Available online at www.sciencedirect.com

ScienceDirect

British Journal of Oral and Maxillofacial Surgery 57 (2019) 873–879

BRITISH Journal of Oral and Maxillofacial Surgery

www.bjoms.com

Validating a risk stratification tool for audit of early outcome after operations for squamous cell carcinoma of the head and neck

D. Tighe[a,*], A.J. Thomas[b], A. Hills[a], R. Quadros[a]

[a] *William Harvey Hospital, EKHUFT, Ashford, KENT TN24 0LZ*
[b] *School of Computing, Engineering and Mathematics, University of Brighton, East Sussex, UK*

Available online 26 July 2019

Abstract

The aim of this study was to validate a case-mix adjustment tool (neural network) for the audit of postoperative outcomes. We tested its calibration and discrimination on two unseen groups of patients being treated for squamous cell carcinoma (SCC) of the head and neck and compared observed complication rates with predicted rates. A total of 196 patients who were treated at two UK NHS institutions between 2016 and 2018 were audited. Preoperative data pertaining to risk (T classification, complexity of operation, and "high-risk" status) were collected, together with data on postoperative complications. Diagnostic test statistics and receiver operating curves (ROC) were used to test the performance of the tool. The score was well calibrated (predicted and observed complication rates both 43%), but discrimination suggested only fair accuracy (ROC 0.66 - 0.68). Adjustment of case mix for the audit of postoperative complications is difficult, although our model suggests that departmental audit is possible, and its accuracy is equivalent to that of other national audits. Further work may elucidate key variables that have not yet been assessed.

Keywords: Audit; Outcomes; Complications; Head & neck squamous cell carcinoma

Introduction

The publication of national audits of surgical outcomes is established in the UK, and datasets on consultants' surgical outcomes routinely report unplanned returns to theatre and deaths within 30 days, usually together with specialty-specific complications. Unlike the other surgical specialties, however, early outcome data on patients operated on for squamous cell carcinoma (SCC) of the head and neck is not currently adjusted for case mix. At present, in national surgical audits, the performance of a unit is measured by mortality, readmission rates, or specific complications. This process is contentious, and the use of risk-adjusted data attempts to mitigate the main concern that surgeons who operate on high-risk cases will be unfairly represented with higher rates of morbidity. The focus on patients' outcomes as part of a quality assurance framework has recently risen on profile within the British Association of Oral and Maxillofacial Surgery.

At present, the effective adjustment for case mix in outcome data in head and neck oncology remains an unmet need. A large national dataset (more than 10 000 patient-care episodes for cancer of the head and neck) based on data derived from hospital episode statistics (HES) has recently attempted to address this issue,[1] and the Association of Surgeons of North America has done the same with the National Surgical Quality Improvement Program (NSQUIP) surgical risk calculator.[2] Both projects, however, have their limitations. Both use post-hoc clinical coding data, and accurate coding of events by clinical coders is not perfect. Logistic

[*] Corresponding author.
E-mail addresses: David.tighe@nhs.net (D. Tighe), alan.j.thomas@gmail.com (A.J. Thomas), Ajh502@gmail.com (A. Hills), r.quadros@nhs.net (R. Quadros).

https://doi.org/10.1016/j.bjoms.2019.07.008
0266-4356/© 2019 Published by Elsevier Ltd on behalf of The British Association of Oral and Maxillofacial Surgeons.

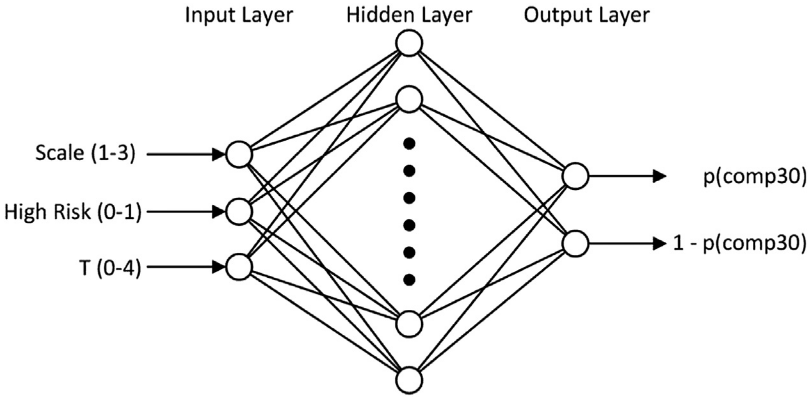

Fig. 1. Diagram of artificial neural network (ANN).

regression, which is currently the method used by these two large projects, can fail when data are not linear, or when categorical relations predominate. The Surgical Outcome Risk Tool (SORT), which was developed by the National Confidential Enquiry into Patient Outcome and Death, calculates the risk of early death after non-cardiac surgery, but as this is a rare event after major operations on the head and neck, it is not suitable to be used to compare the performance of surgical units. When negotiating contracts and payment, commissioners in the UK will increasingly seek evidence of high-quality care, which requires standardisation.

