Royal College of Surgeons
ADVANCING SURGICAL CARE

ORAL AND MAXILLOFACIAL SURGERY

Ann R Coll Surg Engl 2017; **99:** 299–306
doi 10.1308/rcsann.2016.0319

Validating a benchmarking tool for audit of early outcomes after operations for head and neck cancer

D Tighe[1], I Sassoon[2], M McGurk[3]

[1]Queen Victoria Hospital NHS Foundation Trust, UK
[2]King's College London, UK
[3]Guy's and St Thomas' NHS Foundation Trust, UK

ABSTRACT

INTRODUCTION In 2013 all UK surgical specialties, with the exception of head and neck surgery, published outcome data adjusted for case mix for indicator operations. This paper reports a pilot study to validate a previously published risk adjustment score on patients from separate UK cancer centres.

METHODS A case note audit was performed of 1,075 patients undergoing 1,218 operations for head and neck squamous cell carcinoma under general anaesthesia in 4 surgical centres. A logistic regression equation predicting for all complications, previously validated internally at sites A–C, was tested on a fourth external validation sample (site D, 172 operations) using receiver operating characteristic curves, Hosmer–Lemeshow goodness of fit analysis and Brier scores.

RESULTS Thirty-day complication rates varied widely (34–51%) between the centres. The predictive score allowed imperfect risk adjustment (area under the curve: 0.70), with Hosmer–Lemeshow analysis suggesting good calibration. The Brier score changed from 0.19 for sites A–C to 0.23 when site D was also included, suggesting poor accuracy overall.

CONCLUSIONS Marked differences in operative risk and patient case mix captured by the risk adjustment score do not explain all the differences in observed outcomes. Further investigation with different methods is recommended to improve modelling of risk. Morbidity is common, and usually has a major impact on patient recovery, ward occupancy, hospital finances and patient perception of quality of care. We hope comparative audit will highlight good performance and challenge underperformance where it exists.

KEYWORDS
Clinical audit – Outcomes – Complications – Squamous cell carcinoma

Accepted 15 June 2016

CORRESPONDENCE TO
David Tighe, E: dft1@doctors.org.uk

National audit of surgical outcomes is well established in the UK surgical specialties. With the exception of head and neck oncology surgery, all specialties attempt adjustment for patient case mix when presenting outcome data. A large national dataset (>10,000 patient care episodes) based on data derived from Hospital Episode Statistics (HES) has recently attempted to address this issue in presenting complication rates for surgical units treating patients for head and neck cancer.[1]

In 2013 the American College of Surgeons made an online benchmarking tool available to clinicians and the public.[2] Using coding data of 393 participating hospitals (1.4 million operations), it produces a calculated risk of 9 complications based on entry of 22 preoperative patient characteristics. The development of this tool is an ambitious programme of audit of case mix adjusted outcomes, allowing health insurance providers to align reimbursements with outcomes consistent with quality care. A secondary benefit is the provision of information for enhanced patient consent; by adjusting risk for patient factors, clinicians can produce patient specific probability data pertaining to complications and mortality. However, this online score does not as yet allow calculation of risk based on ablative and reconstructive aspects of a major head and neck case.

Both in the US and the UK, work towards risk adjusted audit of outcomes is limited by concerns about the accuracy of data derived from clinical coders and the application of imperfect models to individual patient discussions. We sought to address the need for good quality multicentre audit data in a previous case note audit of three surgical units (sites A–C) in the south-east of England, auditing outcomes of over 1,000 operations, leading to the development of a pilot logistic regression score of three variables.[3] This risk algorithm requires three data points: World Health Organization (WHO) performance status, a description of the complexity of surgery ('minor', 'intermediate' or 'major'; derived

Complication grade	Definition
Grade 1	Any deviation from the expected post operative course that does not require specific treatment
Grade 2	Complications requiring drug therapy, bllod transfusion or nutritional support
Grade 3	Postoperative changes that require invasive treatment (puncture, drainage and re-operations)
Grade 4	Complications with imminent risk of death and need for intensive care
Grade 5	Postoperative death

Figure 1 Clavien–Dindo classification of surgical complications[4]

from the Bupa classification of operative severity) and an estimation of intraoperative blood loss.

