Available online at www.sciencedirect.com

ScienceDirect

British Journal of Oral and Maxillofacial Surgery 57 (2019) 771–777

BRITISH Journal of Oral and Maxillofacial Surgery

www.bjoms.com

Machine learning methods applied to audit of surgical outcomes after treatment for cancer of the head and neck

D. Tighe[a,*], T. Lewis-Morris[a], A. Freitas[b]

[a] *EKHUFT, United Kingdom*

[b] *School of Computing, University of Kent, CT2 7NF, United Kingdom*

Accepted 14 May 2019

Available online 26 July 2019

Abstract

Most surgical specialties have attempted to address concerns about unfair comparison of outcomes by "risk-adjusting" data to benchmark specialty-specific outcomes that are indicative of the quality of care. We are building on previous work in head and neck surgery to address the current need for a robust validated means of risk adjustment. A dataset of care episodes, which were recorded as a clinical audit of complications after operations for squamous cell carcinoma (SCC) of the head and neck (n = 1254), was analysed with the Waikarto Environment for Knowledge Analysis (WEKA) machine learning tool. This produced 4 classification models that could predict complications using data on the preoperative demographics of the patients, operation, functional status, and tumour stage. Three of them performed acceptably: one that predicted "any complication" within 30 days (area under the receiver operating characteristic curve (AUROC) 0.72), one that predicted severe complications (Clavien-Dindo grade 3 or above) within 30 days (AUROC 0.70), and one that predicted a prolonged duration of hospital stay of more than 15 days, (AUROC 0.81). The final model, which was developed on a subgroup of patients who had free tissue transfer (n = 443), performed poorly (AUROC 0.59). Subspecialty groups within oral and maxillofacial surgery are seeking metrics that will allow a meaningful comparison of the quality of care delivered by surgical units in the UK. For these metrics to be effective they must show variation between units and be amendable to change by service personnel. Published baseline data must also be available. They should be modelled effectively so that meaningful comparison, which takes account of variations in the complexity of the patients' needs or care, is possible.

Keywords: Oncology Head & Neck Surgery; HNSCC; Audit; Complications

Introduction

The benchmarking of care in surgical audit requires the definition of outcomes that will enable meaningful comparison of different treatment centres. There is growing consensus that mortality, which is low (0.5% - 2%) in our specialty, is not likely to be a helpful indicator of the quality of surgical care. Morbidity, however, varies between surgical units, and rates can be up to 70% from those that report series of patients who have major operations on the head and neck with free tissue transfer.[1]

The Clavien-Dindo classification system was developed to mitigate some of the subjectivity involved in the recognition of a complication, and to improve consistency.[2] It has been well reported in surgical publications and has been applied to head and neck surgery by many authors.[1,3–6] It allows the clear recording of surgical and systemic complications and facilitates comparative audit. The authors state that complications are "any deviation from the ideal postoperative course that is not inherent in the procedure and does not comprise a failure to cure", and they further underscore their belief that

[*] Corresponding author.

E-mail addresses: David.tighe@nhs.net (D. Tighe), lewis-morris@nhs.net (T. Lewis-Morris), a.a.freitas@kent.ac.uk (A. Freitas).

https://doi.org/10.1016/j.bjoms.2019.05.026

0266-4356/

"the incidence of postoperative complications [should still be] the most frequently used surrogate marker of quality in surgery".[7]

To evaluate more broadly the quality of care after head and neck surgery, additional metrics (such as duration of stay, use of blood products, and adequacy of pathology reports) have been created and tested, in addition to the items contained in the Clavien–Dindo classification, namely return to theatre and mortality.[8]

To be effective, quality metrics must show variation between units and be amendable to change by service personnel. Published baseline data must also be available. We argue that a fourth criterion should be added: that they can be effectively modelled statistically so that meaningful comparison, which takes account of variations in the complexity of the patients' needs or care, is possible.

