Technical appendix - Validating risk-adjustment models used in QOMS

Authors: Tighe DF, Sassoon I, Provost S, Puglia FA, and Ho MWS

1. INTRODUCTION

The Quality and Outcomes in Oral and Maxillofacial Surgery (QOMS) project is the quality improvement and clinical effectiveness programme for Oral and Maxillofacial Surgery (OMFS), initiated by the British Association of Oral and Maxillofacial Surgeons (BAOMS) in 2018. QOMS operates a series of audits across several OMFS subspecialties to assess the quality of care provided to patients in OMFS units in the UK. The QOMS Oncology & Reconstruction and Non-Melanoma Skin Cancer (NMSC) audits are set apart from the other QOMS registries for having published risk-adjustment models built in in their online interface and thus providing users with patient-level risk-adjustment as and when data are entered.

This appendix includes a technical report of the *model development* and *model validation*.

Model development overview

Statistical models are mathematical representations of observed data. They are used to control for other variables (risk-adjustment), to model relationships between variables, and to predict outcomes. Models are selected based on how well they fit the original or development/training dataset. Model development refers to the curating of a dataset of consecutive surgical episodes from one or more NHS trusts that records key 'independent' (i.e., that pertain to risk) and 'dependent' (i.e. metrics or outcomes measured) variables. Independent variables include patient's characteristics, tumour and treatment factors, all generally known before a time point after which outcome variables are captured (e.g. before and after surgery).

Traditional statistical approaches to modelling are logistic and linear regressions for binary and continuous variables, respectively. Alternative statistical methods include Bayesian probabilistic models and recently more complex and computationally intensive methods including decision tree, artificial neural network and automated machine learning techniques.

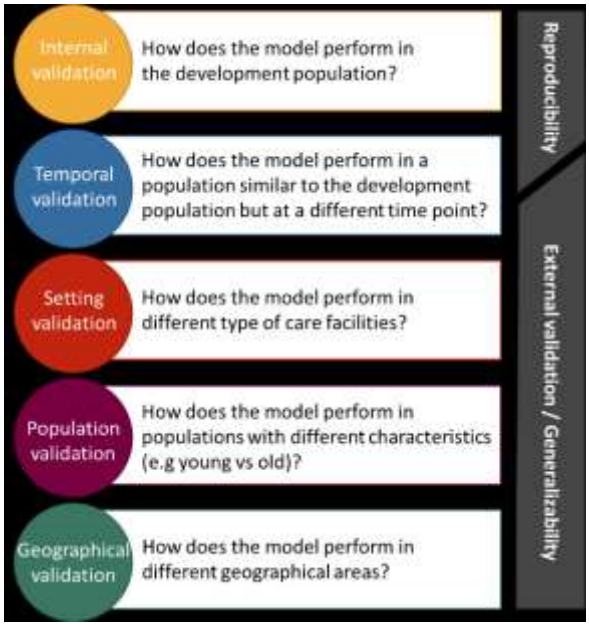
Automated machine learning is a specialist field that seeks to 'turn over' to machines the transposition of data, analysis and validation of model strength in an iterative manner that seeks the 'best' or 'champion' method out of a library of dozens of steps, hundreds of model architectures and a few validation options. This automated process takes hours of computational time, and the insights and boundaries of this exploration are documented in this appendix.

Model validation

Model validation is a series of steps / process carried out on a 'champion' model to verify it achieves its intended purpose, i.e., that the model is predictive under the conditions of its intended use. There are different levels of validation (Figure 1.1), some take place on the development dataset (internal validation) or on a new/blind and independent experimental dataset (external validation). Validation tries to answer the question about how generalisable a model is and explores if it can be reliably applied on new

datasets that may differ on some levels (e.g., patients' characteristics, setting, locations…) from the development dataset.

Figure 1.1. Model validations



*As **internal validation** uses the patients from the development population, it can always be performed but is limited to providing information on the reproducibility of the model. Other types of validations use populations that differ from the development cohort to varying degrees. **Temporal validation** (i.e., the "unseen" population is sampled at an earlier or later time point to the development population) is often considered to lie midway between internal and external validation. It provides some information on a model's reproducibility and generalisability. **External validation** mainly provides evidence on the generalisability to various different patient populations (e.g., geographic validation, populations from different types of care facilities or with different general characteristics). The extent of validation depends on the research question and size of the development cohort and not every model needs to be validated in all the ways depicted. Adapted from Ramspek et all. (2021)* [1]

Early phase of developing risk-adjustment models typically included an 'internal validation' process which involves splitting the dataset into a training subset (70%) and a test subset (30%) after random division of the dataset. Later, "model development" papers describe the use of machine-learning data pipelines, which use k-fold validation in which the user specifies the number of sequential dataset division and re-testing to be done (usually 5-10 'folds'). This increases the chance of reporting a 'fair' assessment of model performance less susceptible to unfavourable or favourable chance division of the dataset, which could result in underfitting or overfitting, respectively.

The models presented here have already been described elsewhere. They were developed independently of QOMS. In other words, the care episodes in QOMS were in no way involved in developing the models, i.e., the data in QOMS is currently 'unseen' data. Therefore, applying those models to the QOMS datasets is part of their external validation.

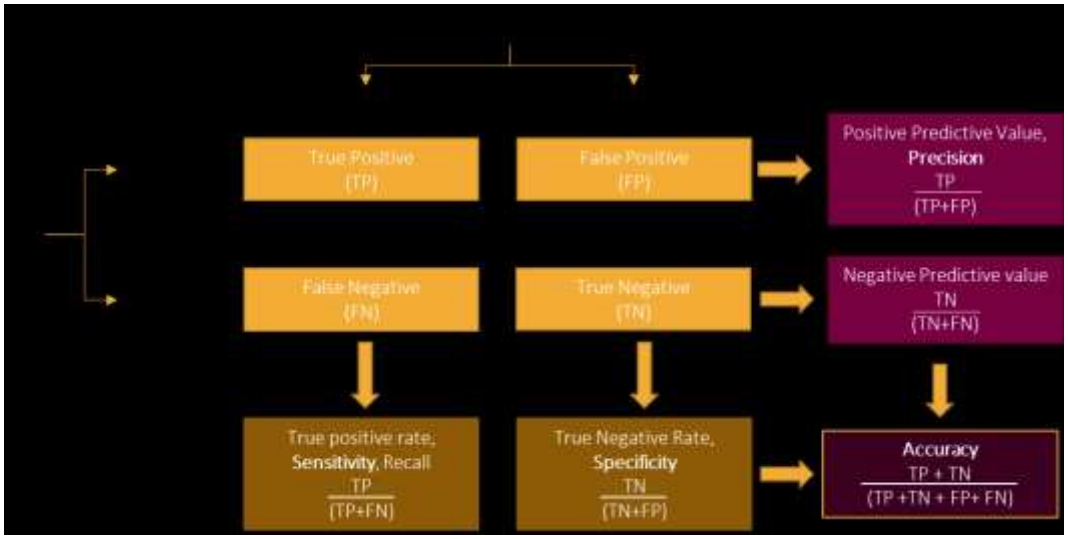
2. MATERIAL AND METHODS

MODEL VALIDATION FOR BINARY OUTCOMES

Confusion matrix

When assessing model performance for binary outcomes, it is common to report the different types of classification errors seen. They can be summarised in a confusion matrix (Figure 2.1). A confusion matrix is similar to a contingency table, where the two different grouping variables used to classify the data are the observed and predicted outcomes. Several measures of model performance (sensitivity, precision…) can be calculated directly from the confusion matrix (Figure 2.1). In addition, the F1-score, which is the harmonic mean between the precision and recall, can also be calculated (formula (2.1)). Machine Learning approaches favour maximising the F1-score, to take into account the trade-off between precision and recall (since an increase in one usually leads to a decrease in the other).

Figure 2.1. Confusion matrix



(2.1) $$F1 - score = \frac{2 \times precision \times recall}{precision + recall}$$

The calculations of precision, recall and F1-score require the user to specify what the positive and negative classes are. Here, we used the common approach to treat each of the two classes (e.g. 'clear margin' and 'positive margin' OR 'complication' and ' no complication') in turn as the positive/negative class and then to compute for each measure the average over those two values, producing the so-called **macro averages** of precision, recall and F1-scores. These macro averages consider the performances in the predictions of both classes as equally important and are considered more informative measures than the corresponding micro averages (micro averages are weighted by the numbers of examples in each class and are

thus overwhelmingly dominated by the performance in the majority class, due to the large class imbalance in our dataset). Hence in our later papers,[2,3] we have reported the macro averages of precision, recall, and F1-score in the Results section for the skin margin and free flap risk adjustment models.

Receiver operating characteristic

A receiver operating characteristic (ROC) curve is a plot of the true positive rate (TPR or sensitivity) versus the false positive rate (FPR):

(2.2) $$FPR = 1 - specificity = \frac{TN}{FP+TN}$$

The ROC curve is produced by varying the value of a classification threshold on the probability of the positive class predicted for each example so that different values of that threshold generate different points (TPR and FPR values) on the curve. The area under that curve, which is a popular measure of predictive performance, is also reported in the result sections of our papers. [2–6]

The model outputs for each care episode can be aggregated as a 'mean' expected outcome for each individual participating institution. The risk-adjusted outcomes for all cases then are calculated using the formula for indirect standardisation (2.3). Indirect standardisation can suffer from 'small numbers' whereby infrequent events in small samples, due to effect of random variation, can disproportionately affect perceived performance.

(2.3) $$DAG\ Risk\ adjusted\ outcome\ rate = \frac{predicted\ outcome\ for\ DAG_i}{mean\ predicted\ for\ all\ DAGs} \times observed\ outcome\ rate for\ DAG$$

(DAG = Data access group. In QOMS, each DAG represent an individual institution)

Graphical representation of comparative data for binary outcomes – funnel plots

Funnel plots are favoured for their ability to graphically demonstrate wider confidence intervals when numbers are small. They however suffer from their inability to identify units in which different 'subgroups' are combined to make an aggregate count. If one hospital has a disproportionate number of particularly high- or low-risk patients, despite risk-adjustment, this may skew the overall observed rates and its graphical representation.

When applying risk-adjustment models with the intention of publishing comparative data, Verburg et al. (2017) [7] recommended to follow the guidance they developed to construct quality assessment graphics (Table 2.1).

Table 2.1. Summary of the guidance to construct a funnel plot [7]

Step 1. Define the policy-level decisions (i.e., contextual decisions and analytical plan)
a. The quality indicator and associated external or internal benchmark * b. The data source, or registry, and patient population, including inclusion and exclusion criteria c. The reporting period d. Control limits and whether data analysts are allowed to inflate them to correct for overdispersion, which occurs when there is true heterogeneity between institutions, over and above the expected level of variation due to randomness. *(* If no external benchmark is available and an internal benchmark is used instead, it needs to be calculated. This will require to have at disposition the observed count of the outcome, the expected count, and the admission count, and to calculate the observed rate / proportion of the outcomes, the standardised rate (SR) of the outcome and its risk-adjusted rate (RAR))*
Step 2. Check the quality of the risk-adjustment model used
Verburg et al. recommended using goodness-of-fit statistics for calibration, the Brier score to indicate overall model performance, and the concordance (or C) statistic for discriminative ability. It is worth mentioning that no consensus exists on the values of these performance measures to indicate whether a model is or is not of 'sufficient' quality for the purpose of benchmarking.
Step 3. Check if there are a sufficient number of observations per hospital (i.e., power)
In funnel plots, users should be able to interpret results and reliably assume that an institution inside or outside of the control limits does not result from chance. Therefore (for a given significance level, effect size, and statistical power) a minimum sample size required should be achieved for each institution. It is important to note that for low-volume institutions, control limits are essentially meaningless.
Step 4. Test for overdispersion of the values of the quality indicator.
A major assumption here is that observed differences are true differences in the quality of care (and random variations). If there is true heterogeneity between hospitals over and above that expected due to random variation (i.e., overdispersion), then this assumption is violated and conclusions from the funnel plot should be drawn carefully.
Step 5. Test whether the values of the quality indicators are associated with institutional characteristics.
Funnel plots can be used to identify institutions whose performance deviate from a benchmark and not to make between-institution comparisons. That's why it is required to ensure there is no association between quality indicator and hospital characteristics.
Step 6. Specify how the funnel plot should be constructed (i.e., how data is presented, units, scales…)

Practically, it means that:

- Step 2. We also added a re-calibration curve (observed vs. predicted outcome / institution) to assess whether the model over- or under-predicted the outcome. If an institution is above the 1:1 line, then the model underpredicts and vice versa,
- Step 3. The minimal sample size was calculated as the number of cases necessary to detect a 50% increase (x1.5) of the proportion or standardized rate from the benchmark with at least 80% power at 95% and 99% control limits.