The risk model was validated internally, and demonstrated good calibration and reasonable discrimination for the purposes of case mix adjustment.[3] The present study applied the logistic regression model to a fourth treatment unit (site D) as a validation of its accuracy and utility.

Methods

A case note audit of a two-year period (June 2013 – June 2015) was undertaken by the lead author at site D. Data collection was retrospective for the first half of this period but prospective for the second half. Complications were recorded according to the Clavien–Dindo classification (Fig 1).[4] Predicted morbidity values were derived for each patient care episode using the previously published algorithm.[3] Data pertaining to patient demographics, indices of functional status, tumour stage, and operative and anaesthetic treatment were also included in the audit for the purpose of further model development. The Waterlow score, which has been shown to be useful as a means of predicting postoperative morbidity,[5] was tested for consideration of incorporation in a future scoring system.

Statistical analysis

Calibration, discrimination and accuracy of the risk model were examined in the form of receiver operating characteristic curves, Hosmer–Lemeshow analysis and Brier scores. Excel® (Microsoft, Redmond, WA, US) was used to generate Brier scores while MedCalc® version 16.2 (MedCalc Software, Mariakerke, Belgium) was used for the remaining analyses (chi-squared tests for categorical data, analysis of variance [ANOVA] and Kruskal–Wallis tests for continuous data).

Table 1 Comparison of patient data by hospital site

	Site A (*n*=180)	Site B (*n*=234)	Site C (*n*=616)	Site D (*n*=188)	Total (*n*=1,218)	λ^2	*p*-value
Age						F-ratio 9.96	**<0.001**
Mean age	65.7	66.6	61.8	66.2			
95% CI	62.9–67.3	64.4–68.2	60.1–62.6	64.5–68.6			
Range	23–93	35–100	22–93	27–96			
Sex						14	**0.003**
Male	119 (66.1%)	170 (72.6%)	370 (60.1%)	130 (72.6%)	789 (64.8%)		
Female	61 (33.9%)	64 (27.4%)	246 (39.9%)	58 (27.4%)	429 (35.2%)		
WHO performance status						347	**<0.00001**
0	29 (16.1%)	16 (6.8%)	399 (64.8%)	110 (58.5%)	554 (45.5%)		
1	105 (58.3%)	139 (59.4%)	141 (22.9%)	36 (19.1%)	421 (34.6%)		
2	30 (16.7%)	61 (26.1%)	42 (6.8%)	24 (12.8%)	157 (12.9%)		
3	8 (4.4%)	14 (6.0%)	16 (2.6%)	18 (9.6%)	56 (4.6%)		
4	0 (0%)	1 (0.4%)	0 (0%)	0 (0%)	1 (0.1%)		
Data missing	8 (4.4%)	3 (1.3%)	18 (2.9%)	0 (0%)	29 (2.4%)		
Scale of surgery						158	**<0.0001**
Minor (<1h)	51 (28.3%)	48 (20.5%)	90 (14.6%)	29 (15.4%)	218 (17.9%)		
Intermediate	73 (40.6%)	110 (47.0%)	174 (28.2%)	36 (19.1%)	393 (32.3%)		
Major (>6h +/- flap)	53 (29.4%)	76 (32.5%)	268 (43.5%)	123 (65.4%)	520 (42.7%)		
Data missing	3 (1.7%)	0 (0%)	84 (13.6%)	0 (0%)	87 (7.1%)		