For the purpose of comparative audit of the quality of care, we have tried to find out if all complications and severe complications (Clavien-Dindo level 3 and above) are valid outcomes to capture, and if duration of hospital stay is a proxy indicator of the quality of care. Finally, we looked into whether it is preferable to model just the subset that has immediate reconstruction with free tissue transfer.

Methods

A total of 1254 admissions for operations (with curative intent) under general anaesthesia for squamous cell carcinoma (SCC) of the head and neck were analysed. This comprised the datasets of 6 units: site 1 (n = 160), site 2 (n = 203), site 3 (n = 521), site 4 (n = 175), site 5 (n = 83), and site 6 (n = 112). A total of 444/1254 patients (35%) had free tissue transfer. Data pertaining to duration of hospital stay (not included in the dataset from site 3) were available on 733/1254 (58%).

Data were preprocessed by the lead author, and experiments done using 3 methods in the "Classify" panel of the Waikarto Environment for Knowledge Analysis (WEKA) data mining (or machine learning) tool (Auto-WEKA),[8] the J48 decision tree algorithm, and the random forest algorithm. Auto-WEKA is an advanced machine learning method that automatically selects the best classification algorithm and its best configuration (hyperparameter settings) for an input dataset, by doing a systematic search of many different types of algorithms and their configurations that are available in WEKA.

Both the J48 and the random forest algorithms are included in the algorithms considered by Auto-WEKA during its search, but we also used them separately for the following reasons. First, J48 learns an interpretable model in the form of a decision tree and highlights the most relevant attributes for classification. Although J48 does not achieve a high predictive performance in the analysed datasets, some parts of a decision tree can be substantially more accurate than the tree as a whole. Some examples of the accurate and interpretable rules that were extracted from a J48 decision tree will be shown later. We used the random forest algorithm because it is one of the state-of-the-art algorithms regarding predictive performance.

As a measure of predictive performance, the results are reported using area under the receiver operating characteristic (AUROC) curves. Receiver operating characteristic (ROC) 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 area under the curve (AUC), which will approach one if it is perfect. An AUC of 0.5 shows performance that is no better than random choice, and will result in an oblique line running across the graph.

Decision tree J48 is the implementation of an open-source algorithm ID3 (Iterative Dichotomiser-3) developed by the WEKA project team. It shows "information gain" graphically in a hierarchical structure, which is easily understood, to show the importance of an attribute in a dataset. Random forests are an ensemble learning method for classification, regression, and other tasks that operate by constructing multiple decision trees.

Auto-WEKA is a non-deterministic method (its results depend on a random seed that is used to initialise the program) and it requires the user to specify a "time limit" – that is, the amount of time the algorithm is allowed to search for the best algorithm and its configuration. We ran Auto-WEKA 10 times, and varied the random seed and the time limit (5 hours in 5 runs, and 20 hours in the other 5 runs). As the best result, we selected the one with the highest AUROC value according to an internal measure of validation. This internal AUROC measure, however, tends to be over optimistic so, to find one that was more realistic, we ran the best algorithm and configuration recommended by Auto-WEKA again, and this time did a 10-fold cross-validation without using Auto-WEKA. Our results refer to this 10-fold cross-validation. To predict the class variable of duration of stay, we tested cut-offs based on previous work (in press, British Journal of Oral and Maxillofacial Surgery), namely more than 15 days and more than 20 days. The AUROC statistics are shown with the ROC diagrams.

Results

Successful analysis of the dataset (with attributes listed in Table 1) allowed us to model the following class variables using the J48 decision tree algorithm, the random forest algorithm, and the Auto-WEKA method: complications with 30 days; complications within 30 days that were Clavien Dindo grade 3 or higher; duration of hospital stay; and complications within 30 days in the immediate free tissue transfer group.