MODEL VALIDATION FOR CONTINUOUS OUTCOMES

As both observed and predicted lengths of stay (LoS) are continuous variables, the prediction accuracy of the linear regression can be assessed by calculating:

- Root-mean-square error (RMSE) as the square root of the mean-square error (MSE):

(2.4) $$MSE = \frac{1}{n}\sum(oberved - predicted)^2 \text{ and } RMSE = \sqrt{MSE}$$

- Residual Standard Error (RSE) or model sigma. RSE is a variant of the RMSE adjusted for the number of predictors in the model. The lower the RSE, the better the model. In practice, the difference between RMSE and RSE is very small, particularly for large multivariate data.
- Mean Absolute Error (MAE), like the RMSE, the MAE measures the prediction error. MAE is less sensitive to outliers compared to RMSE.

(2.5) $$MAE = \frac{1}{n}\sum observed - predicted \text{ V}$$

DATASETS USED

The QOMS metrics for Oncology & Reconstruction currently

- Complications within 30 days of surgery
- Length of postoperative hospital stay (in days)
- Positivity of surgical margins (also applied to NMSC)
- Complete free flap failure.

For the recording of complications, some intra-operative complications, like failure to harvest a viable flap because of technical error or unintentional transection of cranial nerves, merit recording whereas other like chyle leak or haemorrhage controlled with conventional measures do not.

We will now address an overview of each model in turn.

FURTHER READING

The reader is referred to the peer reviewed papers for further information.

30-day complication Model	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. doi: 10.1002/hed.24769.	Categorical
Length of hospital stay	Tighe D, Sassoon I, Hills A, Quadros R. Case-mix adjustment in audit of length of hospital stay in patients operated on for cancer of the head and neck. Br J Oral Maxillofac Surg. 2019 Nov;57(9):866-872. doi: 10.1016/j.bjoms.2019.07.007.	Categorical & continuous

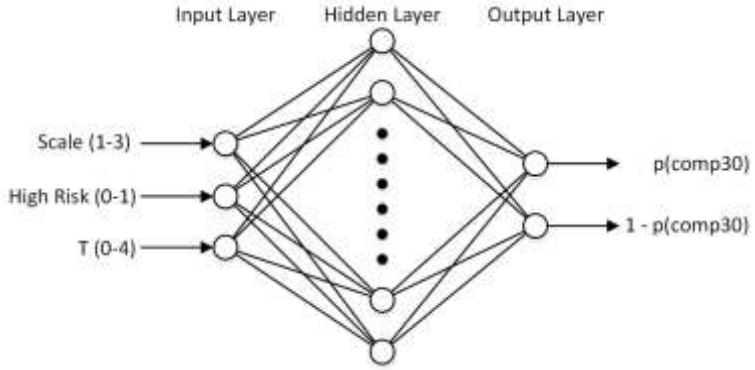
HNSCC surgical margin Model	Tighe D, Fabris F, Freitas A. Machine learning methods applied to audit of surgical margins after curative surgery for head and neck cancer. Br J Oral Maxillofac Surg. 2021 Feb;59(2):209-216. doi: 10.1016/j.bjoms.2020.08.041.	Categorical
Free-flap failure Model	Tighe D, McMahon J, Schilling C, Ho M, Provost S, Freitas A. Machine learning methods applied to risk adjustment of cumulative sum chart methodology to audit free flap outcomes after head and neck surgery. Br J Oral Maxillofac Surg. 2022 Dec;60(10):1353-1361. doi: 10.1016/j.bjoms.2022.09.007.	Categorical
Non-melanoma skin cancer Surgical Margin Model	Tighe D, Tekeli K, Gouk T, Smith J, Ho M, Moody A, Walsh S, Provost S, Freitas A. Machine learning methods applied to audit of surgical margins after curative surgery for facial (non-melanoma) skin cancer. Br J Oral Maxillofac Surg. 2023 Jan;61(1):94-100. doi: 10.1016/j.bjoms.2022.11.280.	Categorical

3. VALIDATING THE ARTIFICIAL NEURAL NETWORK MODEL FOR 30 DAY COMPLICATIONS

INTRODUCTION

Multilayer feed-forward neural networks are an important class of neural networks often used for classification tasks. They consist of an input layer, one or more fully interconnected hidden layers and an output layer (see example in Figure 3.1).

Figure 3.1. Structure of the classifier network developed for 30-day complications



The interconnections represent weights which are randomised when the network is initialised and adjusted during the training phase – most commonly using the backpropagation algorithm or one of its variants. It consists of a forward pass, where the training data is applied to the inputs and processed by the network, followed by a backward pass. This compares the neural network response to the target, generating an error signal which is propagated backwards through the network. The weights are adjusted in a direction such as to reduce the response error. The process is repeated a number of times (epochs) until some termination condition is met.

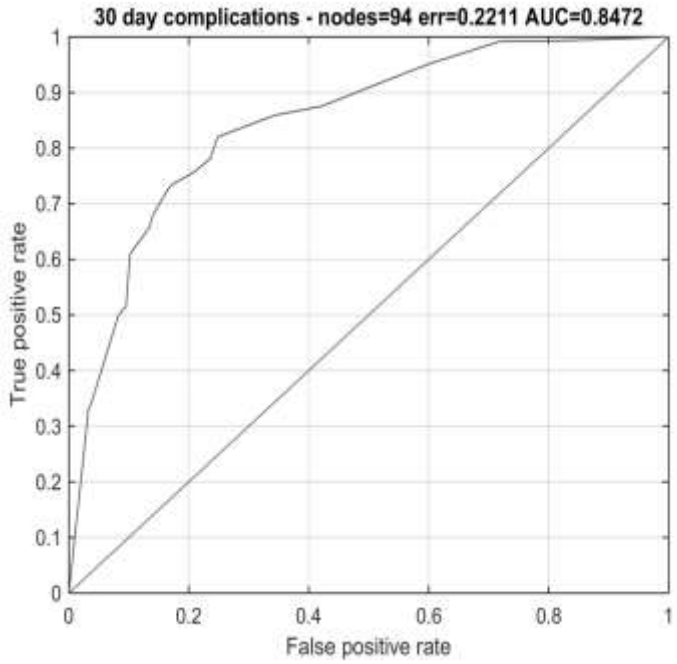
The number of hidden nodes is critical to the performance of the network. Too few, and it may not have the capacity to solve the problem at hand. Too many will result in overfitting, which is characterised by a good response to the training data, but poor response to unseen data (i.e., poor generalisation capacity). Unfortunately, there is no way to determine the best number of hidden nodes *a priori* as it depends on many factors in the problem domain. In our experiments, we adopted a frequently used method of training k networks with n hidden nodes where $1 \leq n \leq n_{max}$. In each case the training (70%) and test datasets (30%) are selected randomly from the full data set. This results in a total of $k.n_{max}$ networks from which the one with the best AUC response to unseen data is selected.

Artificial Neural Network (ANN) model

Referring to Figure 3.1, our network has three inputs: Scale of surgery (Scale, integer between 1 and 3), High Risk (dichotomic, 0 or 1), and pathological T stage (T, integer between 0 and 4). The AUROC of the output p(comp30) was used to select the champion.

MatlabR2014b 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. With $k = 1000$, and $n_{max} = 100$, one of the networks with 94 hidden nodes gave the best response to the test data with an AUC of 0.85 and a misclassification rate of 22% (Figure 3.2).

Figure 3.2. Receiver operating characteristic curve of the selected network



From a practical point of view, for each combination of "Scale of surgery", "High Risk" and "T", the ANN produced a risk of developing complication within 30 days of surgery. The ANN *per se* could not be built into the QOMS registry and applied directly to QOMS dataset, therefore patients were attributed a risk of developing complication according to their combination of "Scale", High Risk" and "T". The output is a probability, which differs for each combination of inputs (Table 3.1).

The reader's attention is drawn to the highest risk groups (of 30 day complications), namely those in which surgery is for the lower jaw / floor or mouth region reaching the neck for a T3,4 tumour (96.5% and 87.4%) which are higher rates that in the group receiving >6-hour surgery including free flaps (Scale 3). The lesson here is that risk is not linear with operative complexity and surgeons are deterred from the most complex operations, usually in the most co-morbid patients.

Table 3.1 Artificial Neural Network outputs

Scale	T	High-risk	%
1	0	0	**0.73**
1	0	1	**77.45**
1	1	0	**11.63**
1	1	1	**25.92**
1	2	0	**17.17**
1	2	1	**46.48**
1	3	0	**23.01**
1	3	1	**96.53**
1	4	0	**11.12**
1	4	1	**87.39**

Scale	T	High-risk	%
2	0	0	**33.23**
2	0	1	**64.53**
2	1	0	**24.97**
2	1	1	**34.89**
2	2	0	**38.48**
2	2	1	**46.23**
2	3	0	**61.01**
2	3	1	**89.74**
2	4	0	**46.44**
2	4	1	**19.51**

Scale	T	High-risk	%
3	0	0	**59.45**
3	0	1	**45.78**
3	1	0	**52.42**
3	1	1	**42.86**
3	2	0	**53.1**
3	2	1	**60.9**
3	3	0	**47.83**
3	3	1	**57.57**
3	4	0	**46.13**
3	4	1	**68.44**

APPLYING THE VERBURG GUIDANCE

Step 1. Define the policy-level decisions

The quality indicator considered here is the rate of complications within 30 days following head and neck surgery. The benchmark of the quality indicator is the cohort average (mean).

The QOMS team considered only capturing data of 'severe complications' (Clavien-Dindo Grade > IIIa) but this was deemed unsuitable by the Steering Committee as similar complications can in certain circumstances be treated at the bedside or dental chair rather than returning the theatre. Thus, the chance of 'gaming' or changing practice to appear to have a lower rate of 'severe complications' was avoided by applying the 'All Complication' model to all post-operative complications.

Data was extracted from the QOMS Oncology & Reconstruction registry on 31/10/2022 for all cases with a cancer diagnosis, created before 01/07/2022 (to maximise the number of cases and allow for data completion of the outcome variables). The dataset contained 1160 records of which 98 were missing the data relative to the actual presence or absence of complications. Of 1062 patients, 246 could not be attributed an ANN risk because at least one of the predictors (scale of surgery, High Risk or pathological T stage) was missing.

Calculating the benchmark (Table 3.2)

Table 3.2 Results of the Verburg guidance when producing funnel plots

Organisations	Observed count	Expected count	Admission count	Observed rate / Proportion	SR	RAR
OMFS-107	4	7.083	19	0.211	0.565	0.192
OMFS-116	4	5.542	9	0.444	0.722	0.245
OMFS-120	36	47.111	112	0.321	0.764	0.260
OMFS-130	123	113.036	228	0.539	1.088	0.370
OMFS-151	12	25.586	50	0.240	0.469	0.160
OMFS-157	20	25.261	49	0.408	0.792	0.269
OMFS-161	58	72.038	203	0.286	0.805	0.274
OMFS-166	10	39.819	84	0.119	0.251	0.085
OMFS-20	32	28.841	68	0.471	1.110	0.377
OMFS-58	17	36.432	89	0.191	0.467	0.159
OMFS-84	3	13.134	27	0.111	0.228	0.078

Finally, the benchmark values for SR and RAR are calculated:

- SR benchmark = 0.7707502
- RAR benchmark = 0.2621208

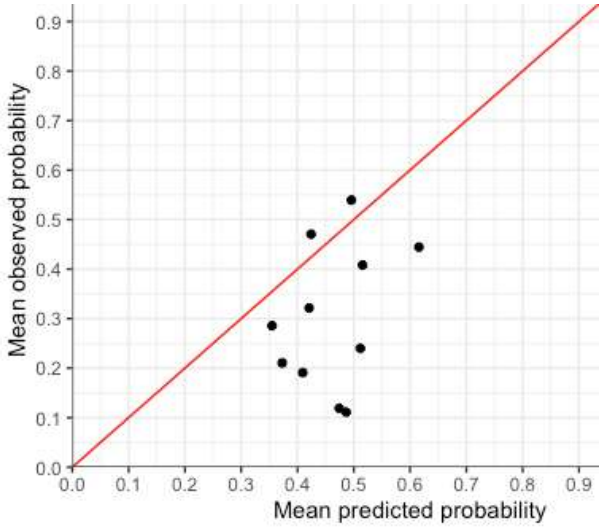
Step 2. Check the quality of the risk-adjustment model used

- Brier Score = 0.2239286
- Scaled Brier Score = 0.003993198
- C statistic = 0.6559194

The calibration of the ANN to QOMS data is weak – moderate.

