Machine Learning methods applied to audit of surgical outcomes after treatment for Head and Neck Cancer

Authors

David Tighe MBChB BDS FRCS FFD-RSCI
Consultant OMFS Surgeon EKUHFT
FFD (RSCI) FRCS- OMFS DOHNS
Dft1@doctors.org.uk

Timothy Lewis-Morris
SpR Nephrology
EKUHFT
Timothy.lewis-morris@nhs.net

Alex Freitas PhD
Professor of Computational Intelligence,
School of Computing,
University of Kent
A.A.Freitas@kent.ac.uk

Abstract

Introduction

Most surgical specialities have attempted to address the concern of unfair comparison of outcomes by ‘risk-adjusting’ data in order to benchmark speciality specific outcomes indicative of quality of care. We are building on previous work to address the current need in Head and Neck surgery for a robust validated means of risk adjustment.

Methods

A dataset of care episodes recorded as a clinical audit of complications after surgery for Head and Neck Squamous Cell Carcinoma (n=1254) was analysed within the WEKA Machine Learning programme. The outcome(s) were 4 classification models that predict for complications using pre-operative patient demographic data, operation data, functional status data and tumour stage data.

Results

Of 4 models developed, 3 performed acceptably. The first model, predicting 'any complication' within 30days (Area Under the Receiver Operator Curve, AUROC 0.72). The second model predicted severe complications (Clavien-Dindo Grade 3 or greater) within 30days was also acceptable (AUROC, 0.70) and the 3rd model predicted prolonged length of hospital stay >15 days, (AUROC, 0.81). The final model, developed on a subgroup of patients who had free tissue transfer (n=443) performed poorly (AUROC, 0.59)

Conclusion

Sub-speciality groups within the Oral & Maxillofacial speciality are seeking metrics that will allow meaningful comparison of quality of care delivered by surgical units in the UK. In order for metrics to be effective they must demonstrate variation between units, be amendable to change by service personnel, and have baseline data available in the literature. We argue metrics can be effectively modelled in order that meaningful benchmarking, which takes account of variation in complexity of patient need / care, is possible.

Introduction

Benchmarking care in surgical audit requires definition of outcomes that allow meaningful comparison of different treatment centres. There is growing consensus that mortality rates, which are low (0.5-2%) in our speciality are not likely to be a helpful indicator of quality of surgical care.

Morbidity rates vary between surgical units and rates can be upto 70% from units reporting series of patients receiving major Head & Neck surgery with free-tissue transfer.[1]

To mitigate some of the subjectivity involved in recognising a complication, and to improve consistency, the Clavien-Dindo classification system was developed [2], which is well covered in the surgical literature and has been applied to Head and Neck surgery by multiple authors[1,3,4,5,6] it allows recording of surgical complications and systemic complications in a clear fashion that 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 further underscore their belief that “the incidence of postoperative complications [should still be] the most frequently used surrogate marker of quality in surgery.”[7]

Work has been done to benchmark quality of care after Head and Neck surgery more broadly by creating and testing additional metrics such as length of stay, use of blood products, adequacy of pathology reports in addition to those items contained in the Clavien – Dindo classification, namely return to theatre rates and mortality rates.[8]

In order for quality metrics to be effective they must demonstrate variation between units, be amendable to change by service personnel, and have baseline data available in the literature. We argue that a fourth criteria be added, that they can be effectively modelled statistically in order that meaningful benchmarking is possible that takes account of variation in complexity of patient need / care.

We attempt to test if all complications and severe complications (Clavien-Dindo Level 3 and above) are valid outcomes to capture for the purpose of comparative audit of quality of care. We also attempt to test if length of hospital stay is a proxy indicator of quality of care. Finally we test if it is preferable to model just the subset that receives immediate free-tissue transfer reconstruction.

Methods

1254 patient care episodes for surgical care (with curative intent) under general anaesthesia for HNSCC were analysed. This represented the combined dataset of 6 treatment units.

