Original article

BJS Open Open Access

Benchmarking of abdominal surgery: a study evaluating the HARM score in a European national cohort

J. Helgeland[1], K. Skyrud[1], A. K. Lindahl[2], D. Keller[4] and K. M. Augestad[3,4]

[1]Division of Health Services, Norwegian Institute of Public Health, and [2]Division of Surgery, Akershus University Hospital, Oslo, and [3]Department of Quality Improvement, University Hospital of North Norway, Tromsø, Norway, and [4]Department of Surgery, Columbia University Medical Center, New York, USA

Correspondence to: Mr J. Helgeland, Norwegian Institute of Public Health, PO Box 222 Skøyen, 0213 Oslo, Norway (e-mail: jon.helgeland@fhi.no)

Background: **Reliable, easily accessible metrics of surgical quality are currently lacking. The HARM (HospitAl length of stay, Readmission and Mortality) score is a composite measure that has been validated across diverse surgical cohorts. The aim of this study was to validate the HARM score in a national population of patients undergoing abdominal surgery.**

Methods: **Data on all abdominal surgery in Norwegian hospitals from 2011 to 2017 were obtained from the Norwegian Patient Registry. Readmissions and 30-day postoperative complications as well as deaths in and out of hospital were evaluated. The HARM scoring algorithm was tested after adjustment by establishing a newly proposed length of stay score. The correlation between the HARM score and complications, as well as the ability of aggregated HARM scores to discriminate between hospitals, were analysed. Risk adjustment models were developed for nationwide hospital comparisons.**

Results: **The data consisted of 407 113 primary operations on 295 999 patients in 85 hospitals. The HARM score was associated with complications and complication severity (Goodman–Kruskal γ value 0·59). Surgical specialty was the dominating variable for risk adjustment. Based on 1-year data, the risk-adjusted score classified 16 hospitals as low HARM score and 16 as high HARM score of the 53 hospitals that had at least 30 operations.**

Conclusion: **The HARM score correlates with major outcomes and is associated with the presence and severity of complications. After risk adjustment, the HARM score discriminated strongly between hospitals in a European population of abdominal surgery.**

Funding information
Northern Norway Regional Health Authority, HST1245-15.

Presented to the annual Norwegian Surgical Meeting (Kirurgisk Høstmøte), Oslo, Norway, October 2019

Paper accepted 25 February 2020
Published online 21 April 2020 in Wiley Online Library (www.bjsopen.com). **DOI:** 10.1002/bjs5.50284

Introduction

Benchmarking of hospitals and hospital departments can help to improve quality and reduce variation in practice. The basic principle of benchmarking consists of identifying the best hospitals from defined outcome measures, identifying their processes and adapting them for quality improvement. The parameters used for hospital benchmarking must be easy identifiable, universal and objective[1].

The most important quality metrics of surgical outcomes include length of hospital stay (LOS), readmission and mortality. In particular, readmissions may occur if patients with a complication are discharged too early[2,3]. The HARM score is based on the quality metrics: length of HospitAl stay, Readmission (30 days) and Mortality rates. First published by Keller and colleagues[2], the HARM score is a simple, reliable and objective measure for assessing quality in colorectal surgery, and has been further refined and validated across gastrointestinal and bariatric surgery[4–6]. The HARM score has been shown to be correlated with complications and increasing complication severity[6].

However, the HARM score has not been validated in gynaecology, urology, gastrointestinal and vascular surgery, and has not been evaluated in a European setting. The objective of this study was to validate the HARM score in a European surgical population. The hypothesis was that the

HARM score would be a reliable quality measure, across surgical specialties, in a national sample.

Methods

Patient administrative data from all publicly financed Norwegian hospitals for the period 2011–2017 were obtained from the Norwegian Patient Registry (NPR). The data set contained type of admission (acute or elective), primary and secondary diagnosis codes according to the Norwegian version of ICD-10, surgical and medical procedures, age, sex, date and time of ward admission and discharge, and procedures, for all department stays[7]. The NPR conducts extensive checks of logical consistency in the data. Surgical procedures were coded according to the Norwegian version of the Nordic Medico-Statistical Committee Classification of Surgical Procedures[7].