- The recalibration curve indicates that the model overestimates the complication rate in 9/11 cases and underestimates for 2/11 cases. (Figure 3.3).

Figure 3.3. Mean predicted vs. observed rates of complication (recalibration curve) by participating hospital



Step 3. Is the number of observations per hospital sufficient?

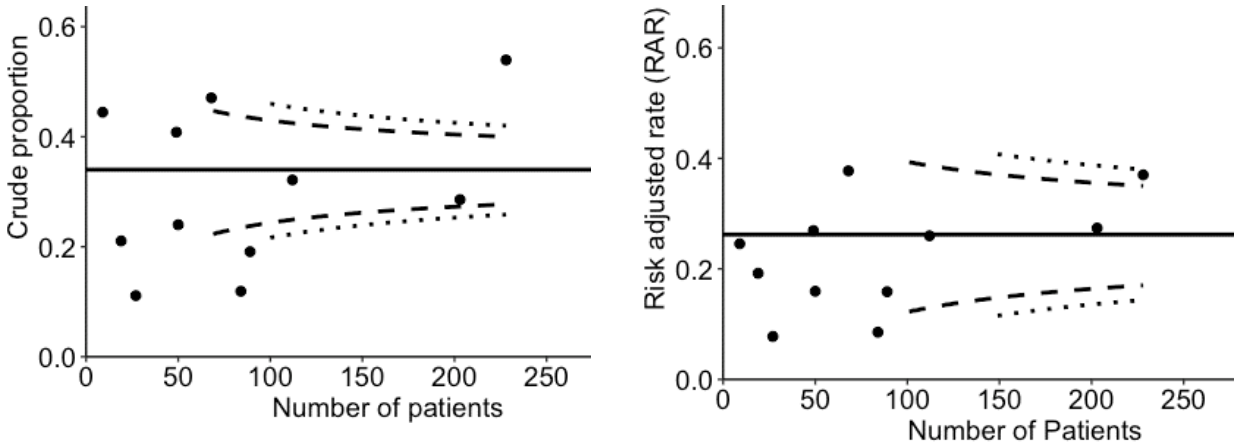
- Minimal sample size for observed rate: n = 69
- Minimal sample size for RAR: n = 101

Verburg et al [7] recommended that if fewer than half of the hospitals had enough cases to fulfil this criteria, funnel plots should not be constructed. In our cohort, 5/10 hospitals had 69 or more records and only 3/10 over 101 records. One hospital (1/11) was excluded for non-participation.

Funnel plots [9]

The complication rates by participating hospital are shown below in the funnel plots for crude and risk-adjusted data (Figure 3.4).

Figure 3.4. Funnel plot for crude probability of complication without correction for overdispersion (left) and for the risk adjusted rate of complication with correction for overdispersion (right)



CHAPTER REFERENCES

Spiegelhalter, David J. 2005. "Funnel Plots for Comparing Institutional Performance." *Statistics in Medicine* 24 (8): 1185–1202.

Tighe, David F, Alan J Thomas, Isabel Sassoon, Robin Kinsman, and Mark McGurk. 2017. "Developing a Risk Stratification Tool for Audit of Outcome After Surgery for Head and Neck Squamous Cell Carcinoma." *Head & Neck* 39 (7): 1357–63.

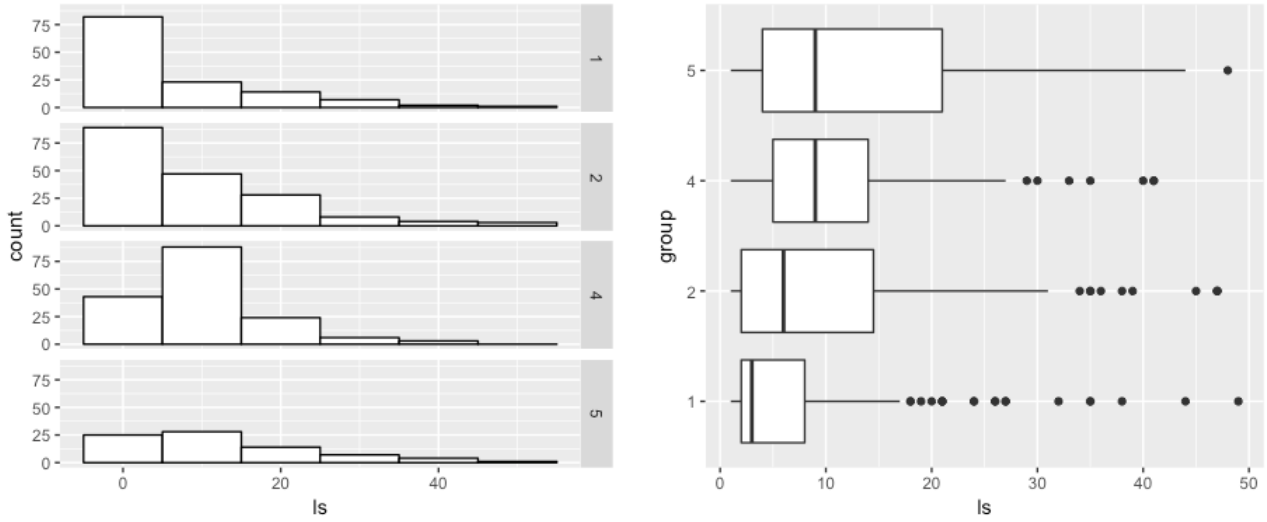
Verburg, Ilona WM, Rebecca Holman, Niels Peek, Ameen Abu-Hanna, and Nicolette F de Keizer. 2017. "Guidelines on Constructing Funnel Plots for Quality Indicators: A Case Study on Mortality in Intensive Care Unit Patients." *Statistical Methods in Medical Research* 27 (11): 3350–66

4. VALIDATING THE LENGTH OF STAY DECISION TREE

INTRODUCTION

The postsurgical length of hospital stay (LoS) risk-adjustment model was built on data from four units (638 patients).[3] The distribution and the boxplot charts of the LoS in the development dataset show wide variations between those units (Figure 4.1).

Figure 4.1. LoS distribution for the 4 units' data used for model development



Initial attempts to model LoS had poor results (mean standard error 55.9 days), even if the data underwent a Poisson transformation. The main issue was that the data distribution suffered from heteroscedasticity (i.e., the variance of the errors is not constant across the range of observations) (Figure 4.2). Thus, two pragmatic decisions were made:

1. To exclude patients with extended LoS (LoS > 50 days). These were 'almost certainly', in the authors' opinion, subsequent to either complications of surgery OR delay in arranging safe domestic return from hospital. As post-operative surgical complications are not 'pre-operative' variables, they are excluded from model design AND as LoS due to domestic needs is often related to hospital-social services issues, extended LoS distorts the data for non-clinical reasons.

2. To limit the linear regression analysis to the first 15 days with improvement in fitting reported (mean standard error 4.8 days).

The following linear regression model resulted after optimising the model to use only the variables independently predicting for increased LoS (Table 4.1). Tumour classification (AJCC v7) was grouped as T1&2 and T3&4. Subsites of anatomical region of the head and neck were grouped as Oral Cavity / Lip, Oropharynx to Larynx and 'Other sites', though as shown, subsite was not included in the final linear regression model. The final inputs into the model were Age, T stage classification, Performance status, Tracheostomy, Scale of surgery and High-risk for saliva egress.

Figure 4.2. Residual distribution showing heteroscedasticity

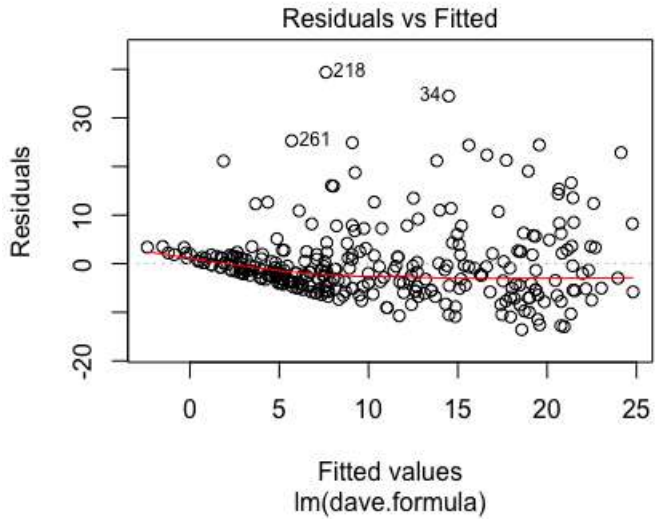


Table 4.1 Linear regression model

Coefficients	Estimate	SD	Error	t	P-value	Reference category
Intercept	-6.9888	2.82555	-2.473	0.01395	<0.05	-
Age	0.10962	0.03654	3	0.00293	<0.01	-
T stage (T3&4)	0.09353	1.10087	0.085	0.93235		-
Performance status 1	1.08614	1.06775	1.017	0.30989		Performance status 0
Performance status 2	2.24747	1.38265	1.625	0.10514		
Performance status 3 & 4	1.6625	2.0878	0.796	0.42651		
Tracheostomy (Yes)	6.0708	1.42193	4.269	0.0000265	<0.001	Tracheostomy (No)
High-risk (Yes)	3.21311	1.23278	2.606	0.00962	<0.01	High-risk (No)
Scale of Surgery 2	3.79137	1.32831	2.854	0.00462	<0.01	Scale of Surgery 1
Scale of Surgery 3	8.85271	1.56941	5.641	0.0000000399	<0.001	
Alcohol Minimal (<14U/week)	-0.89413	1.12228	-0.797	0.42627		No alcohol
Alcohol Moderate (<14U/week)	2.3253	1.47868	1.573	0.1169		
Alcohol Heavy (>40U/week)	2.20704	1.37171	1.609	0.1087		
Alcohol Ex-heavy (previously >40U/week)	3.23868	1.9429	1.667	0.0966		

Multiple R-squared: *0.4452*

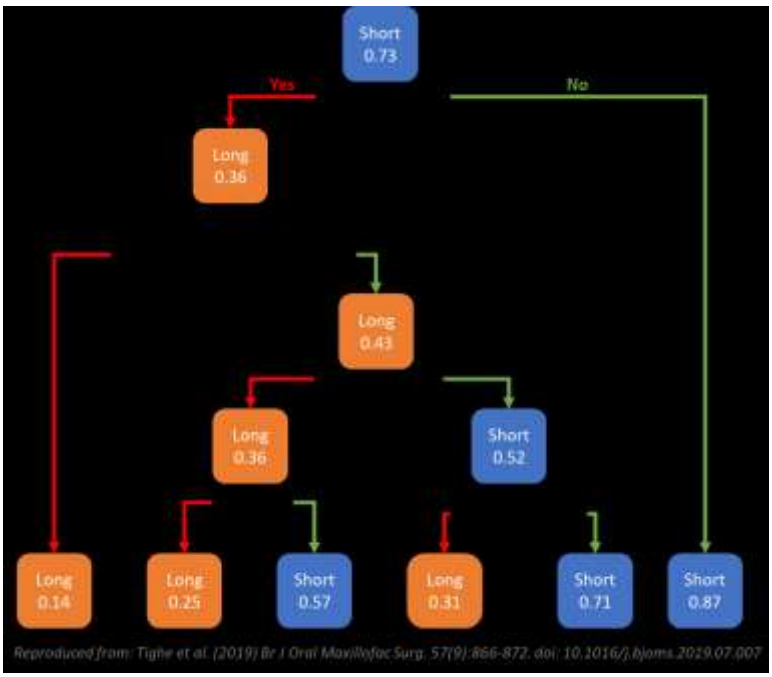
Adjusted R-squared: *0.4206*

F-statistic: *18.09* *On 13 and 293 degrees of freedom (DF)*

p-value: *< 2.2e-16*

There followed an attempt to define, using pre-operative data alone, those patients expected to stay less than or more than 15 days. Based on a 60% training set, and 40% testing set, a decision tree (with an error rate of 0.2) was the champion model chosen in this process (Figure 4.3).