Of these 160 from Site 1, 203 from Site 2, 521 from Site 3, 175 from Site 4, 83 from Site 5 and 112 from Site 6. 444/1254 (35%) had free tissue transfer. Data pertaining to length of hospital stay (not included in the dataset from Site 3) was available on 733/1254 (58%).

Data was preprocessed by the lead author and experiments were performed using three methods in the ‘Classify’ panel of the WEKA data mining (or machine learning) tool: AutoWEKA [8] , the J48 decision tree, and the random forest algorithm. Auto-WEKA is an advanced machine learning

method which automatically selects the best classification algorithm and its best configuration for an input dataset, by doing a systematic search of many different types of algorithms and their configurations available in WEKA. Both the J48 and the random forest algorithms are included in the large set of algorithms considered by Auto-WEKA during its search, but we also used these two algorithms separately for the following reasons. J48 learns an interpretable model in the form of a decision tree, highlighting the most relevant attributes for classification. Although J48 does not achieve very 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 very accurate and interpretable rules extracted from a J48 decision tree will be shown later. The random forest algorithm was used because it is among the state-of-the-art algorithms regarding predictive performance. As a measure of predictive performance, the results are reported using only the Area Under the Receiver Operating Characteristic curve (AUROC). Auto-WEKA is a non-deterministic method (its results depend on a random seed used to initialize the program), hence we ran Auto-WEKA multiple times and report the mean AUROC in a process called 10 fold cross validation set in run-time frames of 5 hours and 20 hours for each experiment. For predicting the length of stay class variable, cut-offs based on previous work were tested, namely >15 days and >20 days. The AUROC statistics are quoted with ROC diagrams.

Results.

Successful analysis of the dataset allowed modelling of the following (Table 1) using the J48 decision tree, random forest plot algorithm and the Auto-WEKA package.

* Complications with 30 days

* Complications within 30days that were Clavien Dindo grade 3 or higher

* Length of hospital stay.

* Complications within 30 days in the immediate free tissue transfer group

The best model predicting ‘Complication within 30 days’ was produced by Auto-WEKA, which selected a ‘Bagging of J48 decision trees’, i.e. a collection of decision trees, with an AUROC of 0.730. It is hard to interpret the entire collection of many decision trees, though; therefore, for interpretation purposes we focus on the single decision tree produced by the J48 algorithm. This model still has an adequate AUROC value of 0.705, and it can be readily embedded within a database for immediate computation of predicted risk.

The J48 decision tree (Figure 1) demonstrates the most important predictors of risk of complication in a graphical and hierarchical form, where in general the most relevant predictors are closer to the top of the decision tree, and less relevant predictors are further down in the tree. Use of tracheostomy, immediate free tissue transfer and complexity of surgery appear at the top of the tree.

The AUROC was higher with Auto-Weka, though it is harder to interpret the model, unlike those of the J48 decisiion tree which are directly interpretable.

A basic classification rule that can be derived from this model states that: ‘IF Scale of Surgery = Minor (1) and Tracheostomy = Absent (0) and Free Flap = Absent (0) (i.e. ‘no’), THEN the patient should have no complication’. This rule is 85% accurate, correctly classifying 204.38 out of the 240.38 patients in our dataset. Therefore, a threshold complication rate of 15% should be tolerated in this ‘low-risk’ group.

The best model predicting ‘Complications within 30 days graded as Clavien – Dindo 3 or greater’ was produced by the random forest algorithm. This model was acceptable with AUROC of 0.70.

The Random Forest algorithm however demonstrates a substantial shortfall, the model is consistently predicting 'no complication' and only predicting a complication in about 1% of the cases (11/1322), despite the fact that the average observed complication rate across all sites is 14.4% – though the range varies significantly (Site 1: 6%, Site 2: 9%, Site 3: 11%, Site 4: 30%, Site 5: 21%, and Site 6 14%).

Classification models were built for predicting the Length of stay using 15 day and 20 day cut-offs. The best models for both cut-offs were produced by Auto-WEKA, in both cases with a good AUROC value greater than 0.8. and the model structure is interpretable. A highly reproducible rule is available from the decision tree produced by J48 for predicting < 20 days of stay, namely:

IF (Tracheostomy = 0) THEN (InPatient Days < 20 days)

.This rule has an accuracy of 90.2%, correctly classifying 515.5 out of 571.25 patients.