The models used must be transparent, and surgeons should be informed about their limitations. Different statistical methods are used to analyse large datasets that have accrued from many years of national audit data which, historically, have not been complete enough in audits of head and neck oncology. Logistic regression, Bayesian classifiers, and decision trees have the virtue of being transparent and easily conceptualised, and logistic regression techniques predominate in current national audits.

Artificial neural networks (ANN) are self-learning computer programs that have an increasing role in "big data" problems, and have been applied to the calculation of clinical risk, diagnostics, and prognostics. Neural networks are more tolerant of "missingness" and the non-linear nature of clinical data, but are criticised for the "black-box" nature of the processes, which remain hidden. For a deeper discussion of the strengths and weaknesses of these data-mining techniques the reader is directed to specialist resources.[3]

We have developed a neural network (Fig. 1) that allows for the adequate adjustment of case mix when auditing complications within 30 days of operations for SCC of the head and neck. It has been internally validated,[4] having been developed from the clinical audit data of four NHS cancer networks that treat these patients in the south-east of England (Sites 1 – 4), and has out-performed alternative methods of risk stratification. In this study we sought to validate its utility on two external surgical units (Sites 5 and 6) to test the potential for "over-fitting". This is when a model maps the outcomes on a dataset on which it is based too closely, and the performance is not maintained on new "unseen" groups.

We chose to report hospital outcome data rather than consultant outcome data because operations on the head and neck are often done by consultants who work in teams across multiple departments including wards and intensive care. The reporting of hospital outcomes as the Head and Neck Audit (HANA) should encourage a response by the hospital that reinforces local clinical governance.

Methods

Two datasets of 24 months of clinical activity at each unit were completed as a mix of retrospective and prospective audits (Sites 5 and 6). The lead author (Sites 5 and 6) with 2 coauthors (Site 6) collaborated to collect pertinent preoperative and postoperative data from the case notes (electronic and paper). Prospective episodes of case were followed after histopathological confirmation of SCC of the head and neck, and surgical treatment had been agreed by the multidisciplinary team. Retrospective cases were identified by clinical coders and validated by the team dataset (Site 5) and previous multidisciplinary outputs (Site 6). Uncertainty about complications was resolved by the lead author who had previously collected data at four hospitals. No further validation processes were employed. Data were collected by the lead author with the permission of the treating consultants. All operations were done under general anaesthetic by maxillofacial or ear, nose, and throat surgeons, or both, and were all done using conventional surgical techniques (excluding laser cases). The audit datasets were registered with the respective clinical audit departments, and data were presented at hospital audit meetings. Ethics approval was sought for this validation stage of the clinical audit because the results could potentially be generalised. Anonymised data were used by collaborators independently.

The structure of the neural network was T classification (integer between 0 and 4), high-risk (Boolean), and scale of surgery (integer between 1 and 3). The neural network required tumours to be classified according to the American Joint Committee on Cancer (AJCC) TNM classification (V7); data on operations, which included the complexity, were derived from the BUPA severity of surgery index (minor = less than 1 hour; intermediate = less than 6 hours; and major or major complex = free tissue transfer or 6 hours), and a derived variable, the high-risk status, which used Operation Procedure Codes version 4 (OPC4) included any procedure that required mucosal closure in association with a neck dissection that could lead to the escape of saliva (such as mandibulectomy, glossectomy, floor or mouth excisions, and pharyngectomy with or without laryngectomy). The area under the curve (AUC) of the output probability of a complication within 30 days (0 - 1) was used to select the best one ("champion").

Receiver operating characteristic curves are a measure of score accuracy (discrimination). The plot compares the sensitivity against the false positive rate (1-specificity) of the model at different probability thresholds. The more curved the plot, the greater the AUC, which will approach 1 if it is perfect. An AUC of 0.5 shows performance that is no better than random choice, and will result in the oblique line running across the graph.