Figure 4.3. Decision tree for < 15 days and ≥ 15 days



Thus, the risk adjustment process for LoS is three simple steps:

1. Exclude patients who had a LoS beyond 50 days
2. Apply the Decision tree model on pre-operative data to predict LoS < 15 days or ≥ 15 days
3. On the cohort predicted to have a short LoS (< 15 days), apply the linear regression equation.

APPLYING THE VERBURG GUIDANCE

Step 1. Define the policy-level decisions

The quality indicator considered here is the rate of short (< 15 days) LoS following head and neck surgery. The benchmark of the quality indicator is the cohort average (i.e., internal benchmark).

Data was extracted from the QOMS Oncology & Reconstruction registry on 31/10/2022 for all cases with a cancer diagnosis, created before 01/07/2022 (to maximise the number of cases and allow for data completion of the outcome variables).

Of the 1160 QOMS cases retrieved, 190 were excluded for being incomplete (n = 172) or having extended LoS (n =18), the final sample size for the analysis of n = 970. Applying the decision tree, 805 and 165 patients were predicted to have a short (< 15 days) or a long (≥ 15 days) LoS, respectively. The confusion matrix is shown below (Table 4.1).

Table 4.1. Confusion matrix for the LoS decision tree

N	Predicted Short LoS	Predicted Long LoS	Total
Observed Short LoS	684	81	*765*
Observed Long LoS	121	84	*205*
Total	*805*	*165*	*970*

The overall accuracy of the model was 79%.

- Sensitivity: 0.84
- Specificity: 0.50
- Positive predictive value: 0.89
- Negative predictive value: 0.41

Calculating the benchmarks

Two sites (OMFS-147 and OMFS-145) had significant missing data and were excluded from this analysis onward. At the hospital level, the prediction accuracy ranged from 44.4% to 96.4% (Table 4.2).

Table 4.2 Results of step of the Verburg guidance when producing funnel plots

Organisations	Observed count	Expected count	Admission count	Observed rate / Proportion	SR	RAR
OMFS-107	29	28	28	1.036	1.036	1.285
OMFS-116	14	9	9	1.556	1.556	1.930
OMFS-120	140	132	122	1.148	1.061	1.316
OMFS-130	341	319	245	1.392	1.069	1.327
OMFS-151	66	77	46	1.435	0.857	1.064
OMFS-157	57	51	46	1.239	1.118	1.387
OMFS-161	279	257	253	1.197	1.086	1.347
OMFS-166	83	76	70	1.186	1.092	1.355
OMFS-20	92	84	76	1.211	1.095	1.359
OMFS-58	100	99	91	1.099	1.010	1.254
OMFS-84	40	39	34	1.176	1.026	1.273

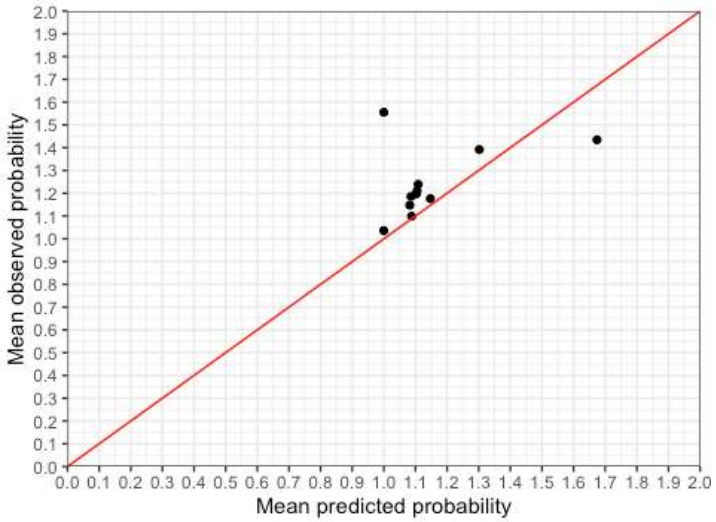
At this point in time, we were unable to calculate the benchmark values. The implementation of the Length of Stay Decision Tree model in REDCap produces a binary (0 or 1) outcome whereas we need the probability calculated in the model to obtain SR and RAR. This will be addressed in the forthcoming year.

Step 2. Check the quality of the risk-adjustment model used

- Brier Score, not calculated
- Scaled Brier Score, not calculated
- C-statistic = 0.68

The model seems to under-estimates the LOS classification for 10 hospitals and over-estimates for one (Figure 4.4).

Figure 4.4 Mean predicted vs. observed rate of complication (recalibration curve) for each hospital



Step 3. Is the number of observations per hospital sufficient? Not computed on this data

Verburg et al (2018) [7] recommended that if fewer than half of the hospitals had enough cases to fulfil this criteria, funnel plots should not be constructed

Funnel plots [9]

Short vs. long LoS rates by participating hospital are showed below in the following funnel plots for crude and risk-adjusted data.

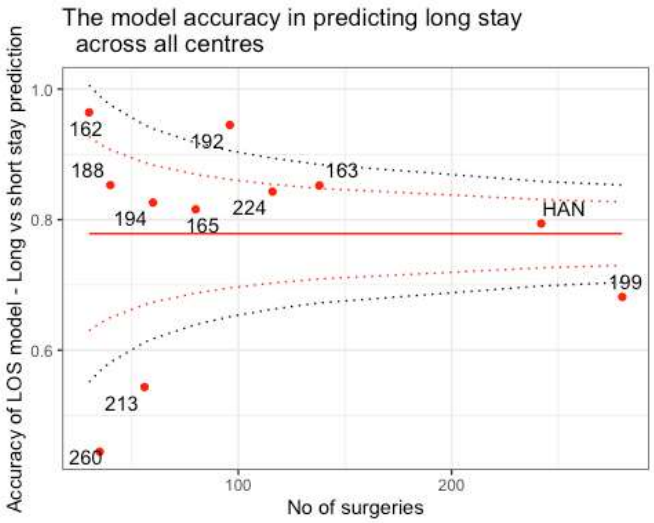


Figure 4.5 Funnel plot for long LoS prediction across the centres (right)

5. VALIDATING THE LINEAR REGRESSION FOR SHORT LENGTH OF STAY

ALL ELIGIBLE CASES IN THE QOMS ONCOLOGY & RECONSTRUCTION REGISTRY

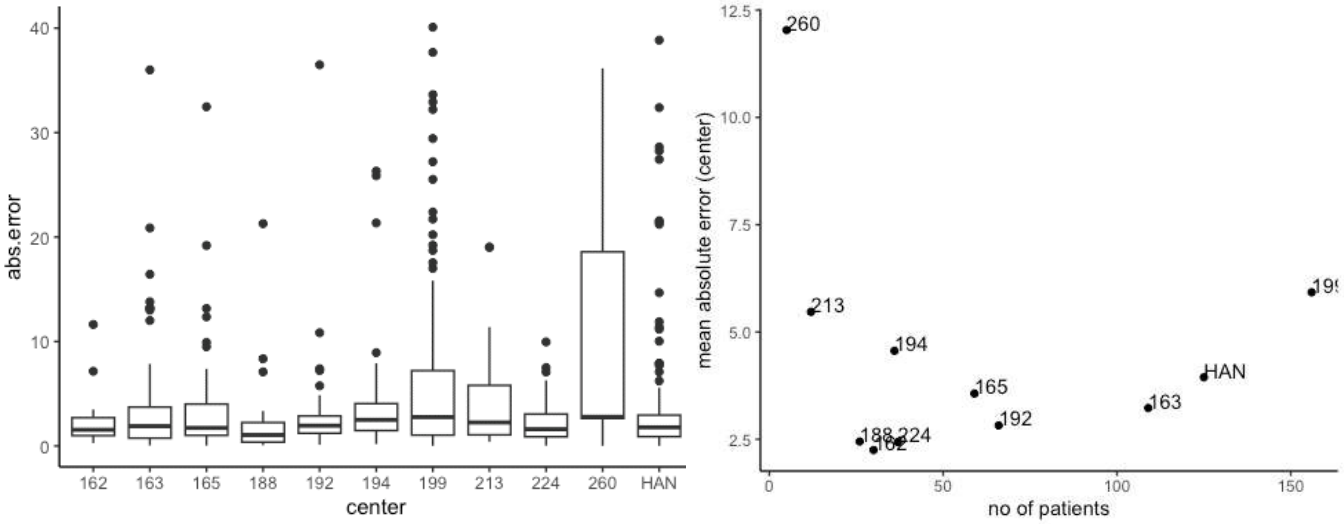
For this stage of the analysis, data were extracted from an updated version of the dataset, after contacting data coordinators, encouraging them to further populate critical fields to boost LoS prediction calculations. Data were extracted from the QOMS Oncology & Reconstruction registry on 22/11/2022. The dataset contained 1216 records from 13 hospitals. Of which, 843 patients were predicted to have a short LOS and LoS was missing for 182 patients, thus leaving 661 observations for the analysis.

When a patient is predicted to have a stay <15 days, the predicted LoS is calculated using the short-LoS linear model and then compared to the observed LoS to assess the performance of the model. The mean absolute error (MAE), root-mean-square error (RMSE) and residual standard error (RSE) for each hospital are showed in Table 5.1. The distribution of the MAE and the relationship between MAE and sample size by participating hospital are showed in Figures 5.1a and 5.1b respectively. The MAE are in proportion to the LoS, i.e., where there is larger spread there is also larger prediction error.

Table 5.1 Mean absolute error, root-mean-square error and residual standard error by participating hospital

Organisations	Frequency (N)	LoS median (days)	MAE (days)	RMSE (days)	RSE
OMFS-107	28	2	2.25	62.99	2.25
OMFS-116	5	11	12.03	60.17	12.03
OMFS-120	108	5	3.23	348.76	3.23
OMFS-130	150	10	5.93	888.77	5.93
OMFS-151	12	7.5	5.47	65.64	5.47
OMFS-157	36	8.5	4.56	164.20	4.56
OMFS161	124	2	3.94	489.15	3.94
OMFS-166	35	8	2.44	85.39	2.44
OMFS-20	59	7	3.57	210.38	3.57
OMFS-58	66	4	2.82	186.38	2.82
OMFS-84	26	3.5	2.45	63.65	2.45

Figure 5.1 a) Left. Boxplot of the mean absolute error by participating hospital and b) Right. Mean absolute error vs. number of cases by participating hospital



ALL ELIGIBLE RECONSTRUCTIVE CASES IN THE QOMS ONCOLOGY & RECONSTRUCTION REGISTRY

A similar analysis to the one above was performed on a subsample containing all eligible reconstructive cases.

Data was extracted from the QOMS Oncology & Reconstruction registry on 22/11/2022. Out of the 1216 existing records (from 13 hospitals), 358 were predicted to have a short LOS with immediate reconstruction with free and pedicled flaps. The predicted LOS was missing for 65 patients, leaving 303 observations in the analysis. The MAE were computed and plotted (Figure 5.3). Finally, we present a box plot of Observed LoS within the immediate construction group, with superimposed median predicted LoS as a graphical way of showing deviation from 'expected' performance (Figure 5.4).