Alcohol consumption						190	**<0.00001**
No alcohol	78 (43.3%)	97 (41.5%)	67 (10.9%)	60 (31.9%)	302 (24.8%)		
Light	37 (20.6%)	65 (27.8%)	236 (38.3%)	55 (29.3%)	393 (32.3%)		
Moderate	7 (3.9%)	17 (7.3%)	88 (14.3%)	39 (20.7%)	151 (12.4%)		
Heavy	35 (19.4%)	36 (15.4%)	121 (19.6%)	21 (11.2%)	213 (17.5%)		
Ex-heavy	5 (2.8%)	14 (6.0%)	61 (9.9%)	13 (6.9%)	93 (7.6%)		
Data missing	18 (10.0%)	5 (2.1%)	43 (7.0%)	0 (0%)	66 (5.4%)		
Smoking history						36	**<0.0001**
Never smoked	63 (35.0%)	96 (41.0%)	138 (22.4%)	69 (36.7%)	366 (30.0%)		
Smoker/ex-smoker	100 (55.6%)	134 (57.3%)	447 (72.6%)	119 (63.3%)	800 (65.7%)		
Data missing	17 (9.4%)	4 (1.7%)	31 (5.0%)	0 (0%)	52 (4.3%)		
ACE-27 score						358	**<0.0001**
0	73 (40.6%)	7 (3.0%)	279 (45.3%)	40 (21.3%)	399 (32.8%)		
1	64 (35.6%)	148 (63.2%)	215 (34.9%)	105 (55.9%)	532 (43.7%)		
2	36 (20.0%)	73 (31.2%)	48 (7.8%)	36 (19.1%)	193 (15.8%)		
3	1 (0.6%)	5 (2.1%)	3 (0.5%)	7 (3.7%)	16 (1.3%)		
Data missing	6 (3.3%)	1 (0.4%)	71 (11.5%)	0 (0%)	78 (6.4%)		
MUST score						948	**<0.0001**
0	56 (31.1%)	168 (71.8%)	0 (0%)	151 (80.3%)	375 (30.8%)		
1	7 (3.9%)	16 (6.8%)	0 (0%)	12 (6.4%)	35 (2.9%)		
2	2 (1.1%)	16 (6.8%)	0 (0%)	12 (6.4%)	30 (2.5%)		
3	0 (0%)	5 (2.1%)	0 (0%)	6 (3.2%)	11 (0.9%)		
4	0 (0%)	0 (0%)	0 (0%)	3 (1.6%)	3 (0.2%)		
5	0 (0%)	0 (0%)	0 (0%)	2 (1.1%)	2 (0.2%)		
Data missing	115 (63.9%)	29 (12.4%)	616 (100%)	2 (1.1%)	762 (62.6%)		
T stage						125	**<0.0001**
0/x	30 (16.7%)	61 (26.1%)	38 (6.2%)	36 (19.1%)	165 (13.5%)		
1	69 (38.3%)	62 (26.5%)	164 (26.6%)	38 (20.2%)	333 (27.3%)		
2	33 (18.3%)	38 (16.2%)	175 (28.4%)	35 (18.6%)	281 (23.1%)		
3	9 (5.0%)	14 (6.0%)	82 (13.3%)	10 (5.3%)	115 (9.4%)		
4	32 (17.8%)	53 (22.6%)	131 (21.3%)	67 (35.6%)	283 (23.2%)		
Data missing	7 (3.9%)	6 (2.6%)	26 (4.2%)	2 (1.1%)	41 (3.4%)		
N stage						71	**<0.0001**
0/x	97 (53.9%)	127 (54.3%)	386 (62.7%)	105 (55.9%)	715 (58.7%)		
1	24 (13.3%)	29 (12.4%)	108 (17.5%)	20 (10.6%)	181 (14.9%)		
2a	15 (8.3%)	17 (7.3%)	31 (5.0%)	6 (3.2%)	69 (5.7%)		
2b	29 (16.1%)	33 (14.1%)	29 (4.7%)	40 (21.3%)	131 (10.8%)		
2c	6 (3.3%)	10 (4.3%)	16 (2.6%)	9 (4.8%)	41 (3.4%)		
3	3 (1.7%)	7 (3.0%)	17 (2.8%)	5 (2.7%)	32 (2.6%)		
Data missing	6 (3.3%)	11 (4.7%)	29 (4.7%)	3 (1.6%)	49 (4.0%)		