The most accurate model that predicted a "complication within 30 days" was produced by Auto-WEKA, which selected a "bagging of J48 decision trees", that is, a collection (ensemble) of decision trees with an

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

AUROC of 0.730. It is, however, hard to interpret the entire collection of many decision trees so, for the purposes of interpretation, we focused on the single decision tree produced by the J48 algorithm that was used separately from Auto-WEKA. This model still had an adequate AUROC value of 0.705, and it can readily be embedded within a database for immediate computation of predicted risk.

The J48 decision tree (Fig. 1) shows the most important predictors of risk of complications in a graphical and hierarchical form, in which the most relevant predictors are generally closer to the root node (attribute “Tracheostomy”) of the decision tree. Note that as the root node, “Tracheostomy” is the most relevant attribute, since the classifications of all patients use this value. In addition, when “Tracheostomy” takes a value of zero (“no”), the most relevant attributes for classification are “Flap” and “Scale of surgery”; whilst when “Tracheostomy” takes a value of one (“yes”), the most relevant attributes are “Duplicate?” and “High risk”.

As noted earlier, the AUROC value was higher with Auto-WEKA, though it was harder to interpret the model, unlike that of the J48 decision tree, which could be interpreted directly.

An easily interpretable classification rule that can be derived from this decision tree model states that:

“IF Scale of Surgery = Minor (1) and Tracheostomy = Absent (0) and Free Flap = Absent (0) (“no”), THEN the patient should have no complication”.

This was 85% accurate, and it correctly classified 204/240 patients in our dataset who satisfied all the conditions in the “IF” part of the rule. A threshold complication rate of 15% should therefore be tolerated in this “low-risk” group.

The most accurate model that predicted “Complications within 30 days graded as Clavien–Dindo 3 or greater” was produced by the random forest algorithm that was used separately from Auto-WEKA. It was acceptable with an AUROC value of 0.70. This algorithm, however, showed a substantial shortfall. The model consistently predicted “no complication” and predicted a complication in only about 1% of the cases (11/1322). This was despite the fact that the observed complication rate across all sites was 14.4% – though the range varied considerably (site 1: 6%; site 2: 9%; site 3: 11%; site 4: 30%; site 5: 21%; and site 6: 14%) (Table 2).

```
Tracheostomy = 0
| Flap = 0
| | ScaleofSurgery = 1: 0 (240.38/36.0)
| | ScaleofSurgery = 2
| | | Age <= 85: 0 (389.99/101.38)
| | | Age > 85: 1 (31.3/12.3)
| | ScaleofSurgery = 3
| | | Age <= 61: 1 (23.07/7.69)
| | | Age > 61: 0 (27.46/5.69)
| | ScaleofSurgery = 4: 0 (0.0)
| Flap = 1
| | Bilateral Neck = 0
| | | CVS = 0
| | | | ACE_27 = 0
| | | | | T = 0: 0 (4.36)
| | | | | T = 1: 0 (8.19/1.0)
| | | | | T = 2: 0 (20.91/6.9)
| | | | | T = 3: 0 (6.87/2.07)
| | | | | T = 4: 1 (23.17/7.46)
| | | | ACE_27 = 1: 0 (41.5/14.88)
| | | | ACE_27 = 2: 1 (7.36/1.22)
| | | | ACE_27 = 3: 0 (0.0)
| | | CVS = 1: 1 (78.06/26.17)
| | Bilateral Neck = 1: 1 (10.71/2.02)
Tracheostomy = 1
| duplicate? = 0
| | High risk = 0
| | | Smoking = 1: 0 (16.25/5.31)
| | | Smoking = 2: 1 (51.75/20.13)
| | High risk = 1: 1 (310.13/94.11)
| duplicate? = 1: 0 (30.55/14.31)
Number of Leaves : 20
Size of the tree : 32
```

Fig. 1. Decision tree for “all complications” within 30 days.