Table 5.2

Organisations	Frequency (N)	LoS median (days)	MAE (days)	RMSE (days)	RSE
OMFS-107	5	5	2.249643	14.95	2.99
OMFS-116	3	11	12.034	38.95	12.98333
OMFS-120	39	10	3.229259	181.13	4.644359
OMFS-130	112	12	5.925133	757.48	6.763214
OMFS-151	6	17.5	5.47	55.5	9.25
OMFS-157	19	11	4.561111	103.73	5.459474
OMFS-161	40	13	3.944758	342.05	8.55125
OMFS-166	24	11	2.439714	70.32	2.93
OMFS-20	26	10	3.565763	142.25	5.471154
OMFS-58	18	10	2.823939	52.57	2.920556
OMFS-84	11	9	2.448077	45.44	4.130909

Figure 5.3 Boxplot of the mean absolute error by participating institution

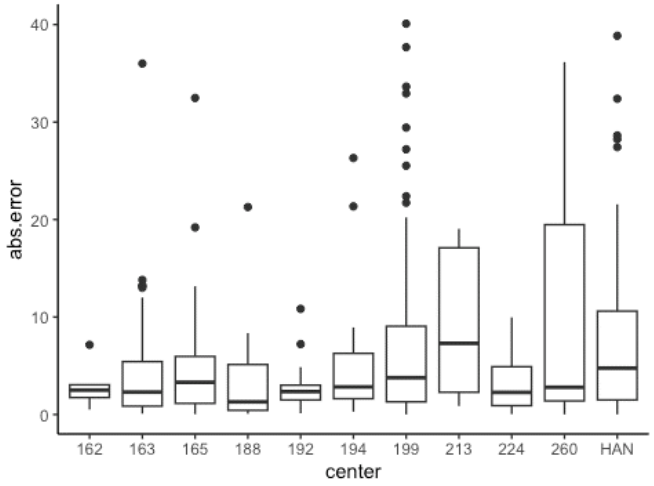
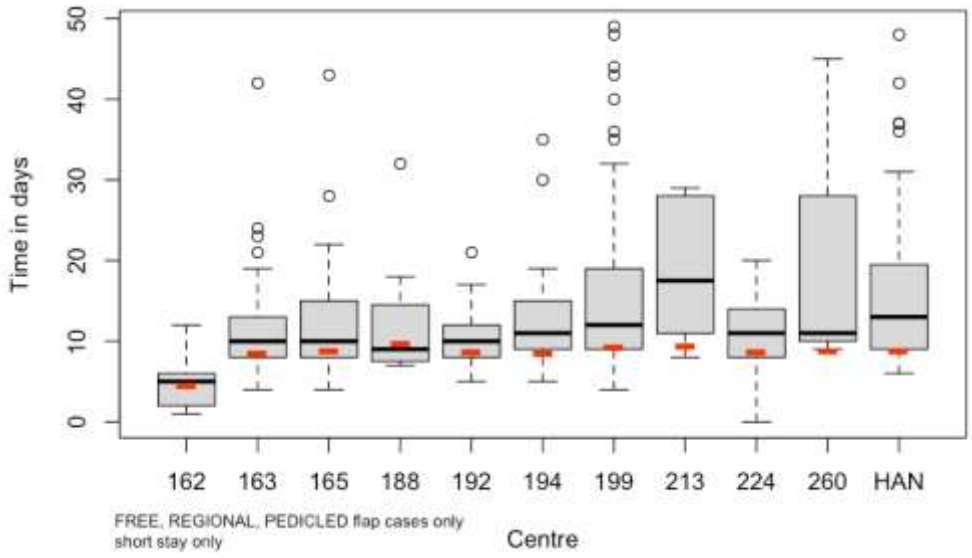


Figure 5.4 Boxplots of actual LoS by participating institution combined with the median predicted LoS (red)



CHAPTER REFERENCES

Tighe, David F, Alan J Thomas, Isabel Sassoon, Robin Kinsman, and Mark McGurk. 2017. "Developing a Risk Stratification Tool for Audit of Outcome After Surgery for Head and Neck Squamous Cell Carcinoma." *Head & Neck* 39 (7): 1357–63.

6. VALIDATING THE ONCOLOGY MARGINS PREDICTION MODEL

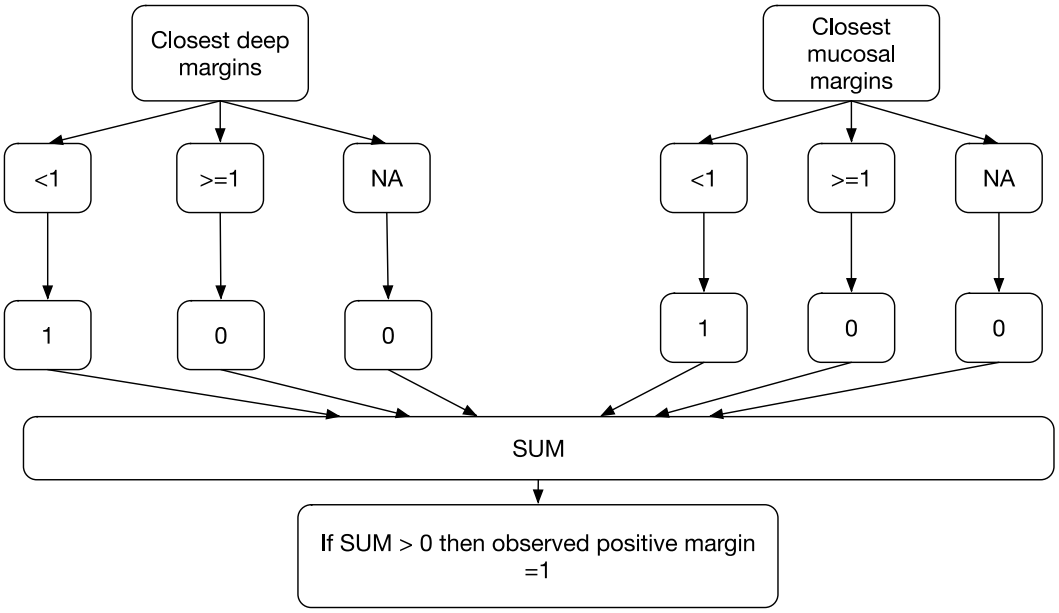
INTRODUCTION

The quality indicator considered here is the rate of positive margins.

For most tumours, the values for deep and mucosal margins should be collected. Therefore, to predict whether a patient is at risk of having positive margins (i.e., <1mm), both needs to be considered and merged into one variable. The following logic was applied (Figure 6.1):

- If either deep OR peripheral mucosal margin is <1mm, then the overall margin status is positive, even if one of the margins is missing.
- If both deep AND peripheral mucosal margins are >1mm, then the overall margin status is "clear".

Figure 6.1 Combined margin process



The model uses a Bayes probability table. The weightings are combined to produce a 'posterior probability' following Bayes formula.[10] The predicted margin status can be calculated for each case depending on pathological T stage classification, presence of extracapsular spread and anatomical subsite of the head and neck tumour, using the probability table below (Table 6.1).

Table 6.1 Bayes probability table

Site	Lip	Oral cavity	Pharynx (inc. tonsils)	Nasopharynx	Hypopharynx	Supraglottis	Larynx	Subglottis	Paranasal sinuses	Neck only	Salivary gland	Other
Margin												
Clear or close	0.00	0.05	0.74	0.07	0.01	0.00	0.00	0.07	0.01	0.03	0.02	0.01
Positive	0.00	0.03	0.58	0.07	0.01	0.01	0.00	0.1	0.01	0.15	0.01	0.02

Stage	T0	T1	T2	T3	T4
Clear or close	0.05	0.36	0.27	0.09	0.23
Positive	0.11	0.17	0.22	0.1	0.41

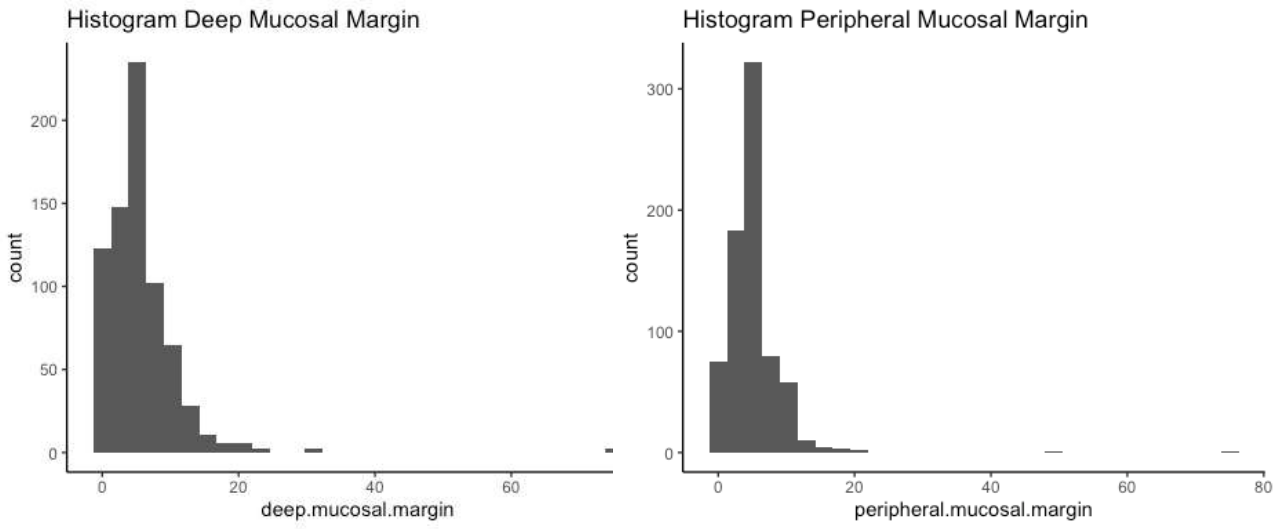
ECS	No ECS	ECS
Clear or close	0.83	0.17
Positive	0.63	0.37

ECS = Extracapsular spread

DISTRIBUTION OF MUCOSAL AND DEEP MARGINS

The distribution of mucosal margins (n = 711) is displayed in Figure 6.2 (left). The size of mucosal margins ranged from 0 to 75mm, with a median of 5mm and an average of 5.6mm. The distribution of deep margins (n = 703) is displayed in Figure 6.3 (right). The size of deep margins ranged from 0 to 75mm, with a median of 5mm and an average of 5.0mm. In fact, three cases of margins over 25mm were challenged with email communication to the data co-ordinators.

Figure 6.2 Distribution of mucosal margins (left) and deep margins (right)



RESULTS OF THE PREDICTION MODEL

The margins prediction model was developed by Tighe D et al. [2]

Data were extracted from the QOMS Oncology & Reconstruction registry on 22/10/22 created before 01/07/2022. Out of the available 1160 records, 495 cases missing either one margin value (and the other one being 'clear') or both and were excluded from the analysis, thus leaving 695 cases. Of 695 cases with complete data 134 (19%) could be classified as 'positive' or <=1mm margins. In future work, this issue around missingness needs to be addressed. The variable predicted risk of positive margins is scaled between 1-30% (Table 6.3).

Table 6.3 Predicted risk of positive margins and their corresponding sample size

Predicted risk	n	Predicted risk	n	Predicted risk	n	Predicted risk	n	Predicted risk	n	Predicted risk	n
4.6	7	**10.1**	8	**13.1**	1	**16.7**	1	**22.1**	2	**28.8**	1
5.7	226	**10.5**	96	**13.7**	1	**17.2**	2	**25.7**	25	**29.6**	1
6.9	13	**11.4**	4	**13.9**	95	**19.4**	3	**26.4**	1		
8.5	157	**11.5**	4	**15.9**	8	**20**	38	**27.9**	4		
8.6	4	**12.5**	3	**16.5**	12	**21.7**	1	**28.5**	1		

APPLYING THE VERBURG GUIDANCE

Step 1. Define the policy-level decisions

The benchmark of the quality indicator is the cohort average (i.e., internal benchmark) of positive margin status (<+1mm). Data was extracted from the QOMS Oncology & Reconstruction registry on 22/10/22 for all cases with oral lip, oral cavity and oropharynx SCC, created before 01/07/2022 (to maximise the number of cases and allow for data completion of the outcome variables).

Calculating the benchmark (Table 6.4)

Table 6.4 Results of step of the Verburg guidance when producing funnel plots

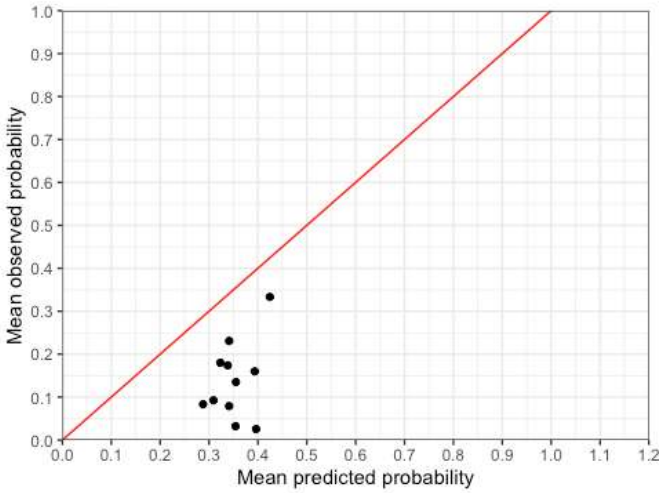
Institutions	Observed count	Expected count	Admission count	Observed rate / Proportion	SR	RAR
162	1	3.453	12	0.083	0.290	0.037
163	21	31.047	91	0.231	0.676	0.087
165	5	16.693	54	0.093	0.300	0.038
188	4	9.837	25	0.160	0.407	0.052
192	8	15.563	46	0.174	0.514	0.066
194	1	15.457	39	0.026	0.065	0.008
199	14	60.380	177	0.079	0.232	0.030
213	1	10.983	31	0.032	0.091	0.012
224	9	16.160	50	0.180	0.557	0.071
260	3	3.820	9	0.333	0.785	0.100
HAN	25	65.673	185	0.135	0.381	0.049

Step 2. Check the quality of the risk-adjustment model used

- Brier Score: 0.1520967
- C statistic: 0.6944473

The recalibration curve indicates that the model over-estimates the rate of positive margins for all hospitals (Figure 6.4).