ACE = Adult Co-morbidity Evaluation; CI = confidence interval; MUST = Malnutrition Universal Screening Tool; WHO = World Health Organization

Results

During the study period, 172 consecutive patients received 185 curative operations for head and neck squamous cell carcinoma at site D. Thirteen of these had recurrent disease. When added to the datasets for sites A–C, this totalled 1,075 patients receiving 1,218 operations. Of these, 96 patients had missing demographic data or were lacking recorded outcomes. The majority (n=83) of these cases were from the historical cohort. This left 979 complete patient care episodes for the later part of the analysis (model development). Case mix and surgical management was heterogeneous for all criteria (p<0.05) (Table 1). Across all four sites, 218 operations (18%) were classified as minor, 393 (32%) as intermediate and 520 (43%) as major operations.

For the 979 patient care episodes analysed, postoperative complications within 30 days were frequent overall (43%) and the raw complication rates differed significantly between sites. Site A had the fewest complications (35%) and site D had the most (51%) (λ^2=10.3, p=0.016). Overall, wound complications, namely dehiscence and infections (5.6% and 4.3% respectively), and pneumonia (4.9%) were the most common (Table 2). Complete flap failure rates were similar across the sites (site A: 2/160 [1.3%], site B: 3/208 [1.4%], site C: 10/439 [2.3%], site D: 8/172 [4.7%]; λ^2=4.4, p=0.21) but severe complications (Clavien–Dindo grade ≥III)

Table 2 Complications by hospital site

Complications	Site A (n=160)	Site B (n=208)	Site C (n=439)	Site D (n=172)	Total(n=979)
Wound					
Loss of flap	2 (1.3%)	3 (1.4%)	10 (2.3%)	8 (4.7%)	23 (2.3%)
Partial loss of flap	0 (0%)	0 (0%)	8 (1.8%)	2 (1.2%)	10 (1.0%)
Wound dehiscence	11 (6.9%)	8 (3.8%)	21 (4.8%)	15 (8.7%)	55 (5.6%)
Wound infection	9 (5.6%)	7 (3.8%)	8 (1.8%)	18 (10.5%)	42 (4.3%)
Haematoma	4 (2.5%)	4 (1.9%)	9 (2.1%)	11 (6.4%)	28 (2.9%)
Orocutaneous fistula	1 (0.6%)	0 (0%)	6 (1.4%)	4 (2.3%)	11 (1.1%)
Chyle leak	1 (0.6%)	3 (1.4%)	3 (0.7%)	2 (1.2%)	9 (0.9%)
Neck abscess	0 (0%)	0 (0%)	3 (0.7%)	0 (0%)	3 (0.3%)
Cardiovascular					
Atrial fibrillation	2 (1.3%)	4 (1.9%)	5 (1.1%)	5 (2.9%)	16 (1.6%)
Cardiac arrest	1 (0.6%)	2 (1.0%)	5 (1.1%)	0 (0%)	8 (0.8%)
Myocardial infarction	0 (0%)	2 (1.0%)	3 (0.7%)	2 (1.2%)	7 (0.7%)
Congestive cardiac failure	2 (1.3%)	1 (0.5%)	1 (0.2%)	2 (1.2%)	6 (0.6%)
Carotid blowout	1 (0.6%)	0 (0%)	1 (0.2%)	1 (0.6%)	3 (0.3%)
Respiratory					
Pneumonia	8 (5.0%)	10 (4.8%)	19 (4.3%)	11 (6.4%)	48 (4.9%)
Pneumothorax	2 (1.3%)	1 (0.5%)	1 (0.2%)	0 (0%)	4 (0.4%)
Pulmonary embolism	0 (0%)	1 (0.5%)	1 (0.2%)	0 (0%)	2 (0.2%)
Unspecified respiratory failure	0 (0%)	0 (0%)	4 (0.9%)	0 (0%)	4 (0.4%)
Gastrointestinal					
Upper gastrointestinal bleed	2 (1.3%)	1 (0.5%)	0 (0%)	0 (0%)	3 (0.3%)
Pancreatitis	1 (0.6%)	0 (0%)	1 (0.2%)	1 (0.6%)	3 (0.3%)
Genitourinary					
Urinary retention	0 (0%)	4 (1.9%)	4 (0.9%)	1 (0.6%)	9 (0.9%)
Other					
Delirium	1 (0.6%)	0 (0%)	0 (0%)	1 (0.6%)	2 (0.2%)
Mortality					
30-day mortality	3 (1.9%)	3 (1.4%)	11 (2.5%)	1 (0.6%)	18 (1.8%)