Finally the 'Complications within 30 days model in the immediate free tissue transfer group' had acceptable discrimination in AutoWeka (AUROC of 0.71) which degraded once validated (AUROC dropped to 0.6, best algorithm recommended by Auto-WEKA, LWL (Locally Weighted Learning) with Decision Stump). LWL uses a nearest-neighbour algorithm to assign instance weights which are then used by a specified base classifier, which in this case was a simple Decision Stump (a decision tree with just one node). This accuracy is barely acceptable as an algorithm for the purpose of risk adjustment in audit. The dataset for this group was smaller (n=444) and modelling may improve with more data, if performance of contributing units is similar to the 6 units contributing data thus far.

The confusion matrix of the 4 models are shown (Figure 3). A composite on the AUROC are also shown (Figure 4).

Discussion

The results suggest that acceptable performance, for the purposes of risk adjustment is found in the models predicting any complication in the period of 30 days after surgery and predicting length of hospital stay (using cut-offs of 15 and 20 days). There is some concern that the model for predicting severe complications (Clavien Dindo 3 or greater) is insufficiently discriminating, as it is tending to predict all patients as not having a severe complication with few exceptions. Finally, the model predicting any complications in the period of 30 days after free flap surgery is at present insufficient for the purpose of risk-adjustment in an audit process.

Models developed by other surgical specialities use risk-adjustment algorithms with AUROC between 0.65 and 0.85. There must be some acceptance that discrimination is imperfect and predicted morbidity scores should therefore not be used to plan individual patient care. However, calibration for the purpose of comparative audit can nevertheless be acceptable and previous work using a neural network in this setting demonstrates weak discrimination but excellent calibration when predicting for all complications in the entire cohort.[3]

Interestingly, while excellent discrimination may remain an unmet objective of this work, useful classification rules have emerged from the analysis, and these rules may present easily measured means of highlighting differences in performance. The use of such rules has not been seen in other national audits and may provide a bridge in the Quality Assurance process, whilst the development of better performing algorithms developed from national datasets progresses.

A strength of this work is that the analysis captures the activity of 6 units with demonstrably different case mix and outcomes. This makes modelling intrinsically more robust relative to units publishing algorithms developed on a single unit's activity, where a phenomenon termed 'over-

fitting' may suggest over-optimistic results that only become evident when external validation of the algorithms is carried out.

Assessing quality of care after Head and Neck surgery is not limited to complication rates or length of hopsital stay or similar surgical metrics. Additional metrics should be developed with an attempt to risk-adjust for surgical care. Quality of Life measurement focusing in particular on the multiple domains covered in the University of Washington Quality of Life questionnaire or the EATOC H&N .[9-11] Objective Functional Outcomes such as the water swallow test (WST) or the Penetration Aspiration Scale have focused on the cohort receiving chemo-radiotherapy[12] though exceptions are in the literature they focus on primary surgery for advanced oropharyngeal disease. [13,14] Early first reports are suggesting an interest in composite outcome reporting[14,15] which remains for the foreseeable future fragmented.

Collaboration with computer scientists with statistical and machine learning training is becoming established in modern medicine. Whilst over-complex and obscure processes may not win confidence of medical professionals, 'the black box effect', collaboration can help choose lesser models that are intuitive and therefore more acceptable because they demonstrate in a transparent fashion the relative weights that different pre-morbid factors carry.

This work has been supported by a Research and Innovation Grant from EKHUFT

There are no conflict of interests to declare.

Ethical approval was given for this audit, as the results are potentially generalizable, by the Grey Area Project group, EKHUFT.