All permanent residents in Norway have a personal identification number (PIN). The NPR encrypted the PIN for all patients with a valid PIN, allowing tracking of patients over time and between hospitals. Hospital data were linked with the NPR to provide date of death where applicable.

The initial data set of primary operations included all abdominal procedures. Procedures were classified by surgical specialty: gastrointestinal, gynaecological, urological and vascular procedures in the abdomen. Procedure codes and types are shown in *Table S1* (supporting information).

Diagnosis and procedure codes signifying complications were scored according to the Clavien–Dindo method[8] (*Tables S2* and *S3*, supporting information). Clavien–Dindo grade V (death) was not used in this study, as the data did not include cause of death, and death within 30 days is a component of the HARM score.

Procedures were included if the recorded procedure date was within the departmental stay and merged into operations when the time intervals did overlap, thus forming the unit of analysis for the study. In case of missing procedure start or end time of day, these were imputed as 0900 and 1500 hours respectively. Clavien–Dindo grades for the primary stay and any subsequent hospital stays within 30 days of the primary operation, but before any new primary operation, were merged with the primary operation data set. When there was more than one complication code, the highest Clavien–Dindo grade was retained. Departmental stays, possibly at different hospitals, were linked into hospital episodes if the time interval was less than 8 h[9].

The Charlson Co-morbidity Index, as revised to ICD-10, was determined from previous admissions 3 years before, but not including, the current episode of care[9,10]. The number of hospital episodes in the year before the operation was also calculated.

All primary operations were included if performed on patients aged at least 18 years and the hospital episode included one night of stay. Operations with a missing PIN, admission type or vital status, or with recorded date of death more than 24 h before department admission, were excluded.

All time intervals used in the analysis were counted from the start date and time of day of the primary operation: LOS, 30-day readmission in the same or a different hospital (R30), and 30-day mortality, regardless of place of death (D30). LOS was counted from operation start to the end of the complete hospital episode, regardless of any intervening transfers between departments or hospitals. Any all-cause emergency admissions occurring within 30 days of the operation start, but later than 8 h after the end of the hospital episode, were counted as readmissions. All time computations were exact to the nearest second.

Statistical analysis

For descriptive statistics, mean(s.d.) values were determined for continuous variables. LOS was scored from 0 to 5. To adapt LOS scoring to the Norwegian data, log-normal distributions were fitted to LOS, separately for emergency and elective operations. The score cut-off points were then derived from the 30, 60, 75, 90 and 95 per cent log-normal percentiles, rounded to whole numbers of days. Cut-off points longer than 30 days were set to 30 days. Eventually, following the method of Keller *et al.*[2], the HARM score was defined as: HARM = LOS score + 5 × D30 + (R30 and not D30).

To adjust hospital HARM scores for patient risk factors, regression models were fitted to the individual HARM scores (after logarithmic transformation), using the Bayesian information criterion and stepwise selection[11]. Three risk adjustment models were explored, based on different sets of explanatory variables: model 0: age (modelled by natural splines) and sex; model 1, as model 0, with Charlson Co-morbidity Index, number of previous hospital episodes and admission type (emergency or elective); model 2, as model 1, with surgical specialty. Two-way interactions between admission type and the other patient specific variables were included as candidate variables in models 1 and 2.

The risk-adjusted aggregate score for a hospital was calculated by first calculating, for each case in the sample, the expected HARM score in the model, but with the parameters for the particular hospital in question. These expected scores were then averaged, and the process was repeated for all hospitals in turn.