This neural network had 94 hidden nodes. There is no way to ascertain the best number of hidden nodes as it depends on the many factors in the problem domain. In our experiments, we adopted a commonly-used method of training networks with different structures and different numbers of hidden nodes where, in each case, the training (70%) and test data (30%) were randomly selected from the full dataset. This resulted in a plethora of networks from which the network with the best AUC response to unseen data was selected. The Matlab R2014b classifier network "patternnet" was used with the Scaled Conjugate Gradient training algorithm without a validation dataset, and with cross-entropy as the error function. Training was stopped when the minimum gradient reached 0.06, or the number of training epochs reached 200. The "champion" network with 94 hidden nodes gave the best response to the test data with an AUC of 0.85 and a misclassification rate of 22%.

Complications were recorded according to the Clavien-Dindo classification,[5] but for consistency, all complications (whatever the severity) were considered, in keeping with the development of the model. If pertinent data points were missing from a case (as selected by the model) it was excluded. We tested the calibration and discrimination of the model with receiver operating curves (ROC), and reported the sensitivity and specificity. Finally, for meaningful comparison of performance between surgical units, we used a funnel plot to report both the raw data on morbidity and those adjusted for case mix.

We present the ability of the neural network to predict outcome accurately in unseen groups, and compare the performance of the units.

Results

The dataset included 83 patients at Site 5 and 112 patients at Site 6. All patients had index operations with curative intent for squamous cell carcinoma (SCC) of the head and neck, and all were done under general anaesthesia. The age range was 19 – 97 years, (mean age 65 years at Site 5 and 69 years at Site 6). A total of 126 (65%) were men. Smoking status, alcohol intake, performance status, previous radiotherapy to the operative site, tumour stage, nodal stage, complexity of surgery, and status of high-risk procedure, were significantly different, which confirmed a variation in case mix ($p<0.05$) when compared with the populations treated at the initial four units (from which the neural network was developed), and the need for the stratification of risk (Table 1).

For the 196 care episodes analysed, postoperative complications within 30 days were common (35% at Site 5, 30% at Site 6) but did not differ significantly ($\chi^2 = 0.53$ (DF = 1), p = 0.5).

The incidence of recorded complications at these centres is presented with data from the other units (Table 2). Complete flap failure rates were similar: Site 1: 2/39 (5%); Site 2: 3/47 (6%); Site 3: 8/86 (9%); Site 4: 7/106 (7%); Site 5: 0/32; and Site 6: 0/41 ($\chi^2 = 6$ (DF = 5), p = 0.3), but severe complications (Clavien-Dindo grade 3 or more) varied significantly (χ^2 47.7 (DF = 5), p = <0.0001): Site 1: 10/160 (6%); Site 2: 20/208 (9%): Site 3: 48/439 (11%); Site 4: 52/172 (30%); Site 5: 18/84 (21%); and Site 6: 16/112 (14%). Mortality within 30 days was also similar at 0.6% – 2% (χ^2 5.4 (DF = 5) p = 0.4).

Discrimination of the ANN was good on internal validation (AUC 0.85).[4] On testing its calibration and discrimination on the new datasets, the overall calibration was remarkably close (both the observed and predicted complication rates were 43%) but there was a mismatch in the calibration plot (Fig. 2). Discrimination was fair, the AUC was 0.68, and specificity was 0.73, and sensitivity 0.51 (Fig. 3).

The role of the ANN in the adjustment of case mix and operations is shown on a funnel plot (Fig. 4) that compares observed complication rates with risk-adjusted rates. When datasets are small, rare events can skew the data, so the confidence intervals, which initially are wide, will taper as the size of the hospital's audit group grows. All units performed within the 95% bounds of expected morbidity.

Discussion

The benchmarking of surgical performance in other specialties has been characterised by the development of logistic regression risk models such as the EuroSCORE,[6] Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (POSSUM) score,[7] and Glasgow aneurysm score.[8] Currently in the UK, no such risk

Table 1
Patients' characteristics.