References

1. McMahon J, Handley TPB, Bobinskas A, Elsapagh M, Anwar HS, Ricciardo PV, McLaren A, Davis R, Syyed N, MacIver C, Wales C, Hislop WS, Thomson E, Thomson S, Fitzpatrick K, Rae A, Campbell R. Postoperative complications after head and neck operations that require free tissue transfer - prevalent, morbid, and costly. Br J Oral Maxillofac Surg. 2017 Oct;55(8):809-814. doi: 10.1016/j.bjoms.2017.07.015. Epub 2017 Aug 12.
2. Dindo D, Demartines N, Clavien PA (2004) Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. Ann Surg 240:205–213
3. Tighe DF Thomas AJ Sassoon I Kinsman R McGurk M. Developing a risk stratification tool for audit of outcome after surgery for head and neck squamous cell carcinoma. Head Neck. 2017 Jul;39(7):1357-1363
4. Hay A1, Migliacci J1, Karassawa Zanoni D1, Boyle JO1, Singh B1, Wong RJ1, Patel SG1, Ganly I2.Complications following transoral robotic surgery (TORS): A detailed institutional review of complications.Oral Oncol. 2017 Apr;67:160-166.
5. O'Connell DA1, Barber B2, Klein MF3, Soparlo J4, Al-Marzouki H5, Harris JR6, Seikaly H7. 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 Nov 2;44:45. doi: 10.1186/s40463-015-0090-6.
6. Complications after free flap surgery: do we need a standardized classification of surgical complications? Perisanidis C1, Herberger B, Papadogeorgakis N, Seemann R, Eder-

Czembirek C, Tamandl D, Heinze G, Kyzas PA, Kanatas A, Mitchell D, Wolff KD, Ewers R.Br J Oral Maxillofac Surg. 2012 Mar;50(2):113-8. doi: 10.1016/j.bjoms.2011.01.013. Epub 2011 Feb 22.

7. Dindo D, Clavien PA. What is a surgical complication? World J Surg. 2008 Jun;32(6):939-41
8. Shellenberger TD1, Madero-Visbal R, Weber RS. Quality indicators in head and neck operations: a comparison with published benchmarks. Arch Otolaryngol Head Neck Surg. 2011 Nov;137(11):1086-93. doi: 10.1001/archoto.2011.177.
9. C. Thornton, F. Hutter, H. H. Hoos, K. Leyton-Brown, K. L.-B. Chris. Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2013), pp. 847-855.
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 May;54(4):415-21.
11. Rogers SN1, Lowe D, Fisher SE, Brown JS, Vaughan ED. Health-related quality of life and clinical function after primary surgery for oral cancer.Br J Oral Maxillofac Surg. 2002 Feb;40(1):11-8.
12. Høxbroe Michaelsen S, Grønhøj C, Høxbroe Michaelsen J, Friborg J, vonBuchwald C. 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, Carding PN, Hamilton D, Wilson JA.The clinical application of the 100mL water swallow test in head and neck cancer. Oral Oncol. 2011 Mar;47(3):180-4. doi: 10.1016/j.oraloncology.2010.11.020. Epub 2011 Jan 12.
14. Seikaly H, Biron VL, Zhang H, O'Connell DA, Côté DWJ, Ansari K, et al. Role of primary surgery in the treatment of advanced oropharyngeal cancer Head Neck. 2016;38(Suppl 1)
15. 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 Marzouki HZ1, Biron VL2, Dziegielewski PT3,4, Ma A2, Vaz J2, Constantinescu G2, Harris J2, O'Connell D2, Seikaly H2. J Otolaryngol Head Neck Surg. 2018 Sep 19;47(1):58.

Extra refs

16. Quality of life after free flap surgery for cancer of the head and neck in patients with or without postoperative complications. Lahtinen S, Koivunen P, Ala-Kokko T, Laurila P, Kaarela O, Liisanantti JH. Eur Arch Otorhinolaryngol. 2018 Oct;275(10):2575-2584. doi: 10.1007/s00405-018-5103-4. Epub 2018 Aug 24.
17. Complications and outcome after free flap surgery for cancer of the head and neck. Lahtinen S, Koivunen P, Ala-Kokko T, Kaarela O, Ohtonen P, Laurila P, Liisanantti JH. Br J Oral Maxillofac Surg. 2018 Oct;56(8):684-691. doi: 10.1016/j.bjoms.2018.07.009. Epub 2018 Aug 11.