Table 1 Descriptive statistics for the full operation data set, by surgical specialty

	Several†	Gastrointestinal	Gynaecology	Urology	Vascular	All specialties
Age (years)*	58·0(16·0)	58·8(19·6)	52·6(15·8)	66·2(14·7)	71·5(11·0)	60·0(18·1)
Sex (%)						
F	83·6	48·1	99·9	23·6	35·2	52·2
M	16·4	51·9	0·1‡	76·4	64·8	47·8
Admission type (%)						
Emergency	18·8	55·2	15·1	15·3	30·1	34·4
Elective	81·2	44·8	85·0	84·7	69·91	65·6
Charlson Co-morbidity Index*	1·1(2·2)	1·0(2·1)	0·2(0·9)	1·0(1·9)	0·8(1·5)	0·9(1·9)
No. of previous admissions*	1·7(7·1)	1·8(7·1)	0·7(3·2)	2·3(10·3)	3·3(16·5)	1·8(8·1)
30-day readmission rate (%)	10·6	11·5	4·7	11·5	13·1	10·2
30-day mortality rate (%)	1·6	4·0	0·2	0·9	5·2	2·4
Clavien–Dindo complication grade (%)						
No complication	64·4	63·0	88·8	67·9	43·8	68·6
I	0·4	0·7	0·3	1·0	2·4	0·7
II	16·3	16·2	6·8	19·1	27·9	15·6
III	15·0	13·3	3·6	8·5	19·4	10·4
IV	3·9	6·9	0·6	3·6	6·5	4·7
No. of operations	16 131	188 545	77 180	111 158	14 099	407 113

*Values are mean(s.d.). †Procedures from more than one specialty occurring at one operation. ‡Related to gender affirmation surgery.

The estimated hospital effects in risk adjustment model 2 to assess the ability of the HARM score were used to discriminate between hospitals, based on the most recent 1-year data set only, after excluding hospitals with 30 or fewer operations. One measure of discriminatory power is the ratio of between-hospital variance to the median total variance, $\tau^2/(\tau^2 + \text{median}(\sigma_1{}^2))$, where τ is the between-hospital standard deviation and $\sigma_1, \ldots, \sigma_H$ are the standard deviations of the H hospital effect estimates, in this context denoted rankability[12]. To identify low- and high-HARM score outliers, the deviations of the estimated hospital effects from their 25 per cent trimmed mean were tested for being significantly below or above zero respectively[13]. The Guo–Romano procedure was used, with false discovery rate not exceeding 0·01 as the criterion for significance, thus correcting for multiple testing[14].

The degree of association between the HARM score and the Clavien–Dindo grade was measured by Goodman and Kruskal's γ[15]. The quintiles of mean hospital HARM scores and the percentage of operations with serious complications, defined as Clavien–Dindo grades III and IV, for each hospital, were cross-tabulated. Only hospitals with 100 operations or more were included, and γ was used to measure the strength of the association.

An alternative way of constructing the LOS score was evaluated, by fitting a cumulative logistic regression model to the Clavien–Dindo grades, with LOS as a continuous variable (modelled by splines), and admission type, R30 and D30 as explanatory variables. This analysis directly estimates candidate cut-off points that give high correlation with the response variable.

Relationships between missing procedure start times and other characteristics were explored by tabulation. To give an indication of the effect of missing times, LOS and total departmental length of stay were summarized for both complete and incomplete cases. As the total LOS is derived from complete data, this would give an indication of any systematic bias due to missing times. Two alternative ways of handling missing times were studied and compared with the imputation method described above: imputing all start times with the median start time (separately for emergency and elective procedures) (method 1); and using only calendar days to calculate LOS (method 2). The corresponding HARM scores were computed using the previously established LOS scoring, the association with Clavien–Dindo grades was evaluated, and outlier hospitals under model 2 were identified.

All data preprocessing and statistical analyses were performed in R version 3.5.1 (The R Foundation, Vienna, Austria)[16].