	Hospital						Totals
	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	
Mean (range) age (years)	66 (63.6-68.1)	67 (64.9-68.6)	61 (60.1-62.6)	66 (64.4-68.1)	62(59.7-65.2)	69 (66.6 – 71.4)	
Sex:							
Male	109	147	323	122	61	65	
Female	51	56	198	53	22	47	
Total	160	203	521	175	83	112	1254
Alcohol:							
1	69	99	54	56	28	28	
2	31	54	93	48	20	46	
3	7	14	76	38	23	17	
4	31	32	105	20	5	8	
5	5	10	54	13	7	6	
Total	143	209	382	175	83	105	1097
Smoking:							
Current	56	83	110	66	16	46	
Ex-current or non-smoker	88	117	384	109	67	59	
Total	144	200	494	175	83	105	1201
ACE 27:							
0	62	7	239	39	35	29	
1	56	123	215	97	35	51	
2	35	67	48	32	12	20	
3	1	5	3	7	1	7	
Total	154	202	505	175	83	107	1226
Performance status:							
0	25	14	–	102	28	47	
1	90	119	–	36	47	35	
2	29	54	–	23	4	18	
3	8	13	–	14	2	5	
Total	152	200	–	175	81	105	713
Flap:							
0	118	156	353	69	51	79	
1	39	47	86	106	32	31	
Total	157	203	439	175	83	110	1167
Tracheostomy:							
0	124	145	229	128	48	87	
1	32	56	178	47	35	20	
2	156	201	407	175	83	107	1129
Scale of operation							
1	41	35	72	27	3	36	
2	65	96	128	32	28	31	
3	51	72	242	116	52	45	
Total	157	203	442	175	83	112	1172
High risk:							
0	96	132	208	101	29	64	
1	61	71	231	74	54	48	
Total	157	203	439	175	83	112	1169
T classification:							
0	26	50	25	30	14	13	
1	57	55	124	35	16	36	
2	32	33	149	35	19	29	
3	9	12	76	10	7	2	
4	30	47	122	63	27	29	
Total	154	197	496	173	83	109	1212
N classification:							
0	88	108	315	98	42	71	
1	19	24	96	20	16	7	
2a	14	14	26	6	0	1	
2b	27	30	26	37	18	26	
2c	5	9	15	6	4	1	
3	1	7	11	5	0	0	
Total	154	192	492	172	80	106	1196

Table 1 (*Continued*)

	Hospital						Totals
	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	
Previous radiotherapy:							
0	147	159	–	148	50	99	
1	13	44	–	27	33	13	
Total	160	203	–	175	83	112	733
Previous operation:							
0	134	166	–	138	52	79	
1	26	37	–	37	31	33	
Total	160	203	–	175	83	112	733

Table 2
Complications at 30 days. Data are number (%).

Complication	Site 1 (n = 160)	Site 2 (n = 208)	Site 3 (n = 428)	Site 4 (n = 171)	Site 5 (n = 84)	Site 6 (n = 112)	Total (n = 1163)	Overall (%)
No. of flaps lost/No. of flaps	2/39 (5)	3/47 (6)	8/93 (9)	7/106 (7)	0/39 (0)	0/41	20/365	5
Partial loss of flap	–	–	10/93 (11)	2/106 (2)	1/39 (1)	0/41	11	1
Haematoma	4 (2)	4 (2)	9 (2)	11(6)	12 (14)	5 (4)	45	4
Wound dehiscence	11 (7)	8 (4)	21 (4)	15 (8)	4 (5)	2 (2)	61	5
Orocutaneous fistula	1 (0.5)	0 (0)	6 (1)	4 (2)	6 (7)	5 (4)	23	2
Wound infection	9 (6)	7 (4)	8 (1)	18 (10)	7(8)	3 (3)	59	5
Neck abscess	–	–	3 (0.5)	–	1 (0.5)	–	4	0
Chyle leak	1 (0.5)	3 (2)	3 (0.5)	2 (1)	–	1 (1)	10	1
Carotid blowout	1 (0.5)	0 (0)	1 (0)	1 (0.5)	–	–	3	0
Atrial fibrillation	2 (1)	4 (2)	5 (1)	5 (3)	–	1 (1)	17	1
Myocardial infarction	–	2 (1)	3 (0.5)	2 (1)	–	3 (3)	10	1
Cardiac arrest	1 (0.5)	2 (1)	5 (1)	–	–	–	8	1
Congestive cardiac failure	2 (1)	1 (0.5)	1 (0)	2 (2)	–	1 (1)	7	1
Pulmonary embolism	–	1 (0.5)	1 (0)	–		–	2	0
Pneumonia	8 (5)	10 (5)	19 (4)	11 (6)	4 (5)	4 (4)	56	5
Urinary retention	–	4 (2)	4 (1)	1 (0.5)	–	–	9	1
Delirium	1 (0.5)	0 (0)	0 (0)	1 (0.5)	2 (3)	1 (1)	3	0
Slow wean	–	–	–	–	–	2 (2)	–	0
30-day mortality	3 (1)	3(1)	11 (2)	1 (0.5)	–	1 (1)	19	2