varied significantly (site A: 10/160 [6.3%], site B: 19/208 [9.1%], site C: 48/439 [10.9%], site D: 48/172 [27.9%]; λ^2=47.7, p<0.0001). The 30-day mortality rate was similar at 0.6–2.5% (λ^2=2.7, p=0.42).

No length of hospital stay (LOS) data were available for site C, and date of discharge was missing for some patients at sites A and B. LOS data for site D, however, were complete. For sites A, B and D, there was little difference in LOS for procedures of minor and intermediate complexity (Table 3). For major operations, however, the mean LOS varied between the sites (site A: 24 days, site B: 27 days, site D: 15 days; p<0.0001).

Risk adjustment algorithm

The data for each site were analysed to compare the mean rates of observed complications against the spread of predicted probabilities of complications (Fig 2). For site D, the logistic regression score suggested an expected complication rate (adjusted for case mix) of 47% while the observed complication rate was 51%. Good calibration was confirmed with the Hosmer–Lemeshow test, which found no significant difference between predicted and observed complication rates for groups divided into deciles of risk (χ^2=6.514, df=7, p=0.48) (Table 4). The receiver operating characteristic curve (area under the curve [AUC]: 0.70) indicated that discrimination was poor to reasonable (Fig 3) and the Brier score was unacceptably high (0.35), implying poor accuracy.

The risk adjustment algorithm allowed relative morbidity rates for the four sites to be compared in the form of a funnel plot (Fig 4). All units were performing within the 95% bounds of expected morbidity rates. The Waterlow score

Table 3 Length of stay data by scale of surgery

Scale of surgery	Length of stay (days)											
	Site A				Site B				Site D			
	n	Mean	95% CI	Median	*n*	Mean	95% CI	Median	*n*	Mean	95% CI	Median
Minor (<1h)	45	1.4	1.0–1.7	1.0	40	2.6	1.0–4.2	1.0	27	2.6	1.0–4.1	1.0
Intermediate	54	7.0	1.7–12.2	3.0	82	7.6	5.2–10.0	4.0	31	4.6	3.8–5.4	4.0
Major (>6h +/- flap)	35	24.6	17.7–31.5	18.0	64	26.7	20.0–33.5	16.5	114	14.8	12.5–17.1	12.0