Results

Of a total of 588 232 operations, 437 420 included one night of stay. After exclusion of reoperations and operations

Table 2 Lower cut-off points for length of stay score categories

	LOS (days)	
LOS score	Emergency admission	Elective admission
0	–	–
1	1	1
2	4	3
3	8	4
4	19	7
5	30	10

LOS, length of stay.

Table 3 Distribution of HARM scores in present and previously published series

	HARM score			
	≤2	3	4	>4
Crawshaw *et al.*[5] (2017)	49·1	12·0	9·8	29·2
Brady *et al.*[4] (2018)	55·8	10·0	9·4	24·7
Present series	69·3	13·5	6·8	10·4

Values are percentages.

on patients aged less than 18 years, 409 483 remained. A further 2370 were excluded because of missing or inconsistent variables. The final analysis data set consisted of 407 113 primary operations on 295 999 patients in 85 hospitals. Operation start time was missing and consequently imputed in 26·6 per cent of the cases. In 2016, some hospital trusts did not identify individual hospitals correctly in the data reported to the NPR. After exclusion of operations with incorrect hospital names or in hospitals with fewer than 100 operations, 55 hospitals remained with 379 428 operations and 279 192 patients for hospital-specific analyses. The 1-year data set used to investigate the discriminatory power of the HARM score comprised 53 hospitals with 58 811 operations on 49 971 patients during 2017. All other analyses were performed on the full sample.

Table 4 Mean HARM score by age quartile, sex and surgical specialty

	HARM score
Age quartile (years)	
18–47	1·4
48–63	1·9
64–74	2·2
75–100	2·4
Sex	
F	1·9
M	2·0
Surgical specialty	
Several	3·0
Gastrointestinal	2·4
Gynaecology	1·1
Urology	1·5
Vascular	3·0

Descriptive statistics are shown in *Table 1*. Mean LOS for elective and emergency operations was 4·4 and 7·5 days respectively. The procedure for determining

Fig. 1 Proportions of Clavien–Dindo grades by HARM score, per operation

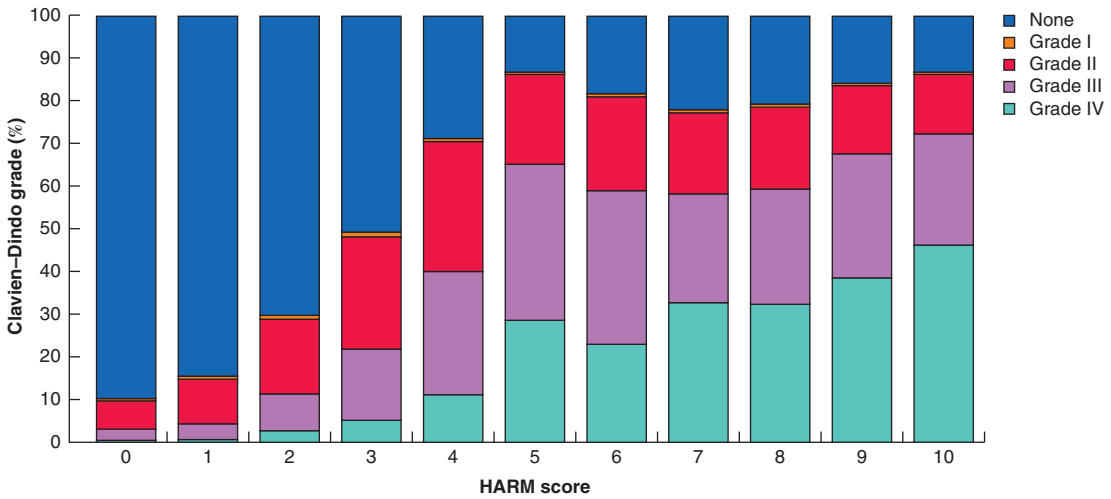


Table 5 Clavien–Dindo grade III–IV *versus* quintiles of mean HARM score per hospital

Quintile of hospital HARM score	Clavien–Dindo grade III–IV
1	4·9
2	11·0
3	12·4
4	14·7
5	16·6

Values are percentages.