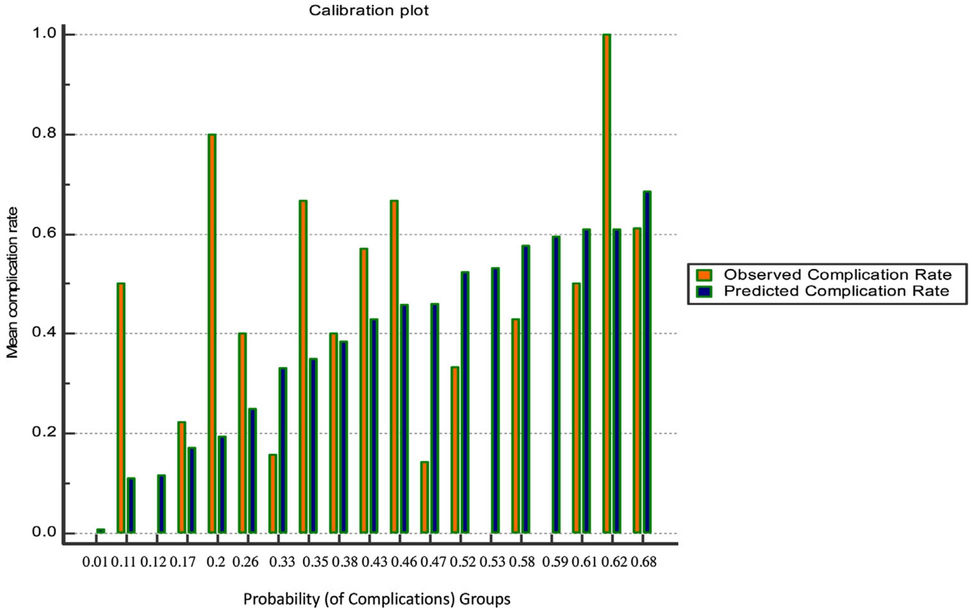

Fig. 2. Observed and predicted complications.

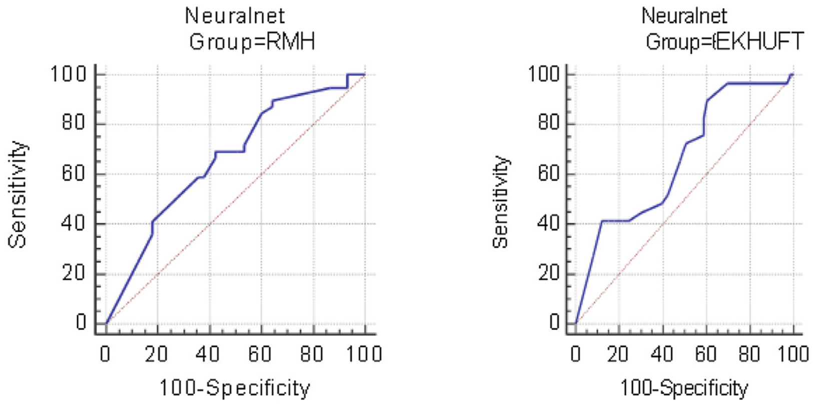

Fig. 3. Comparison of receiver operating characteristic (ROC) curves (RMH = Royal Marsden Hospital; EKHUFT = East Kent Hospitals University Foundation Trust).

