interdisciplinary teams.

### Limitations

Our study has several important limitations. First, we do not draw direct conclusions about appropriateness or quality of care. Although we identify large variations, we acknowledge that the extremes (eg, lowest mortality or shortest length of stay) may not reflect the highest quality care. Second, we collected data about readmissions at any participating study site and thus miss readmissions to other hospitals. In our region, 82% of hospital readmissions occur to the same site.<sup>44</sup>

Our data likely capture substantially more than 80% of all readmissions because we included readmissions to any participating hospital. It is unlikely that the limitations in this measure would affect physicians within a hospital differently and bias within-hospital comparisons. Third, we were unable to measure out-of-hospital mortality and thus choices about end-of-life care (eg, choosing to palliate patients who are terminally ill in hospital rather than transfer to another facility) may have affected physician-level differences. We attempted to address this by adjusting for validated predictors of mortality and by excluding patients with ‘palliative’ diagnostic codes, and results were consistent across these analyses. Fourth, we did not have sufficient sample size at the physician level to determine whether variations in mortality or readmission were stable over time. Fifth, resident physicians play an important role in the delivery of GIM care at most participating study sites and test ordering is known to vary among residents.<sup>45</sup> We were unable to include trainee schedules and thus could not account for this potential source of variation. However, residents in participating hospitals rotate across clinical teams and are not paired directly with individual attending physicians. Thus, over large samples, resident-driven differences would likely be non-differential between physicians. Physician-level variability was similar across all participating sites, including the two community hospitals where residents are less involved in clinical care. Moreover, the attending physician is ultimately responsible for care that is delivered by the residents under their supervision. Sixth, we were unable to measure important patient factors, such as socioeconomic status, functional status, cognition or patient preferences, which could vary across physicians and affect outcomes. Given these limitations, we believe that data about physician-level practice variations in hospital medicine should be used for formative feedback and to identify opportunities for quality improvement but not for formal evaluations or remuneration.

### CONCLUSION

Our study demonstrates that patient outcomes and resource use in inpatient medical care vary substantially across physicians. Physician-level variations in length of stay and imaging use are unlikely to be explained by patient factors and further research is needed to establish the stability of variations in mortality and readmission over time. Physician-level variations may represent practice differences that highlight quality improvement opportunities.

### Author affiliations

<sup>1</sup>Department of Medicine, University of Toronto, Toronto, Ontario, Canada

<sup>2</sup>Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Ontario, Canada

<sup>3</sup>Department of Medicine, St. Michael's Hospital, Toronto, Ontario, Canada

<sup>4</sup>Institute for Health Policy, Management, and Evaluation, University of Toronto, Toronto, Ontario, Canada



<sup>5</sup>Department of Medicine, Sinai Health System, Toronto, Ontario, Canada

<sup>6</sup>Department of Medicine, University Health Network, Toronto, Ontario, Canada

<sup>7</sup>Department of Medicine, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada

<sup>8</sup>Institute for Better Health, Trillium Health Partners, Mississauga, Ontario, Canada

**Correction notice** The article has been corrected since it was published online first. The Funding and the Disclaimer statements have been updated.

**Twitter** Amol A Verma @AmolAVerma, Allan S Detsky @Adetsky and Fahad Razak @FahadRazak

**Contributors** The study was designed by AAV and FR with input from all coauthors. YG and HYJ performed statistical analysis. The manuscript was drafted by AAV and all coauthors provided critical revision for important intellectual content and input in writing. AAV, YG, HYJ, AW, TT, SR, LLS, JLK and FR were involved in collecting data.

**Funding** This study was funded by Green Shield Canada Foundation and University of Toronto Division of General Internal Medicine. FR is supported by an award from the Mak Pak Chiu and Mak-Soo Lai Hing Chair in General Internal Medicine, University of Toronto.

**Disclaimer** The funding agencies and Ontario Health had no role in the design, conduct or interpretation of this study, and the views expressed herein do not reflect the views of the organisations.

**Competing interests** AAV and FR are employees of Ontario Health.

**Patient consent for publication** Not required.

**Ethics approval** Research ethics board approval was obtained from all participating hospitals. A waiver of participant consent was obtained from research ethics boards of all participating hospitals.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request. The study's lead investigators will make data for this manuscript available upon request as possible in compliance with local research ethics board requirements and data sharing agreements.