LOS scores resulted in the cut-off points shown in *Table 2*.

The distribution of HARM scores is shown in *Table 3*. The overall mean was 1·94. For comparison with previous studies[4,5], corresponding percentage values are also shown. Notably, the proportion of scores exceeding 4 (death within 30 days, readmission with LOS score of 4, LOS score of 5) was markedly lower for the present study. The relative frequencies of these events were 2·4, 1·4 and 6·6 per cent respectively.

Table 4 shows mean HARM scores for age quartile, sex and surgical specialty.

The correspondence between HARM score and Clavien–Dindo grades per operation is shown in *Fig. 1* (Goodman–Kruskal γ 0·59). *Table 5* shows the correspondence between mean HARM scores and rate of serious complications per hospital (Goodman–Kruskal γ 0·14).

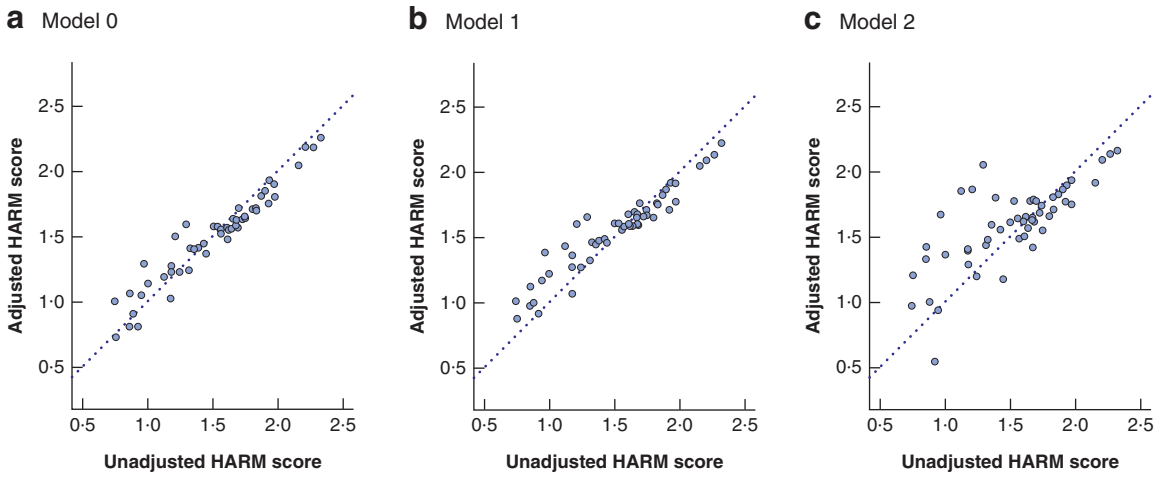
For the three risk adjustment models 0, 1 and 2, R^2 values were 0·10, 0·16 and 0·25 respectively. The corresponding risk-adjusted hospital mean HARM score for each model was calculated and compared with the unadjusted score (*Fig. 2*).

After fitting model 2 to the data for 2017 only, the rankability was 0·98; of the 53 hospitals, 16 hospitals were declared low-mortality outliers and 16 high-mortality outliers. Risk-adjusted aggregate scores for the 1-year data, with outlier status indicated, are shown in *Fig. 3*.

The alternative method for estimating LOS scores to give the best correspondence between the HARM score and Clavien–Dindo grade, using cumulative logistic regression, yielded very large LOS cut-off points, and thus did not discriminate between the main body of cases.