Classification models were built to predict the duration of stay using 15-day and 20-day cut-offs. The most accurate models for both were produced by Auto-WEKA (in both cases with a good AUROC value of more than 0.8). Although the J48 algorithm that was used separately from Auto-WEKA obtained lower accuracies overall (AUROC 0.73 and 0.68 for

Table 2
Complications by hospital site. Data are number (%).

Complications	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. flap loss/flap no.	2/39 (5)	3/47 (6)	8/93 (9)	7 /106 (7)	0 /39 (0)	0/41 (0)	20/365	5
Partial loss of flap	-	-	10 /93 (11)	2/106 (2)	1/39 (1)	0/41(0)	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
Clavien-Dindo >3	10 (6)	20 (9)	48 (11)	52 (30)	18 (21)	16 (14)	-	14
30-day mortality	3 (1)	3(1)	11 (2)	1 (0.5)	-	1 (1)	19	2

15-day and 20-day cut-offs, respectively), some parts of the generated decision tree were highly accurate, and could easily be interpreted. In particular, a simple and highly reproducible rule that was produced by J48 predicted more than 20 days of stay, namely:

IF (Tracheostomy = 0) THEN (Inpatient days < 20 days).

This had an accuracy of 90.2%, and correctly classified 515/571 patients who satisfied the condition in the "IF" part of the rule.

Finally, the most accurate model that predicted "Complications within 30 days in the immediate free tissue transfer group" was produced by Auto-WEKA, but it had a relatively low AUROC of 0.590. This level of accuracy is not acceptable as an algorithm for the purpose of risk adjustment in audit (note that an AUROC of 0.5 indicates predictive performance equivalent to a random classifier). The dataset for this group was much smaller (n = 444), and modelling may improve with more data if the performance of other contributing units is similar to that of the 6 units that have contributed data thus far.

A truncated list of complication rates at each site, including a Clavien-Dindo grade of more than 3, is shown in Table 3. Fig. 2 shows a composite graph of the AUROC values, and Fig. 3, the confusion matrices for the 4 models.

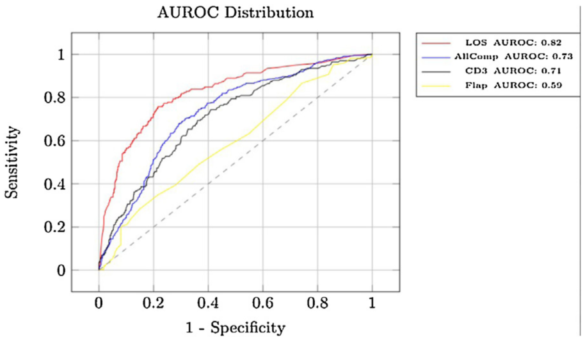


Fig. 2. Composite graph of AUROC values.

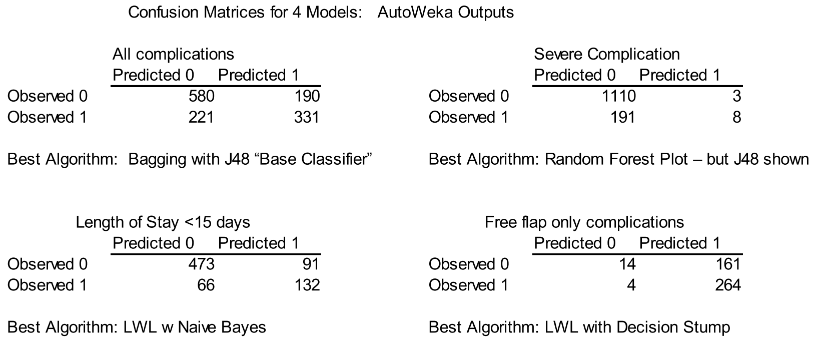
Discussion

Our results suggest that acceptable performance, for the purposes of risk adjustment, is found in the models that predict any complication in a period of 30 days after operation, and duration of hospital stay (using cut-offs of 15 and 20 days). There is some concern that the model used to predict severe complications (Clavien-Dindo 3 or more) did not discriminate sufficiently, as it tended to predict that all patients would not have severe complications, with few exceptions. Finally, the model that predicted any complications in a period of 30 days after free flap surgery is, at present, insufficient for the purpose of risk-adjustment in an audit.