#### ORCID iDs

Amol A Verma <http://orcid.org/0000-0001-6857-931X>

Allan S Detsky <http://orcid.org/0000-0003-2772-3121>

Shail Rawal <http://orcid.org/0000-0002-8466-1193>

#### REFERENCES

- Jha AK, Perlin JB, Kizer KW, *et al.* Effect of the transformation of the Veterans Affairs health care system on the quality of care. *N Engl J Med* 2003;348:2218–27.
- Fisher ES, Bynum JP, Skinner JS. Slowing the growth of health care costs—lessons from regional variation. *N Engl J Med* 2009;360:849–52.
- Rathore SS, Foody JM, Wang Y, *et al.* Race, quality of care, and outcomes of elderly patients hospitalized with heart failure. *JAMA* 2003;289:2517–24.
- Chan PS, Nichol G, Krumholz HM, *et al.* Racial differences in survival after in-hospital cardiac arrest. *JAMA* 2009;302:1195–201.
- Burwen DR, Galusha DH, Lewis JM, *et al.* National and state trends in quality of care for acute myocardial infarction between 1994–1995 and 1998–1999: the Medicare health care quality improvement program. *Arch Intern Med* 2003;163:1430–9.
- Wang DE, Tsugawa Y, Figueroa JF, *et al.* Association between the centers for Medicare and Medicaid services Hospital STAR rating and patient outcomes. *JAMA Intern Med* 2016;176:848–50.
- Kohn LT, Corrigan JM, Donaldson MS. *To err is human*. Washington (DC), 2000.
- Jha AK, Epstein AM. The predictive accuracy of the new York state coronary artery bypass surgery report-card system. *Health Aff* 2006;25:844–55.
- Pierce O, Allen M. Assessing surgeon-level risk of patient harm during elective surgery for public reporting. *ProPublica* 2015:1–21.
- Birkmeyer JD, Finks JF, O'Reilly A, *et al.* Surgical skill and complication rates after bariatric surgery. *N Engl J Med* 2013;369:1434–42.
- Tsugawa Y, Jha AK, Newhouse JP, *et al.* Variation in physician spending and association with patient outcomes. *JAMA Intern Med* 2017;177:675.
- McAlister FA, Youngson E, Bakal JA, *et al.* Physician experience and outcomes among patients admitted to general internal medicine teaching wards. *CMAJ* 2015;187:1041–8.
- Tsugawa Y, Newhouse JP, Zaslavsky AM, *et al.* Physician age and outcomes in elderly patients in hospital in the US: observational study. *BMJ* 2017;357:j1797.
- Tsugawa Y, Jena AB, Figueroa JF, *et al.* Comparison of hospital mortality and readmission rates for Medicare patients treated by male vs female physicians. *JAMA Intern Med* 2017;177:206–13.
- Tsugawa Y, Jena AB, Orav EJ, *et al.* Quality of care delivered by general internists in US hospitals who graduated from foreign versus US medical schools: observational study. *BMJ* 2017;356:j273.
- Stevens JP, Nyweide DJ, Maresh S, *et al.* Comparison of hospital resource use and outcomes among hospitalists, primary care physicians, and other generalists. *JAMA Intern Med* 2017;177:1781–7.
- Singh S, Lin Y-L, Kuo Y-F, *et al.* Variation in the risk of readmission among hospitals: the relative contribution of patient, hospital and inpatient provider characteristics. *J Gen Intern Med* 2014;29:572–8.
- Goodwin JS, Lin Y-L, Singh S, *et al.* Variation in length of stay and outcomes among hospitalized patients attributable to hospitals and hospitalists. *J Gen Intern Med* 2013;28:370–6.
- Auerbach AD, Hamel MB, Davis RB, *et al.* Resource use and survival of patients hospitalized with congestive heart failure: differences in care by specialty of the attending physician. support Investigators. study to understand prognoses and preferences for outcomes and risks of treatments. *Ann Intern Med* 2000;132:191–200.
- Ivers N, Jamtvedt G, Flottorp S, *et al.* Audit and feedback : effects on professional practice and healthcare outcomes (Review). *Cochrane Database Syst Rev* 2012;6:CD000259.
- Verma AA, Guo Y, Kwan JL, *et al.* Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the general medicine inpatient initiative (gemini) retrospective cohort study. *CMAJ Open* 2017;5:E842–9.
- Barnett ML, Olenski AR, Jena AB. Opioid-prescribing patterns of emergency physicians and risk of long-term use. *N Engl J Med* 2017;376:663–73.
- Canadian Institute for Health Information. *Dad Abstracting manual, 2015-2016 edition*. Ottawa, ON, 2015.