Figure 6.4 Recalibration curve for the margin status model



Step 3. Estimating minimal sample size

- Minimal sample size for observed rate: n = 255
- Minimal sample size for RAR: n = 770

Following the Verburg's recommendation that "if fewer than half of the hospitals have enough admissions (cases) to fulfil this criterion", the funnel plot should not be constructed. Insufficient hospitals had enough data so the funnel plot was not constructed.[7]

Step 4. Calculating overdispersion

The overdispersion factor estimate is 1.807772. Overdispersion can be detected by dividing the residual deviance by the degrees of freedom. If this quotient is much greater than one, the negative binomial distribution should be used. There is no hard cut-off of "much larger than one", but a rule of thumb is 1.10 or greater is considered large.

CONCLUSION

While the discrimination of the model is satisfactory (AUROC 0.7), the model is over-predicting positive margin status with evidence of over-dispersal that will need addressing when more data is accumulated. Fewer than half the participating hospitals, at the time of the cut-off, had sufficient numbers to present risk-adjusted graphics in the form of a funnel plot. The funnel plot supplied in the QOMS Inaugural Report

demonstrates proof of principle. QOMS will investigate if a new HNSCC model is required in the next annual report.

REFERENCES

Tighe, D, Fabio Fabris, and A Freitas. 2022. "Machine Learning Methods Applied to Audit of Surgical Margins After Curative Surgery for Head and Neck Cancer." *British Journal of Oral and Maxillofacial Surgery* 59 (2): 209–16.

Verburg, Ilona WM, Rebecca Holman, Niels Peek, Ameen Abu-Hanna, and Nicolette F de Keizer. 2018. "Guidelines on Constructing Funnel Plots for Quality Indicators: A Case Study on Mortality in Intensive Care Unit Patients." *Statistical Methods in Medical Research* 27 (11): 3350–66.

7. DESCRIPTION OF THE MODEL FOR FREE FLAP FAILURE

INTRODUCTION

The free flap failure model was developed using a dataset of flap complications after free tissue transfer, collated from eight NHS units. [6]

The model report

The free flap failure model selected was a Deep Forest variant (powerful types of ensemble algorithms). The selected version used Random Undersampling (RUS) integrated in the learning of AdaBoost and a standard Random Forest algorithm as the base ensemble learner of the boosting ensemble (i.e., DF(RUSBoost)). RUA is a Machine Learning technique used to balance datasets in which the event of interest is rare (<5%). Other ways of boosting the proportion of the 'minority class' like random oversampling and synthetic minority oversampling technique (SMOTE) were also tested in the development stage of the work. ADABoost and Random Forest combine other aspects of Machine Learning and the reader can refer to "Deep Forest RUSBoost" for additional information.

This configuration was implemented using DF21 [11,12], Scikit-Learn [13], and Imbalanced-Learn. [14]

Figure 7.1 shows the final configuration of the DF(Rusboost) architecture design. It is composed of four RUSBoost per layer, all containing Random Forest as boosting base learner. The model takes account of 20 inputs concerning patient, tumour and surgical details that contribute to risk. These inputs are re-coded in 'one hot encoding.' The feature importance graphic demonstrates the relative importance of these inputs (Figure 7.2).

The results of the confusion matrix and 10-fold cross-validation of the DF(RusBoost) configuration and the macro averages of precision, recall and F1-scores and AUROC of the resulting model are shown in Tables 7.1 and 7.2, respectively. Figure 7.2 shows the calibration plot of the model on test-data, demonstrating acceptable performance (Hosmer-Lemershow Goodness of fit $\chi^2 = 2\ 6.9$, $p = 0.53$).

Figure 7.1. Final configuration – Architecture

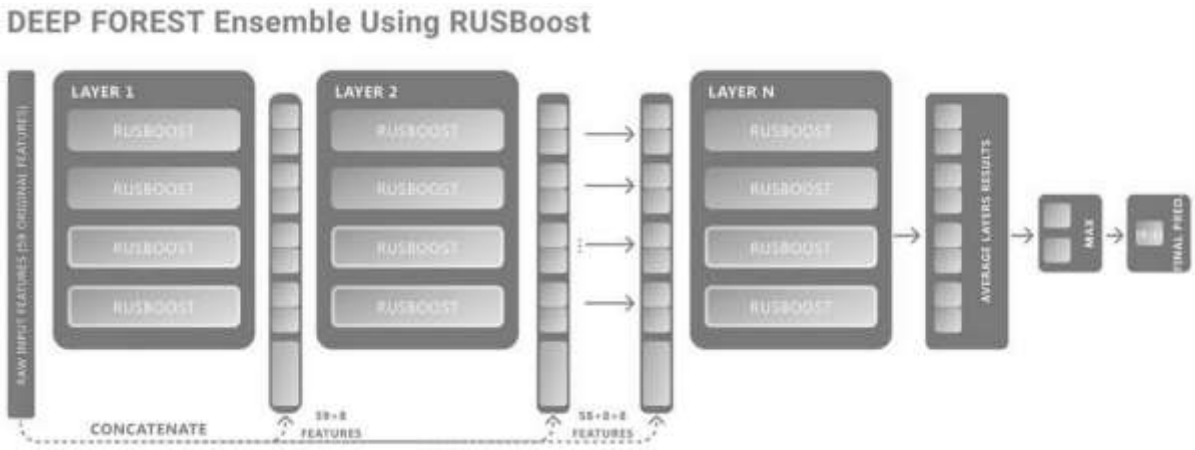


Table 7.1 Confusion matrix of the 10-fold cross-validation results for the best classification model, considering the minority class ('complete flap failure') as the positive class.

		Predicted flap outcome	
		Negative	Positive
True flap outcome	Negative	860	658
	Positive	34	41

Table 7.2 Predictive accuracy results obtained with 10-fold cross-validation for the selected model

Macro AVG Precision	Macro AVG Recall	Macro AVG F1-score	AUROC Score
0.51	0.56	0.53	0.655

Figure 7.2 Calibration plot for observed vs. predicted free flap outcome status (0 = flap success and 1 = flap failure) – Next page

Figure 7.3 Calibration plot for observed vs. predicted free flap outcome status (0 = flap success and 1 = flap failure)

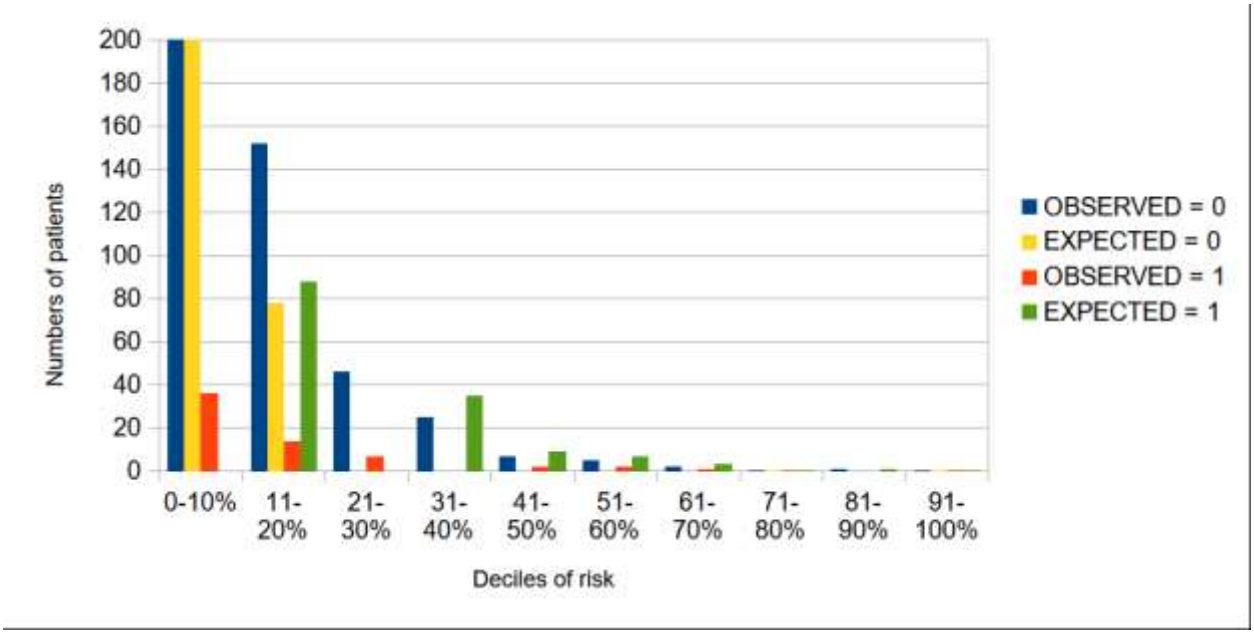
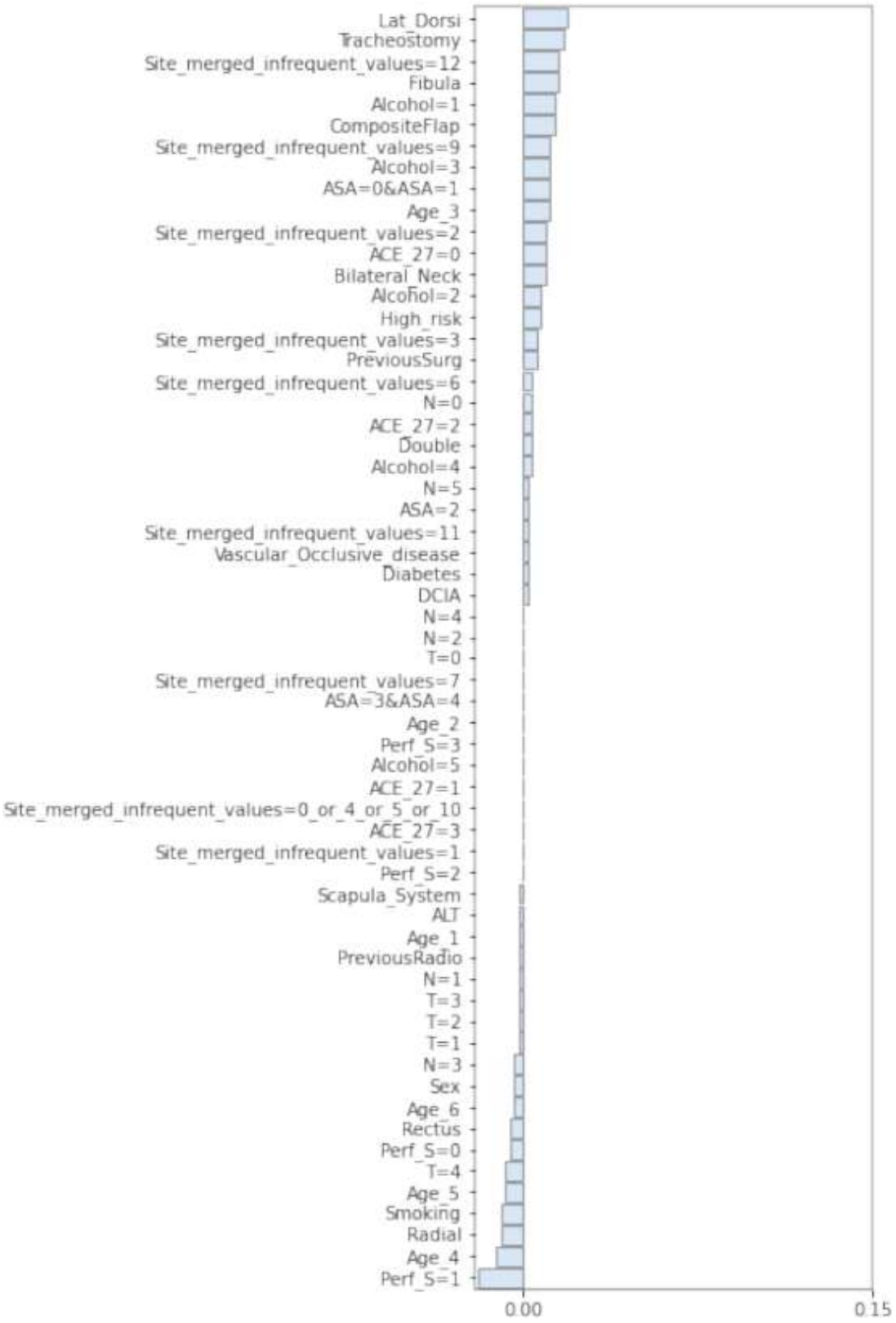


Figure 7.2 Variables inputted in the model and their respective relative importance



Implementation of the free flap failure model

The model was embedded within a Python environment. The model processes free flap care episode data, received from REDCap with all relevant input variables, and outputs a risk (probability of free flap failure) back into REDCap, combined with the observed outcome. The observed outcome and risk adjusted are plotted in a Cumulative Sum (CuSUM) chart.