The decision to report all complications may be considered unduly onerous and less relevant than the collection of data on severe complications which, in general, affect patients more and require more resources. However, donor-site complications, which are often "minor", may require multiple visits to dressing clinics and therefore merit capture. Duration of stay could be considered a good aggregate outcome measure, as it necessarily reflects severe complications, but it could also be affected by non-surgical issues, such as socioeconomic status, the status of carers (next of kin), and distance from home to the hospital, which is the subject of another publication. Another metric, time elapsed from operation to radiotherapy, could also be used to capture the quality of surgical care, but as it is also limited to a subset of patients who need radio-

Table 3
AUROC obtained by the 3 different classifications used in this work, for each dataset (defined by a given class variable).

Class variable and dataset	Best algorithm selected by Auto-WEKA	Random forest without using Auto-WEKA	J48 decision tree without using Auto-WEKA
Complications <30 days (dataset 1)	Bagging with J48: 0.730	0.719	0.705
Clavien-Dindo >=3, (dataset 2)	Random forest: 0.684	0.702	0.501
Inpatient days <15 days (dataset 3)	Locally weighted learning with Naïve Bayes: 0.824	0.814	0.731
Complications <30 days (free-flap only) (dataset 4)	Locally weighted learning with decision stump: 0.590	0.521	0.549



Confusion Matrices for 4 Models: AutoWeka Outputs

All complications

	Predicted 0	Predicted 1
Observed 0	580	190
Observed 1	221	331

Best Algorithm: Bagging with J48 “Base Classifier”

Severe Complication

	Predicted 0	Predicted 1
Observed 0	1110	3
Observed 1	191	8

Best Algorithm: Random Forest Plot – but J48 shown

Length of Stay <15 days

	Predicted 0	Predicted 1
Observed 0	473	91
Observed 1	66	132

Best Algorithm: LWL w Naive Bayes

Free flap only complications

	Predicted 0	Predicted 1
Observed 0	14	161
Observed 1	4	264

Best Algorithm: LWL with Decision Stump

Fig. 3. Confusion matrices for the four models.

therapy, the potential breadth of the activity captured will be compromised.

Models that have been developed by other surgical specialties use risk-adjustment algorithms with an AUROC of between 0.65 and 0.85. It must be accepted that discrimination is imperfect so predicted scores for morbidity should not be used to plan the care of individual patients. For the purpose of comparative audit, however, calibration can be acceptable, and previous work using a neural network in this setting has shown weak discrimination, but excellent calibration when predicting for all complications in the entire group.[3]

Interestingly, while the objective of excellent discrimination may not have been achieved, useful classification rules have emerged that may prove to be an easily-measured means to highlight differences in performance. To our knowledge, these rules have not been seen in other national audits, and may provide a bridge in the quality assurance process while better-performing algorithms are developed from national datasets.

One of the strengths of this work is that the analysis captured the activity of six units with different case mixes and outcomes. This model is intrinsically more robust than the algorithms based on a single unit’s activity, in which a phenomenon termed “over-fitting” may suggest over-optimistic results that become evident only on external validation.

The assessment of the quality of care after head and neck surgery is not limited to complications, duration of hospital stay, or similar surgical metrics. Additional metrics should be developed to risk-adjust for surgical care. Patient-reported outcome measures (PROMS) that have been validated for the assessment of a patient’s quality of life, focus in particular on the multiple domains covered by the University of Washington quality of life questionnaire or the European Organisation for Research and Treatment of Cancer quality of life questionnaire head and neck module (EORTC H&N).[9–11] Objective functional outcomes such as the water swallow test or the penetration aspiration scale have focused on patients who have had chemoradiotherapy,[12] though exceptions have been reported that focus on primary surgery for advanced oropharyngeal disease. [13,14] Early first reports suggest an interest in the reporting of composite outcomes,[14,15] which remains, for the foreseeable future, fragmented.