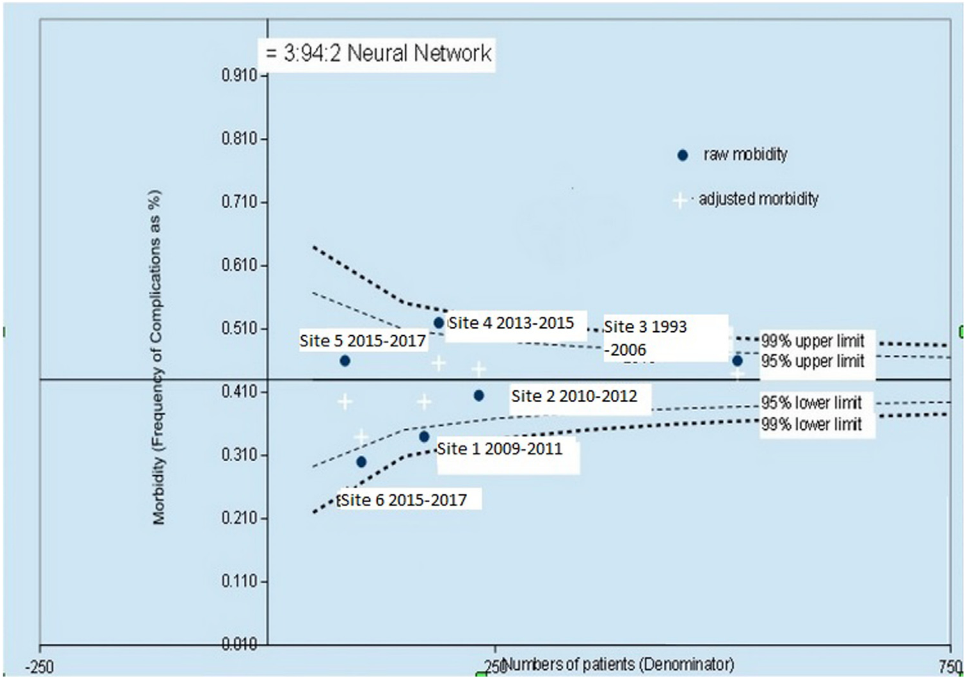

Fig. 4. Funnel plot of observed and adjusted incidences of 30-day complications.

adjustment has been implemented to compare the incidence of postoperative complications in head and neck oncology.

For the comparison of surgical performance to be meaningful, it is essential to adjust for risk given the heterogeneity of case mix and surgical management in different hospitals. Observed morbidity rates varied between all units (30% - 51%), but premorbid characteristics of the patients and tumours also varied significantly (Table 1). Whilst complications were common, decision-making and multidisciplinary interventions seemed to mitigate them. Case-mix adjustment for morbidity showed that all the centres performed within the 95% bounds on the funnel plot.

The ANN was calibrated well for complications, which suggested its potential as an audit tool, but discrimination was much less effective when applied to the new datasets. It is likely that factors that contribute to risk have not yet been adequately captured by the model. Another explanation is the large difference in the performance of the two new units from those previously audited, and the neural network score needs to be updated to reflect this.

The performance of the model is likely to improve if a narrower band of operative risk is selected, perhaps by the identification of indicator procedures or outputs such as free-flap cases only, or pneumonia or wound infections, but this will be the focus of further work.

In the future we will compare the discrimination and calibration of the score against other benchmarking methods currently in use. A large national study of morbidity after major operations on the head and neck was recently published by Nouraei et al,[2] though the complication rate seemed somewhat low at 31% when overall morbidity across the six units we have audited so far was 43%. A subsequent study by the same team acknowledged the limitations of relying on data that had been collected by clinical coders.[9] The audit of outcome using case notes has been suggested as the gold standard method to measure morbidity and mortality, as we have done in our study.

Despite promising results of the internal validation of the neural network (AUC 0.85) the model could be "over-fitted". While the calibration plot suggested that it predicted poorly in low-risk cases, the overall rate of predicted complications was remarkably accurate, which suggests that the model has some potential in disseminated audit. Using the same methods, it should be possible to update (or mature) the network with new datasets, and this will be the focus of further work.

Conflict of interest

The main author is a consultant at East Kent University Hospitals NHS Foundation Trust. The lead author is employed at Site 6.

Ethics statement/confirmation of patients' permission

This audit was registered at each audit department of the respective hospitals. As the culmination of this study has the potential to be generalised, ethics approval was obtained from the EKHUFT, Grey Area Project Group Research and Innovation Department. Patients' permission was not required

Acknowledgement

We thank our consultant surgical colleagues for the care on which this research is based.

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**:502–11.
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**:833–42.
3. Ray S. Essentials of machine learning algorithms (with Python and R codes). Analytics Vidhya, 9 September 2017. Available from URL: http://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/ (last accessed 18 March 2019).
4. Tighe DF, Thomas AJ, Sassoon I, et al. Developing a risk stratification tool for audit of outcome after surgery for head and neck squamous cell carcinoma. *Head Neck* 2017;**39**:1357–63.
5. Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann Surg* 2004;**240**:205–13.
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**:355–60.
8. Samy AK, Murray G, MacBain G. Glasgow aneurysm score. *Cardiovasc Surg* 1994;**2**:41–4.
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**:1096–107.