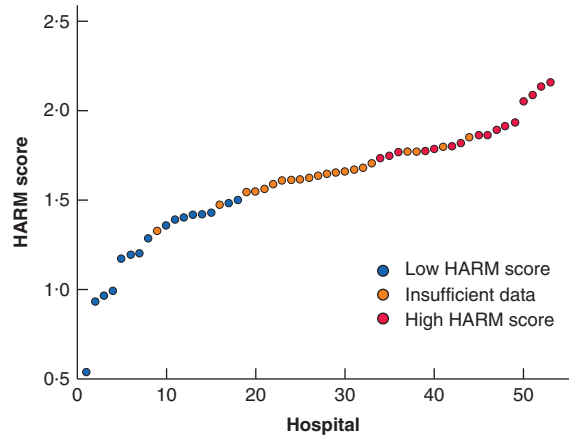
Cases with missing start time of day had somewhat longer LOS than those with complete data. The same increase was found for the total length of stay, counting from admission instead of operation start. The proportion of missing times was somewhat higher for emergency than for elective cases, and appeared to be weakly associated with hospital and surgical specialty. The two alternative ways of handling missing procedure times resulted in very small changes in Goodman–Kruskal γ values and rankability (data not shown). The number of non-outlier hospitals remained small (17 and 26 for the method 1 and 2 respectively).

Fig. 2 Risk-adjusted *versus* unadjusted hospital HARM scores



a Model 0 adjusts for age and sex. **b** Model 1, as model 0 plus adjustment for co-morbidity, admission type (emergency or elective) and previous admissions. **c** Model 2, as model 1 plus surgical specialty.

Fig. 3 Risk-adjusted HARM scores per hospital under model 2, with outlier status



Discussion

In this study, the HARM score was evaluated in patients undergoing abdominal surgery in Norwegian hospitals, after adjustment by establishing a new LOS score. The HARM score increased with age and was lower for gynaecological and urological than for gastrointestinal and vascular procedures. A moderate to strong association between HARM scores and complications, as measured by Clavien–Dindo grades, was documented for individual operations. Going from the hospitals in the lowest quintile of mean HARM score to the highest, the percentage of serious complications increased threefold. The mean HARM score differentiated well between surgical specialties.

Risk adjustment for age, sex, co-morbidity, type of admission and number of previous admissions was also studied and had only a moderate impact, whereas adjusting also for surgical specialty led to large adjustments. In particular, hospitals with low unadjusted scores tended to have substantially increased scores after adjustment. These hospitals had a different case mix, in particular a higher proportion of gynaecological and very few vascular operations.

For comparison of hospitals, the aggregated score discriminated strongly between hospitals. Of 53 hospitals, 16 were identified as low HARM score outliers and 16 as high HARM score outliers, using a multiple testing method and 1-year data.

This study was based on a large and recent sample, covering virtually all relevant abdominal surgeries in the nation over a 7-year period. NPR, the data source, has a high degree of completeness, as shown by comparing coverage of diagnoses and/or procedures with medical quality registries[17]. However, the data set did not cover outpatient visits. According to the NPR's guidelines, complications should not be reported unless they have had consequences for the care given, but actual practice may be variable. On the other hand, by including complications registered in subsequent hospital admissions, there is a risk of including complications unrelated to the primary operation. In addition, conditions counted as complications may have been present before surgery, as the NPR does not have a code for a condition being present on admission. In a recent study[18] in two Norwegian hospitals, the presence of surgical complications in patient administrative systems was found to have a sensitivity and specificity of 56 and 95 per cent respectively, and 76 and 65 per cent after exclusion of complications present on admission. A medical registry study[19] found a 14·0 per cent rate of major complications after elective colonic cancer surgery in Norway. This corresponds well with the finding of a 20·2 per cent rate of complications with Clavien–Dindo grade III–IV for all gastrointestinal operations.

The rate of missing time of day for procedures was relatively high. Cases with imputed start times had longer LOS than the rest. However, the same cases also had longer departmental length of stay. The incompleteness was therefore deemed inconsequential for the present methodological development study, although it could pose a problem in the actual use of the HARM score in benchmarking or quality reporting. However, the results were changed only marginally when the score was computed without exact start times. Other variables showed a high degree of completeness and consistency. Although it could be useful for risk adjustment, this study did not measure the severity of illness or the complexity and duration of the surgery. A recent Norwegian study[20] of gastrointestinal cancer surgery concluded that an appropriate measure of LOS should include not only transfers, but also readmissions.