- 24 Boom NK, Lee DS, Tu JV. Comparison of processes of care and clinical outcomes for patients newly hospitalized for heart failure attended by different physician specialists. *Am Heart J* 2012;163:252–9.
- 25 Jackevicius CA, Tsadok MA, Essebag V, *et al.* Early non-persistence with dabigatran and rivaroxaban in patients with atrial fibrillation. *Heart* 2017;103:1331–8.
- 26 Cujec B, Quan H, Jin Y, *et al.* Association between physician specialty and volumes of treated patients and mortality among patients hospitalized for newly diagnosed heart failure. *Am J Med* 2005;118:35–44.
- 27 Quan H, Li B, Couris CM, *et al.* Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge Abstracts using data from 6 countries. *Am J Epidemiol* 2011;173:676–82.
- 28 Escobar GJ, Greene JD, Scheirer P, *et al.* Risk-adjusting Hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Med Care* 2008;46:232–9.
- 29 van Walraven C, Escobar GJ, Greene JD, *et al.* The Kaiser Permanente inpatient risk adjustment methodology was valid in an external patient population. *J Clin Epidemiol* 2010;63:798–803.
- 30 Verma AA, Guo Y, Kwan JL, *et al.* Prevalence and costs of discharge diagnoses in inpatient general internal medicine: a multi-center cross-sectional study. *J Gen Intern Med* 2018;33:1899–904.
- 31 Healthcare Cost and Utilization Project. Beta clinical classifications software (CCS) for ICD-10-CM/PCS. agency for healthcare research and quality, 2017. Available: <https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp> [Accessed 8 Jul 2019].
- 32 Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun Stat Simul Comput* 2009;38:1228–34.
- 33 Donner A, Birkett N, Buck C. Randomization by cluster. sample size requirements and analysis. *Am J Epidemiol* 1981;114:906–14.
- 34 Wennberg J. Small area variations in health care delivery. *Science* 1973;182:1102–8.
- 35 Riddell CA, Hutcheon JA, Strumpf EC, *et al.* Inter-Institutional variation in use of caesarean delivery for labour dystocia. *J Obstet Gynaecol Can* 2017;39:988–95.
- 36 Ko DT, Tu JV, Austin PC, *et al.* Prevalence and extent of obstructive coronary artery disease among patients undergoing elective coronary catheterization in New York state and Ontario. *JAMA* 2013;310:163–9.
- 37 Newhouse JP, Garber AM. Geographic variation in health care spending in the United States: insights from an Institute of medicine report. *JAMA* 2013;310:24–5.
- 38 Zhang Y, Baik SH, Fendrick AM, *et al.* Comparing local and regional variation in health care spending. *N Engl J Med* 2012;367:1724–31.
- 39 Bragg F, Cromwell DA, Edozien LC, *et al.* Variation in rates of caesarean section among English NHS trusts after accounting for maternal and clinical risk: cross sectional study. *BMJ* 2010;341:c5065.
- 40 Jong P, Gong Y, Liu PP, *et al.* Care and outcomes of patients newly hospitalized for heart failure in the community treated by cardiologists compared with other specialists. *Circulation* 2003;108:184–91.
- 41 Birkhead JS, Weston C, Lowe D. Impact of specialty of admitting physician and type of hospital on care and outcome for myocardial infarction in England and Wales during 2004-5: observational study. *BMJ* 2006;332:1306–11.
- 42 Walker AS, Mason A, Quan TP, *et al.* Mortality risks associated with emergency admissions during weekends and public holidays: an analysis of electronic health records. *Lancet* 2017;390:62–72.
- 43 Hammill BG, Curtis LH, Fonarow GC, *et al.* Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. *Circ Cardiovasc Qual Outcomes* 2011;4:60–7.
- 44 Staples JA, Thiruchelvam D, Redelmeier DA. Site of hospital readmission and mortality: a population-based retrospective cohort study. *CMAJ Open* 2014;2:E77–85.
- 45 Geleris JD, Shih G, Logio L. Analysis of diagnostic test ordering habits among internal medicine residents. *JAMA Intern Med* 2018;178:1719–1283.

© 2021 Author(s) (or their employer(s)) 2021. No commercial re-use. See rights and permissions. Published by BMJ.