A bespoke CuSUM module designed by Queen Mary, University London (QMUL) staff, has been embedded into REDCap. With QMUL's assistance to extract free flap care episode data on a monthly basis, QOMS can present updatable unit- and national-level free-flap success/failure data displayed against time in the CuSUM chart. This is done using Plotly visualisation graphics.[15] The graphical presentation of free flap outcome data has options the user can refine (e.g., presenting non risk-adjusted or risk-adjusted CuSUM charts) (Figure 7.3).

When non risk-adjusted graphical display of outcome data is selected, the default increments reflect the failure rate of 4.7% (95.3% penalty for failure, 4.7% reward for success). When the risk-adjusted graphical display is selected, the baseline increments are adjusted to plot the function:

(7.1) $X_t = \max(0, X_{t-1}, W_t)\,, t = 1, 2, 3, \ldots$

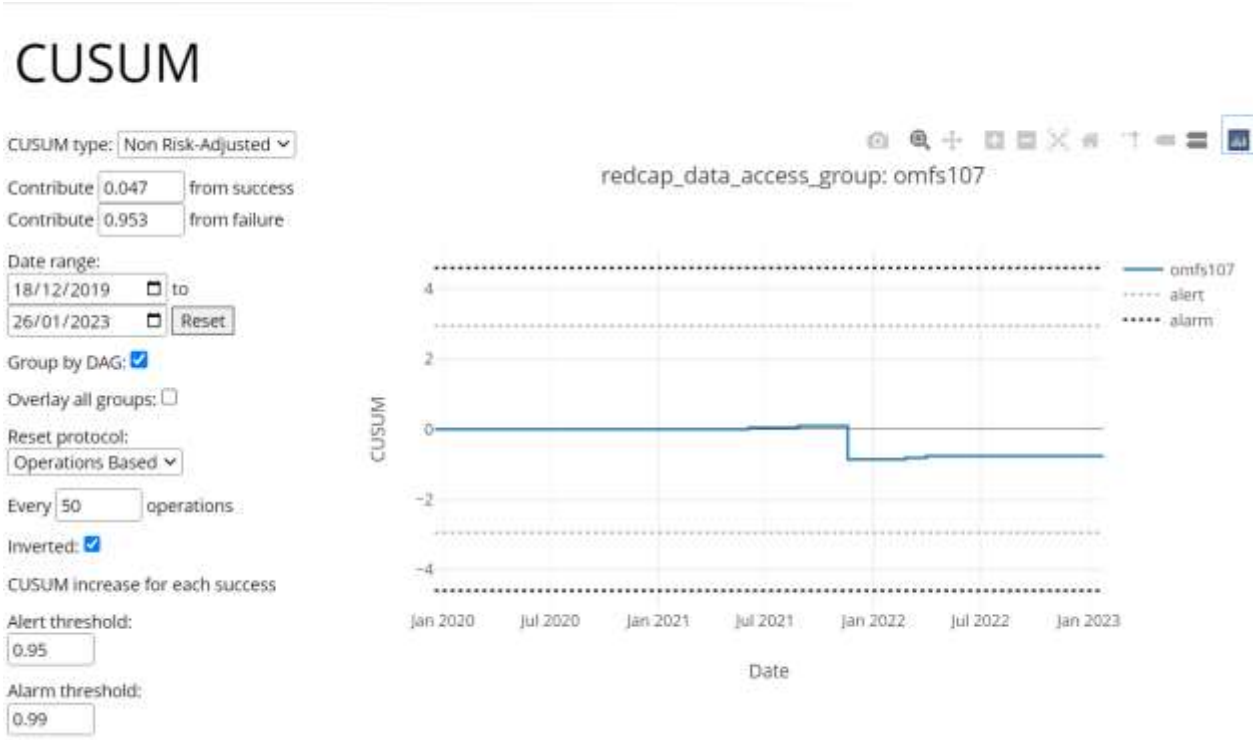
where Wt is a weight assigned to each value of t. In this study, the risk-adjusted CUSUM charts are updated on a patient-to-patient basis, i.e., each value of t corresponds to a new admitted patient. Consequently, the weights Wt are given by:

(7.2) $W_t = Y_t log(RA) - \log(1 - p_t + RAp_t)$

Where, Y_t is the outcome of patient t (free flap failure within 30 days of operation date yes/no), W_t is the expected probability of the outcome estimated from a prediction model based on data from a reference period and RA >1 is a specified Odds Ratio (OR) increase in the outcome rate, as compared to the reference period, that the risk-adjusted CUSUM chart is set to detect. We set it at 2 (or twice the expected rate).

The weight, W_t, is set as positive if the patient did not have the outcome, and negative if they did and its absolute value is large if the outcome is unexpected. Thus, if more patients had free tissue failure than predicted, the CuSUM function would decrease (Figure 7.4).

Figure 7.4 A snapshot of a single DAG's CuSUM chart *(see next page for instructions on how to use the CuSUM module in REDCap)*



VALIDATING THE MODEL FOR FREE FLAP FAILURE

The quality indicator considered here is the rate of free flap failure. Data was extracted from the QOMS Oncology & Reconstruction registry on 22/10/22 created before 01/07/2022.

The overall complete flap failure rate was 72/1159 (4.7%). The partial flap failure rate was lower (2%). There was substantial variation in the coding of partial flap failure, and each had different implications to the patient; because of the relative rarity we decided to not model partial flap failure at this stage. Further, a more clinically useful classification has since been published that looking forward, will improve the coding of partial flap failure within QOMS. [16]

External validation of the free flap model had not been attempted at the time of writing. We will aim to publish an external validation of the free flap model in 2023.

How to use the CuSUM module in REDCap

Using the following interfaces on the left-hand side to adjust the plots:

1. CuSUM type
    a. Select between a regular CuSUM chart or a risk-adjusted CuSUM chart. For risk-adjusted CuSUM, the risk for each operation must be provided.
2. Contribution from success or failure
    a. For the regular CuSUM chart, this adjusts how much the CuSUM chart increases for a failed operation or decreases for a successful operation.
    b. Doe risk-adjusted CuSUM, choose the Odds scale*: set the hypothesis to test if the failure odds ratio is greater than a multiple of the odds ratio expected – this is by default set to 2.
3. Group by DAG
    a. Plot a CuSUM for each DAG (i.e. hospital).
    b. Select "Overlay all groups" to plot in addition to the CuSUM using all DAGs, i.e. national data
4. Reset protocol: reset the CuSUM chart to zero using different protocols:
    a. Non-negative: The CuSUM chart can never be negative
    b. Operation-based: reset the CuSUM chart after a certain number of operations (e.g. 50)
    c. Time-based: reset the CuSUM chart after a certain number of days, requires the operation dates to be provided
5. Inverted: Select this to flip the graph vertically. The default is increment up for 'success', increment' down for failure.
6. Alert and alarm threshold: set the significant level for the alert and alarm lines. Default is set at 2SD and 3SD

8. MODEL FOR NMSC MARGINS

DEVELOPMENT OF THE NMSC MARGIN MODEL

The NMSC margins model was developed with the discreet purpose of 'shrinking' the NMSC dataset to the minimum size necessary to generate risk-adjustment on a parsimonious set of inputs. [2]

A pathology output of consecutive histology reports for a period of three years was requested from three oral and maxillofacial units in the southeast of England. A total of 3354 cases were retrieved and analysed. The dependent variable was a deep margin with peripheral margin clearance at the 0.5 mm threshold. Predictive models, accounting for patient and tumour factors, were developed using automated machine learning (Auto-ML) methods.

Five independent Auto-ML (five-fold cross validation) studies were run with the same dataset for each of the Auto-ML optimisation metric, namely AUROC, F1 score, and Recall. The results for each performance indicator are provided as macro average (i.e., average of the five runs). Machine learning are resource intensive, and the present analysis required 75 hours of Auto-ML analysis when all optimisation metrics are considered (3 metrics x 5-fold cross-validation x 5 hours each = 75 hours). The performance metrics for each Auto-ML analysis were gathered to determine which one was most suitable for handling unobserved data and this led us to concentrate on the winner pipeline to locate the optimal strategy to apply to unseen data.

As soon as a pipeline was picked, a thorough examination of its results was conducted. The algorithm more frequently chosen by the Auto-ML system in the five-fold cross-validation step was selected. Finally, we had to choose a method for each phase of a machine learning model, i.e., Imputation, Rescaling, and Data preprocessor, to apply to the final model's construction phase. Given the "Balancing Strategy" phase for instance, we used the same strategy as for the algorithm to choose which method to select (e.g., if "weighting" or another technique is more significant, it will be selected for the final pipeline). Upon completion, we were able to consider the champion model and its pipeline's phase.

Each column in Tables 8.1a-c is labelled with the cross-validation number, and each row describes the Auto-ML-selected pipeline characteristics (e.g., imputation method, algorithm selection, etc). The final column in each of the subsequent tables reflects the mean of the classification report metrics chosen for N (i.e., N=5) fold validation, such as Precision, Recall, F1, and AUROC. The pipelines were assessed with optimisation strategies prioritising Recall Macro-Average, F1 Score and AUROC, respectively.

Table 8.1 a) Averaged 5-fold cross validation results with Recall-macro average as Auto-ML optimisation metric, b) Averaged 5-fold cross validation results with F1-Macro Average as Auto-ML optimisation metric, and c) Averaged 5-fold cross validation results with AUROC as Auto-ML optimisation metric

(a) Recall	1	2	3	4	5	Average
Balancing Strategy	Weighting	Weighting	Weighting	Weighting	Weighting	
Category Coalescence	No coalescence	Minority coalescer	No coalescence	Minority coalescer	No coalescence	
Imputation	Mean	Mean	Mean	Mean	Mean	
Rescaling	Power Transformer	Power Transformer	Standardize	Quantile Transformer	Quantile Transformer	
Preprocessor	KPCA	Fast ICA	KPCA	Fast ICA	KPCA	
Classifier	SGD	Passive Aggressive	SGD	Passive Aggressive	LibLinear SVC	
Precision	0.337	0.303	0.330	0.318	**0.337**	0.325
Recall	0.598	0.590	0.614	0.647	**0.721**	0.634
F1-SCORE	0.431	0.401	0.429	0.427	**0.459**	0.429
AUROC	0.70	0.688	0.712	0.693	**0.739**	0.7

(b) F1	1	2	3	4	5	Average
Balancing Strategy	Weighting	Weighting	Weighting	Weighting	None	
Category Coalescence	Minority coalescer	No coalescence	Minority coalescer	Minority coalescer	Minority coalescer	
Imputation	Mean	Mean	Mean	Mean	Mean	
Rescaling	Power Transformer	Min Max	Standardize	Power Transformer	Quantile Transformer	
Preprocessor	Fast ICA	Select Rates Classification	Random Trees Embedding	Fast ICA	Random Trees Embedding	
Classifier	Passive Aggressive	Extra Trees	SGD	Extra Trees	Bernoulli NB	
Precision	0.310	0.319	0.324	**0.351**	0.329	0.326
Recall	0.491	0.557	0.598	0.581	**0.663**	0.578
F1-SCORE	0.380	0.405	0.420	0.438	**0.440**	0.416
AUROC	0.688	0.672	0.712	0.699	**0.745**	0.7

(c) AUROC	1	2	3	4	5	Aver-age
Balancing Strategy	None	Weighting	None	Weighting	None	
Category Coalescence	No coalescense	Minority coalescer	Minority coalescer	No coalescense	Minority coalescer	
Imputation	Mean	Mean	Median	Mean	Median	
Rescaling	Quantile Transformer	Normalize	Normalize	Robust Scaler	None	
Preprocessor	Fast ICA	KPCA	SPC	Fast ICA	Random Trees Embedding	
Classifier	LDA	Bernoulli NB	Liblinear SVC	MLP	Multinomial NB	
Precision	0.5	0.389	1	0	**0.527**	0.483
Recall	0.221	**0.360**	0.008	0	0.319	0.181
F1-SCORE	0.306	0.374	0.016	0	**0.397**	0.218
AUROC	0.69	0.66	0.6	0.3	**0.756**	0.6

For every given performance metric, seen as rows from row 6, bold values are the winners for this one (i.e., by metrics we mean first column, by values we mean every float-precision number across the table or classifier's name as an exception for "value")

Abbreviations used in Tables 8.1a-c: PCA = Principal Component Analysis, Kernel PCA = Kernel Principal Component Analysis, Fast ICA = Fast Independent Component Analysis, SGD = Stochastic Gradient Descent, LibLinear SVC = Linear Support Vector Machine, Bernoulli NB = Bernoulli Naive Bayes, MLP = Multi-Layer Perceptron, Multinomial NB = Multinomial Naive Bayes, QDA = Quadratic Discriminant Analysis, SPC = Select Percentile Classification, ICA = independent component analysis, LSVC = Linear Support Vector Classification, and LDA = Linear discriminant analysis.