The present findings are consistent with previous studies of the HARM score. However, a lower occurrence of the highest HARM scores was documented. A reasonable explanation could be the low rate of 30-day mortality and readmission, as well as the short hospital stays in these data.

The HARM score is a composite quality measurement, motivated by the fact that simple rates of risk-adjusted morbidity or mortality may not reliably reflect hospital performance with surgery. In the literature, several composite surgical quality measurements have been described[21–29].

All of these metrics are fairly new, and none seems to have been firmly established.

Acknowledgements

This work was supported by a grant from the Northern Norway Regional Health Authority (project HST1245-15).

Data from the NPR have been used in this article. The interpretation and reporting of these data are the sole responsibility of the authors, and no endorsement by the NPR is intended, nor should it be inferred.

Disclosure: The authors declare no conflict of interest.

References

1 Staiger RD, Schwandt H, Puhan MA, Clavien PA. Improving surgical outcomes through benchmarking. *Br J Surg* 2019; **106**: 59–64.
2 Keller DS, Chien HL, Hashemi L, Senagore AJ, Delaney CP. The HARM score: a novel, easy measure to evaluate quality and outcomes in colorectal surgery. *Ann Surg* 2014; **259**: 1119–1125.
3 Tsai TC, Joynt KE, Orav EJ, Gawande AA, Jha AK. Variation in surgical-readmission rates and quality of hospital care. *N Engl J Med* 2013; **369**: 1134–1142.
4 Brady JT, Ko B, Hohmann SF, Crawshaw BP, Leinicke JA, Steele SR *et al.* Application of a simple, affordable quality metric tool to colorectal, upper gastrointestinal, hernia, and hepatobiliary surgery patients: the HARM score. *Surg Endosc* 2018; **32**: 2886–2893.
5 Crawshaw BP, Keller DS, Brady JT, Augestad KM, Schiltz NK, Koroukian SM *et al.* The HARM score for gastrointestinal surgery: application and validation of a novel, reliable and simple tool to measure surgical quality and outcomes. *Am J Surg* 2017; **213**: 575–578.
6 Janik MR, Mustafa RR, Rogula TG, Saleh AA, Abbas M, Khaitan L. Application of HARM score to measure surgical quality and outcomes in bariatric patients. *Obes Surg* 2018; **28**: 2815–2819.
7 Norwegian Directorate of eHealth. *Helsefaglige kodeverk*: 2018. https://ehelse.no/standarder-kodeverk-og-referansekatalog/helsefaglige-kodeverk [accessed 6 June 2018].
8 Clavien PA, Barkun J, de Oliveira ML, Vauthey JN, Dindo D, Schulick RD *et al.* The Clavien–Dindo classification of surgical complications: five-year experience. *Ann Surg* 2009; **250**: 187–196.
9 Hassani S, Lindman AS, Kristoffersen DT, Tomic O, Helgeland J. 30-day survival probabilities as a quality indicator for Norwegian hospitals: data management and analysis. *PLoS One* 2015; **10**: e0136547.
10 Quan H, Li B, Couris CM, Fushimi K, Graham P, Hider P *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**: 676–682.
11 Schwarz G. Estimating the dimension of a model. *Ann Statist* 1978; **6**: 461–464.
12 van Dishoeck AM, Lingsma HF, Mackenbach JP, Steyerberg EW. Random variation and rankability of hospitals using outcome indicators. *BMJ Qual Saf* 2011; **20**: 869–874.
13 Kristoffersen DT, Helgeland J, Clench-Aas J, Laake P, Veierød MB. Observed to expected or logistic regression to identify hospitals with high or low 30-day mortality? *PLoS One* 2018; **13**: e0195248.