CI = confidence interval

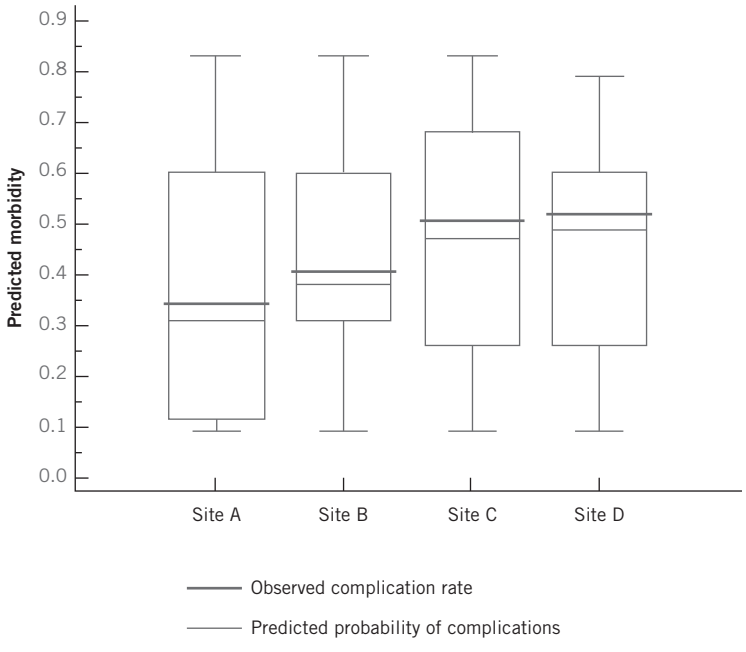

Figure 2 Box and whisker plot comparing the observed mean postoperative complication rates against the spread of predicted probabilities of complications

Table 4 Hosmer–Lemeshow goodness of fit for site D

Morbidity	Complications				No complications		
Risk band	Observed	Expected	(O-E)^2/E	Ratio	Observed	Expected	(O-E)^2/E
0–9% (*n*=14)	2	1.3	0.43	0.34	12	12.7	0.04
10–19% (*n*=13)	3	2.1	0.44	0.21	10	10.9	0.08
20–29% (*n*=17)	4	4.4	0.04	0.01	13	12.6	0.01
30–39% (*n*=10)	3	3.5	0.06	0.02	7	6.6	0.03
40–49% (*n*=4)	2	1.7	0.04	0.02	2	2.3	0.03
50–59% (*n*=52)	35	27.6	2.01	0.07	17	24.4	2.26
60–69% (*n*=49)	32	31.3	0.02	0.00	17	17.7	0.03
70–79% (*n*=10)	6	7.4	0.26	0.03	4	2.6	0.73
Total (*n*=169)*	**87 (51%)**	**79.1 (47%)**	**0.78**	**0.01**	**82**	**89.9**	**0.69**

*3 cases excluded because blood loss data were missing and risk score was not calculated

was tested as an additional predictor variable in the risk model. It correlated positively with adverse outcome (ANOVA, p=0.08), suggesting potential utility in the model.

Discussion

Given the heterogeneity of patient cohorts across different sites, in order to meaningfully compare surgical performance, it is essential to attempt to adjust for risk. Consequently, benchmarking surgical performance in other specialties has been characterised by the development of risk models such as the EuroScore,[6] the POSSUM (Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity)[7] and Glasgow aneurysm score.[8] In the UK, no such risk adjustment is implemented systemically in comparing the incidence of postoperative complications in head and neck oncology.

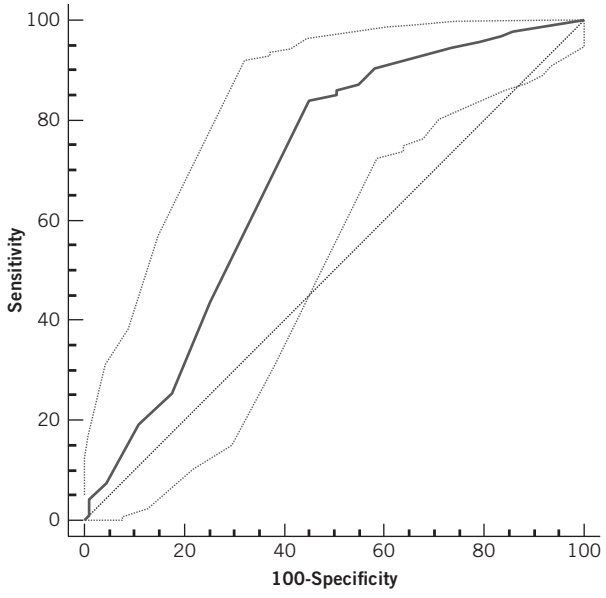

Figure 3 Receiver operating characteristic curve for site D data with 95% confidence interval