Auto-ML analyses comparison

Table 8.2 All three optimisation Auto-ML metric analyses average comparison

Auto-ML analysis with optimisation metric	AVG Precision	AVG Recall	AVG F1	AVG AUROC
Recall Macro-Average	0.325	**0.634**	**0.429**	**0.7**
F1 Macro-Average	0.326	0.578	0.416	0.7
AUROC	**0.483**	0.181	0.218	0.6

For each analysis, the performance metric winners are highlighted in bold (by analyses we mean the first column, by performance metric the second-to-last column, and by values we mean every float-precision integer in the table). AVG= Average.

Champion model *(Imputation, Rescaling, Data preprocessor, Classifier)*

Following the method's section, Table 8.2 shows that the analysis utilising Recall-Macro Average as the Auto-ML optimisation metric appears to be the most effective compared to the others. Theoretically, F1-macro should have been our primary emphasis, but the Auto-ML search optimizer did not identify a way to maximise F1 better than the recall-macro average analysis did by maximising the recall measure alone

resulting in a better F1 overall. Unfortunately, AUROC appears incapable of optimising anything, and a glance at it (Table 8.1c) reveals that it is not an optimisation target worth focusing on.

Given that we selected to concentrate on the Recall Macro Average optimisation metric analysis, Table 8.1a contains the selection of two distinct algorithms, namely SGD and Passive Aggressive. To differentiate, we investigate classification metrics such as Precision, Recall, F1-Score, and AUROC in greater depth. As a final result, SGD is more consistent.

We use a similar approach for the balancing strategy, category coalescence, imputation, and rescaling, as well as the pre-processing phase, while constructing the final Scikit pipeline using SGD as the champion classifier.

The final champion model is represented in Figure 8.1a.

Figure. 8.1

a)

b) SCC

c) BCC

It is necessary to read the schema from left to right. It begins with the initial input and successively proceeds to the classifier selection, passing through an imputer phase, rescaling phase, and pre-processing phase. The small tags at the bottom of a particular phase represent its hyper parameter. At the conclusion, the model is constructed and ready for use, for instance with Python.

RESULTS FOR SCC SEPARATE DATASET

Dataset characteristics

Table 8.3 Basic characteristics of SCC separate dataset

	Value
Number of Features	18
Number of Subjects	980
Number of Classes	2
Number of subject class 1	~79%
Number of subject class 0	~20%

Table 8.4 Percentage of missingness per feature with at least >0 of missing values

Feature's name	% of missingness
Tumour thickness	8
Perineural	1
Lymphovascular	3
Grade	0.51
Level of Invasion	5
High or low	55
Maximum dimension	4

Auto-ML result

Table 8.5 Averaged 5-fold cross validation results SCC separate dataset

	1	2	3	4	5	Average
Balancing Strategy	Weighting	None	None	None	None	
Category coalescence	Minority coalescer	Minority coalescer	None	Minority Coalescer	None	
Imputation	Most Frequent	Mean	Mean	mean	Mean	
Rescaling	Quantile Transformer	none	Power Transformer	None	Quantile Transformer	
Preprocessor	FastICA	KPCA	Polynomial	FastICA	SPC	
Classifier	SGD	Passive Aggressive	SGD	QDA	SGD	
Recall	0.85	0.22	0.65	0.42	0.14	0.436
Precision	0.3	0.52	0.42	0.5	0.6	0.468
F1-Score	0.44	0.31	0.51	0.39	0.23	0.376
AUROC	0.75	0.68	0.76	0.74	0.62	0.71

Champion model (Imputation, Rescaling, Data Preprocessor, Classifier)

Occasionally, we made exceptions. For instance, the winning preprocessor is Fast ICA regarding the (Table 8.5) but after a few trials Select Percentile (SPC) proved to be more responsive to unseen data (see other report), thus we chose to finalise the champion model as described in Figure 8.1b.

RESULTS FOR BCC SEPARATE DATASET

Dataset characteristics

Table 8.6 Basic characteristics of BCC separate dataset

	Value
Number of Features	20
Number of Subjects	2309
Number of Classes	2
Number of subject class 1	~82%
Number of subject class 0	~17%

Table 8.7: Percentage of missingness per feature with at least >0 of missing values

Feature's name	% of missingness
Gender	0.09
Tumour Thickness	64
Level Of Invasion	60
High or Low	72
Maximum Dimension	11

Auto-ML result

Table 8.8: Averaged 5-fold cross validation results BCC separate dataset

	1	2	3	4	5	Average
Balancing Strategy	Weighting	None	Weighting	None	Weighting	
Category coalescence	None	Minority coalescer	Minority coalescer	Minority coalescer	None	N/A
Imputation	mean	mean	mean	mean	median	
Rescaling	none	Power Transformer	Quantile Transformer	standardize	minmax	
Preprocessor	FastICA	PCA	KPCA	FastICA	SPC	
Classifier	Random Forest	Passive Aggressive	LDA	Bernoulli NB	LSVC	
Recall	0.2	0.43	0	0.13	0.34	**0.22**
Precision	0.36	0.4	0	0.28	0.36	**0.28**
F1-Score	0.26	0.42	0	0.18	0.35	**0.242**
AUROC	0.67	0.7	0.67	0.62	0.69	**0.67**

Champion model (Imputation, Rescaling, Data Preprocessor, Classifier)

Occasionally, we made exceptions. For instance, there is not winning rescaler, but we found out that Power Trans- former, regarding Table (8.9) proved to be more responsive to unseen data (see other report). Another exception is that no winning algorithms were identified, however Passive Aggressive appeared to be the least damaging fold, so we selected its classifier. Finally, we decided to finalise the champion model as described in Figure 8.1c.

Confusion matrix

The confusion matrix of the champion models when run on the development dataset (n=3545) for BCC and SCC margin prediction are displayed in Table 8.9a and b, respectively.

Table 8.9 Confusion matrices for the best classifiers found by Auto-sklearn, (a) namely the Passive Aggressive classifier on the BCC dataset and (b) the Stochastic Gradient Descent classifier on the SCC dataset.

(a)		Predicted class label	
		0	1
True class label	0	36	46
	1	52	321

(b)		Predicted class label	
		0	1
True class label	0	66	55
	1	117	350

The predicted probabilities for all cases of BCC and SCC were calculated using the formula below:

(8.1) $$Risk_{adjusted}PMR = \frac{mean\ PMR}{expected\ PMR\ for\ unit} \times observed\ PMR\ for\ unit$$

where PMR = positive margin rate.

APPLYING BCC AND SCC MODELS TO QOMS DATA

The BCC and SCC models were applied to QOMS skin data. The BCC dataset contained 345 cases and the SCC dataset contained 135 cases with complete records on which to calculate a probability of a positive margin at the <0.51mm threshold. (Tables 8.10 and 8.11). As the equation above describes, the risk adjusted positive margin rate was predicted for each unit.

Table 8.10 Risk adjusted rates of <=0.51mm after excision of Basal Cell Carcinoma

Cohort	20%	60	345	10%	27%
Organisation	**Raw <0.51mm Margin**	**Numera-tor**	**Denomina-tor**	**Predicted <0.51mm Margin**	**Risk-adjusted <0.51mm margin**
OMFS-107	6%	4	68	13%	5%
OMFS-130	20%	7	35	6%	36%
OMFS-157	11%	11	96	3%	38%
OMFS-28	32%	6	19	11%	31%
OMFS-58	41%	18	44	27%	15%
OMFS-84	20%	9	45	4%	46%
OMFS-94	13%	5	38	8%	17%

Table 8.11: Risk adjusted rates of <=0.51mm after excision of Squamous Cell Carcinoma

Cohort	26%	35	135	41%	25%
Organisation	**Raw <0.51mm Margin**	**Numera-tor**	**Denomina-tor**	**Predicted <0.51mm Margin**	**Risk-adjusted <0.51mm margin**
OMFS-107	11%	2	18	28%	17%
OMFS-130	18%	2	11	55%	14%
OMFS-157	11%	3	27	44%	10%
OMFS-28	33%	1	3	=0/3	#VALUE!
OMFS-58	61%	14	23	48%	53%
OMFS-84	33%	8	24	37%	37%
OMFS-94	17%	5	29	37%	19%

VALIDATION OF BCC AND SCC MODELS

Validation work on the NMSC margin dataset will be completed in 2024 when 2 further cycles (estimated 1000 further cases are available).

REFERENCES

1. Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C. & Van Diepen, M. External validation of prognostic models: what, why, how, when and where? *Clinical Kidney Journal* **14**, 49–58 (2021).
2. Tighe, D. *et al.* Machine learning methods applied to audit of surgical margins after curative surgery for facial non-melanoma skin cancer. *Brit J Oral Maxillo surg* (2022).
3. Tighe, D., Sassoon, I., Hills, A. & Quadros, R. Case-mix adjustment in audit of length of hospital stay in patients operated on for cancer of the head and neck. *British Journal of Oral and Maxillofacial Surgery* **57**, 866–872 (2019).
4. Tighe, D. F., Thomas, A. J., 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* **39**, 1357–1363 (2017).
5. Tighe, D., Lewis-Morris, T. & Freitas, A. Machine learning methods applied to audit of surgical outcomes after treatment for cancer of the head and neck. *British Journal of Oral and Maxillofacial Surgery* **57**, 771–777 (2019).
6. Tighe, D. *et al.* Machine Learning methods applied to risk adjustment of Cumulative Sum chart methodology to audit free flap outcomes after Head and Neck Surgery. *British Journal of Oral and Maxillofacial Surgery* S0266435622002698 (2022) doi:10.1016/j.bjoms.2022.09.007.
7. Verburg, I. W., Holman, R., Peek, N., Abu-Hanna, A. & de Keizer, N. F. Guidelines on constructing funnel plots for quality indicators: A case study on mortality in intensive care unit patients. *Statistical Methods in Medical Research* 096228021770016 (2017) doi:10.1177/0962280217700169.
8. Spiegelhalter, D. J. Handling over-dispersion of performance indicators. *Quality and Safety in Health Care* **14**, 347–351 (2005).
9. Spiegelhalter, D. J. Funnel plots for comparing institutional performance. *Statist. Med.* **24**, 1185–1202 (2005).
10. Witten, I. H., Frank, E. & Hall, M. A. *Data mining: practical machine learning tools and techniques*. (Morgan Kaufmann, 2011).
11. Zhou, Z.-H. & Feng, J. Deep Forest: Towards An Alternative to Deep Neural Networks. in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* 3553–3559 (International Joint Conferences on Artificial Intelligence Organization, 2017). doi:10.24963/ijcai.2017/497.
12. Zhou, Z.-H. & Feng, J. Deep forest. *National Science Review* **6**, 74–86 (2019).
13. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
14. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* **18**, 1–5 (2017).
15. Plotly: Low-Code Data App Development. https://plotly.com/.
16. Ho, M. W. *et al.* Results of flap reconstruction: categorisation to reflect outcomes and process in the management of head and neck defects. *British Journal of Oral and Maxillofacial Surgery* S0266435619303213 (2019) doi:10.1016/j.bjoms.2019.08.005.