14 Guo W, Romano JP. On stepwise control of directional errors under independence and some dependence. *J Stat Plan Infer* 2015; **163**: 21–33.
15 Goodman LA, Kruskal WH. Measures of association for cross classifications. *J Am Stat Assoc* 1954; **49**: 732–764.
16 R Core Team. *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, 2019.
17 Bakken IJ, Ariansen AMS, Knudsen GP, Johansen KI, Vollset SE. The Norwegian patient registry and the Norwegian registry for primary health care: research potential of two nationwide health-care registries. *Scand J Public Health* 2020; **48**: 49–55.
18 Storesund A, Haugen AS, Hjortås M, Nortvedt MW, Flaatten H, Eide GE *et al.* Accuracy of surgical complication rate estimation using ICD-10 codes. *Br J Surg* 2019; **106**: 236–244.
19 Nymo LS, Norderval S, Eriksen MT, Wasmuth HH, Kørner H, Bjørnbeth BA *et al.* Short-term outcomes after elective colon cancer surgery: an observational study from the Norwegian registry for gastrointestinal and HPB surgery, NoRGast. *Surg Endosc* 2019; **33**: 2821–2833.
20 Lassen K, Nymo LS, Olsen F, Søreide K. Benchmarking of aggregated length of stay after open and laparoscopic surgery for cancers of the digestive system. *BJS Open* 2018; **2**: 246–253.
21 Dimick JB, Staiger DO, Hall BL, Ko CY, Birkmeyer JD. Composite measures for profiling hospitals on surgical morbidity. *Ann Surg* 2013; **257**: 67–72.
22 Karthaus EG, Lijftogt N, Busweiler LAD, Elsman BHP, Wouters MWJM, Vahl AC *et al.*; Dutch Society of Vascular Surgery, the Steering Committee of the Dutch Surgical Aneurysm Audit, the Dutch Institute for Clinical Auditing. Textbook outcome: a composite measure for quality of elective aneurysm surgery. *Ann Surg* 2017; **266**: 898–904.
23 Merath K, Chen Q, Bagante F, Beal E, Akgul O, Dillhoff M *et al.* Textbook outcomes among Medicare patients undergoing hepatopancreatic surgery. *Ann Surg* 2018: https://doi.org/10.1097/SLA.0000000000003105 [Epub ahead of print].
24 Lingsma HF, Bottle A, Middleton S, Kievit J, Steyerberg EW, Marang-van de Mheen PJ. Evaluation of hospital outcomes: the relation between length-of-stay, readmission, and mortality in a large international administrative database. *BMC Health Serv Res* 2018; **18**: 116.

25 Hofstede SN, Ceyisakar IE, Lingsma HF, Kringos DS, Marang-van de Mheen PJ. Ranking hospitals: do we gain reliability by using composite rather than individual indicators? *BMJ Qual Saf* 2019; **28**: 94–102.
26 Clavien PA, Vetter D, Staiger RD, Slankamenac K, Mehra T, Graf R *et al.* The Comprehensive Complication Index (CCI®): added value and clinical perspectives 3 years 'down the line'. *Ann Surg* 2017; **265**: 1045–1050.
27 Slankamenac K, Graf R, Barkun J, Puhan MA, Clavien PA. The comprehensive complication index: a novel continuous scale to measure surgical morbidity. *Ann Surg* 2013; **258**: 1–7.
28 Dimick JB, Birkmeyer NJ, Finks JF, Share DA, English WJ, Carlin AM *et al.* Composite measures for profiling hospitals on bariatric surgery performance. *JAMA Surg* 2014; **149**: 10–16.
29 Rajaram R, Barnard C, Bilimoria KY. Concerns about using the patient safety indicator-90 composite in pay-for-performance programs. *JAMA* 2015; **313**: 897–898.

Supporting information

Additional supporting information can be found online in the Supporting Information section at the end of the article.