The need to pursue risk adjustment is seen in the paradox of complications and observed LOS data. Observed morbidity rates varied between the units (35–51%) but premorbid patient and tumour characteristics also varied significantly. Site D had the shortest mean LOS for patients undergoing major surgery (15 days) despite its cohort having higher frequencies of increased WHO performance status and microvascular reconstructions (Table 1). The site D complication rate was in keeping with the predicted probabilities of postoperative complications. Furthermore, using the risk algorithm, the morbidity rates for the different sites can be compared meaningfully on a funnel plot, which demonstrates risk adjusted performance within the 95% bounds of expected morbidity rates for all four sites (Fig 4).

Although Site D had the highest frequency of complications (51%), the 30-day mortality rate (0.6%) was lower than for the other sites. However, as death was a rare event, statistical significance for this finding is uncertain. Clavien–Dindo grades III and IV denote severe surgical or medical morbidity (Fig 1). Such complications were also seen more frequently in patients at site D.

Interpreting the paradox of more complications and shorter LOS requires risk adjustment as well as an assessment of possible differences in service provision. We believe that an increased level of involvement of senior anaesthetic staff, supporting the surgical team, allowed prompt recognition of clinical deterioration and a lower threshold for patient transfer to a high dependency unit (HDU), where complications are managed aggressively, and that this led to a significant reduction in LOS. More information about LOS in the HDU setting and comparison of staffing reviews recorded in the case notes around the time of deterioration is needed to test this assertion.

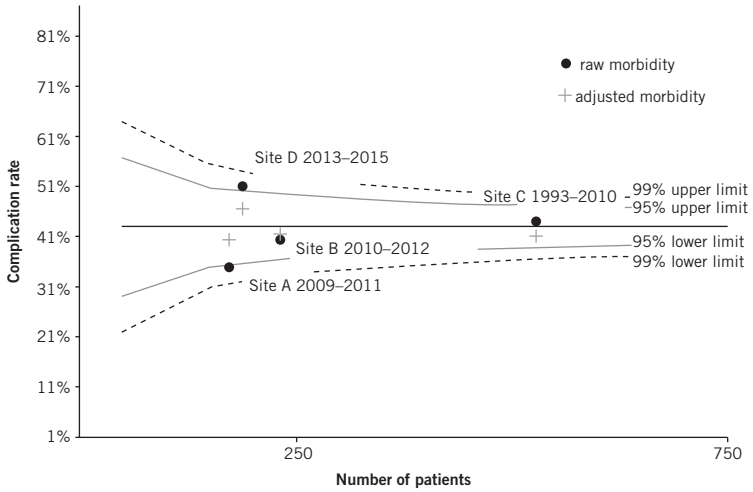

Figure 4 Funnel plot comparing frequency of complications between the four sites

Another untested assertion is that delayed discharge secondary to lack of family support and/or social services provision is conspicuous in its absence at site D. The catchment area covered by this hospital is larger than for the other sites. We speculate that the increased distance families must travel to visit the patient makes the social network more 'willing' to receive the patient back to the home setting when medically fit for discharge.