Collaboration with computer scientists who are trained in statistics and machine learning is becoming established in modern medicine. Whilst over-complex and obscure models may not win the confidence of medical professionals (a problem called “the black box effect”), collaboration can help us to choose models that may not be as accurate overall, but are intuitive and more acceptable, because they show transparently the relative weights carried by different premorbid factors.

Ethics statement/confirmation of patients’ permission

Grey Area Project Ethics Board Approval, EKHUFT. Patients’ permission was not applicable.

Conflict of interest

The author is employed by East Kent Hospital University Foundation Trust, one of the units in the analysis.

Acknowledgement

We thank Dr Fabio Fabris for his help in preparing the ROC graphs presented in this work. The collaborative project was supported with a mini-grant from the East Kent Hospitals University Foundation Trust (EKHUFT) Research and Innovation Fund

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.bjoms.2019.05.026.

References

1. McMahon J, Handley TP, Bobinskas A, et al. Postoperative complications after head and neck operations that require free tissue transfer — prevalent, morbid, and costly. *Br J Oral Maxillofac Surg* 2017;**55**:809–14.
2. 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.
3. 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.
4. Hay A, Migliacci J, Karassawa Zanoni D, et al. Complications following transoral robotic surgery (TORS): a detailed institutional review of complications. *Oral Oncol* 2017;**67**:160–6.
5. O'Connell DA, Barber B, Klein MF, et al. Algorithm based patient care protocol to optimize patient care and inpatient stay in head and neck free flap patients. *J Otolaryngol Head Neck Surg* 2015;(44):45.
6. Perisanidis C, Herberger B, Papadogeorgakis N, et al. Complications after free flap surgery: do we need a standardised classification of surgical complications? *Br J Oral Maxillofac Surg* 2012;**50**:113–8.
7. Dindo D, Clavien PA. What is a surgical complication? *World J Surg* 2008;**32**:939–41.
8. Shellenberger TD, Madero-Visbal R, Weber RS. Quality indicators in head and neck operations: a comparison with published benchmarks. *Arch Otolaryngol Head Neck Surg* 2011;**137**:1086–93.
9. Thornton C, Hutter F, Hoos HH, et al. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2013)* 2013:847–55.
10. Rogers SN, Lowe D, Kanatas A. Suitability of the Patient Concerns Inventory as a holistic screening tool in routine head and neck cancer follow-up clinics. *Br J Oral Maxillofac Surg* 2016;**54**:415–21.
11. Rogers SN, Lowe D, Fisher SE, et al. Health-related quality of life and clinical function after primary surgery for oral cancer. *Br J Oral Maxillofac Surg* 2002;**40**:11–8.
12. Høxbroe Michaelsen S, Grønhøj C, Høxbroe Michaelsen J, et al. Quality of life in survivors of oropharyngeal cancer: a systematic review and meta-analysis of 1366 patients. *Eur J Cancer* 2017;(78):91–102.
13. Patterson JM, Hildreth A, McColl E, et al. The clinical application of the 100mL water swallow test in head and neck cancer. *Oral Oncol* 2011;**47**:180–4.
14. Seikaly H, Biron VL, Zhang H, et al. Role of primary surgery in the treatment of advanced oropharyngeal cancer. *Head Neck* 2016;**38**(Suppl. 1). E571-9.
15. Marzouki HZ, Biron VL, Dziegielewski PT, et al. The impact of human papillomavirus (HPV) status on functional outcomes and quality of life (QOL) after surgical treatment of oropharyngeal carcinoma with free-flap reconstruction. *J Otolaryngol Head Neck Surg* 2018;**47**:58.