Case note audit of outcomes has been suggested as the gold standard for collecting data for the purposes of morbidity and mortality benchmarking. Ideally, internal verification of data accuracy would have strengthened the perceived integrity of the data. HES is an alternative source of data but the authors of a paper on the largest series of UK head and neck patients using HES data concede that over half of care episodes contain details that do not reflect data held by the multidisciplinary team and that complications are significantly underreported with the existing coding framework.[9]

Logistic regression analysis is a traditional statistical multivariate classifier technique that manages categorical, ordinal and continuous data well. The score algorithm in this study had internal validation (sites A–C), confirmed by an AUC of 0.76,[5] but for the external validation dataset (site D), discrimination worsened (AUC: 0.70), suggesting overfitting of the original model or a genuine increase in morbidity that was adjusted for inadequately by the current model. Modification of the score with more parameters that correlate with perioperative risk should bring further improvement.

To this end, our study also investigated the Waterlow score as a potential additional predictor variable for the logistic regression model. A high preoperative Waterlow score has previously been found to be associated with increased postoperative morbidity.[5] In terms of its utility for the risk algorithm, the Waterlow score looks promising as it appears to correlate with risk on initial testing. However, some discrimination will be lost through surgical case selection. As with the WHO performance status, if patients are very frail, they are less likely to be offered radical surgery. Patients with a performance status of 3 or 4 accounted for less than 5% of our cohort. Much bigger datasets are therefore required to capture those few who do have an operation in order to allow effective enumeration of added risk for the purposes of improved score discrimination.

Further effort is necessary to acquire new validation datasets in order to refine the algorithm before one can focus on the comparatively high morbidity rates at site D because the discriminatory power is not adequate (AUC: 0.70). Different forecasting classifier methods such as Bayesian models, decision trees or artificial neural networks should be tested as they have some pedigree in clinical contexts, especially in diagnostics and outcome prediction.[10] This will be the subject of a further publication.

Conclusions

Multicentre audit of early outcomes in head and neck surgery is in its infancy still. This validation study of a new case mix adjustment algorithm suggests further development is required and that quality of care cannot be distilled as a surgical (or surgeon) issue. Data mining techniques may yield improvement in score performance; Bayesian models, decision trees and artificial neural networks merit investigation although specialist knowledge is a prerequisite. Early morbidity after head and neck surgery is frequent, and contributes to mortality in around 1–2% of patients. Morbidity increases psychosocial distress, hospital bed occupancy rates and healthcare costs, and needs to be minimised where possible by highlighting and spreading good institutional practices.

References

1. Nouraei SA, Middleton SE, Hudovsky A *et al*. A national analysis of the outcome of major head and neck cancer surgery: implications for surgeon-level data publication. *Clin Otolaryngol* 2013; **38**: 502–511.
2. Bilimoria KY, Liu Y, Paruch JL *et al*. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg* 2013; **217**: 833–842.
3. Tighe D, Sassoon I, Kwok A, McGurk M. Is benchmarking possible in audit of early outcomes after operations for head and neck cancer? *Br J Oral Maxillofac Surg* 2014; **52**: 913–921.
4. Dindo D, Demartines N, Clavien PA. Classification of surgical complications. *Ann Surg* 2004; **240**: 205–213.
5. Thorn CC, Smith M, Aziz O, Holme TC. The Waterlow score for risk assessment in surgical patients. *Ann R Coll Surg Engl* 2013; **95**: 52–56.
6. Nashef SA, Roques F, Michel P *et al*. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999; **16**: 9–13.
7. Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical audit. *Br J Surg* 1991; **78**: 355–360.
8. Samy AK, Murray G, MacBain G. Glasgow aneurysm score. *Cardiovasc Surg* 1994; **2**: 41–44.
9. Nouraei SA, Hudovsky A, Frampton AE *et al*. A study of clinical coding accuracy in surgery: implications for the use of administrative big data for outcomes management. *Ann Surg* 2015; **261**: 1,096–1,107.
10. Barbini E, Cevenini G, Scolletta S *et al*. A comparative analysis of predictive models of morbidity in intensive care unit after cardiac surgery – Part I: model planning. *BMC Med Inform Decis Mak* 2007; **7**: 35.