Appendix 1: Patient data

Appendix table 1 summarises the patient data used in the study.

Appendix table 1: Clinical trials making up the dataset.

Trial	Patients available	Primary disease site	Radiotherapy technique	Radiotherapy dose-fractionation	Concurrent chemotherapy
COSTAR (Phase III, multicentre; CRUK/08/004)	78	Parotid gland	Unilateral; conventional, IMRT	65 Gy / 30 # (definitive RT), 60 Gy / 30 # (post-operative RT)	No
PARSPORT (Phase III, multicentre; CRUK/03/005) [1]	71	Oropharynx, hypopharynx	Bilateral; conventional IMRT	65 Gy / 30 # (definitive RT), 60 Gy / 30 # (post-operative RT)	No
Dose Escalation (Phase II, single centre) [2]	30	Larynx, hypopharynx	Bilateral; IMRT	67.2 Gy / 28 #, 63 Gy / 28 #	Yes
Midline (Phase II, single centre) [3]	117	Oropharynx	Bilateral; IMRT	65 Gy / 30 # (definitive RT), 60 Gy / 30 # (post-operative RT)	Yes
Nasopharynx (Phase II, single centre; NCT02149641)	36	Nasopharynx	Bilateral; IMRT	65 Gy / 30 # (definitive RT), 60 Gy / 30 # (post-operative RT)	Yes
Unknown Primary (Phase II, single centre; NCT02112344)	19	Unknown primary	Bilateral; IMRT	65 Gy / 30 # (definitive RT), 60 Gy / 30 # (post-operative RT)	Yes

Detailed descriptions of the patients included and treatment planning and delivery protocols can be seen in the trial references. IMRT - intensity-modulated radiotherapy; # - fractions; RT – radiotherapy; Unilateral – treatment delivered to ipsilateral parotid bed only; Bilateral – treatment delivered to ipsilateral and contralateral mucosa of relevant subsite (e.g. nasopharynx, oropharynx or larynx).

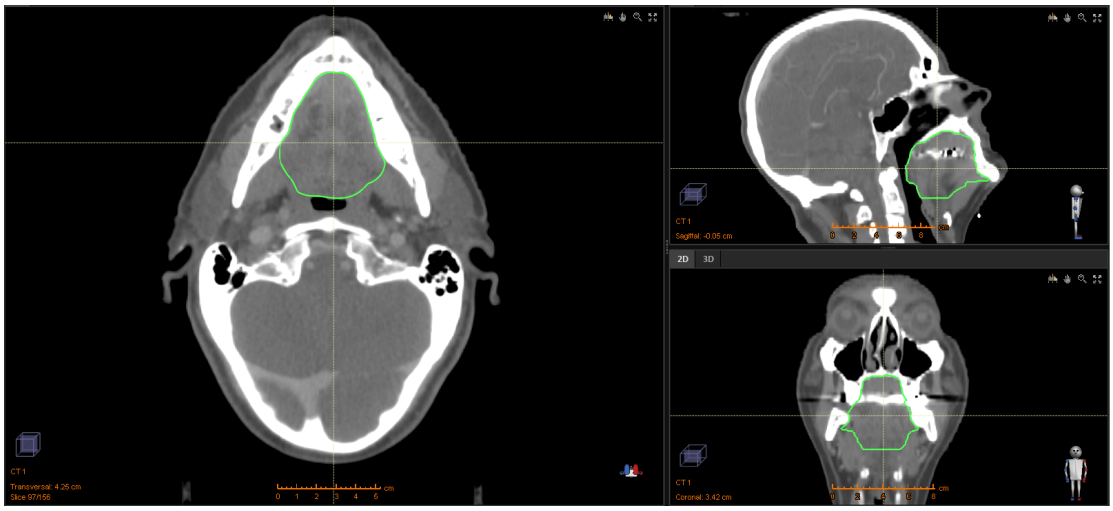
Unilateral versus bilateral irradiation was not explicitly included as a covariate in the models since it correlates perfectly with parotid gland primary disease site.

The radiotherapy treatment planning techniques are detailed in [1,4,5] and the COSTAR trial (CRUK/08/004) protocol. No oral cavity dose constraints were used in treatment planning.

Concurrent chemotherapy was administered in two cycles, on day 1 and day 29 of radiotherapy.

Appendix 2: Oral cavity contouring

Appendix figure 1 shows an example of the oral cavity contouring approach used.



Appendix figure 1: Axial (left), sagittal (top right) and coronal (bottom right) views of an example of the oral cavity structure used.

Appendix 3: In house software

Software for reading in the DICOM RT data and calculating the radiotherapy dose metrics was developed using the Python (version 2.7.9) programming language [6] and the NumPy (version 1.9.2) [7], SciPy (version 0.16.1) [7], Matplotlib (version 1.4.3) [8], Seaborn (version 0.6.0) [9] and PyDicom (version 0.9.9) [10] modules to allow novel dose metrics (which are not calculated by any of the treatment planning systems) to be calculated from the dose distributions. Statistical analysis was performed using the Pandas (version 0.17.1) [11] and scikit-learn version (0.17.1) [12] Python modules.

Appendix 4: Hyper-parameter tuning and internal validation

Model hyper-parameter tuning was carried out using a 100 iteration shuffled stratified cross-validated (with a train/test split of 80/20) grid-search. The grid-search method involves fitting models (including covariate transformation to standardised scores within each cross-validation fold) using every combination of hyper-parameters and measuring their performances (in this case using AUC). The combination of hyper-parameters with the highest AUC (in this case mean of 100 cross-validation iterations) was selected for the model. The possible hyper-parameters over which the cross-validated grid-searchers were performed were:

• PLR: regularisation = [LASSO (L1), ridge (L2)]; inverse regularisation strength (C) = [0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0].

• SVC: kernel = [linear, radial basis function]; C = [0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0]; kernel coefficient for radial basis function = [0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0].

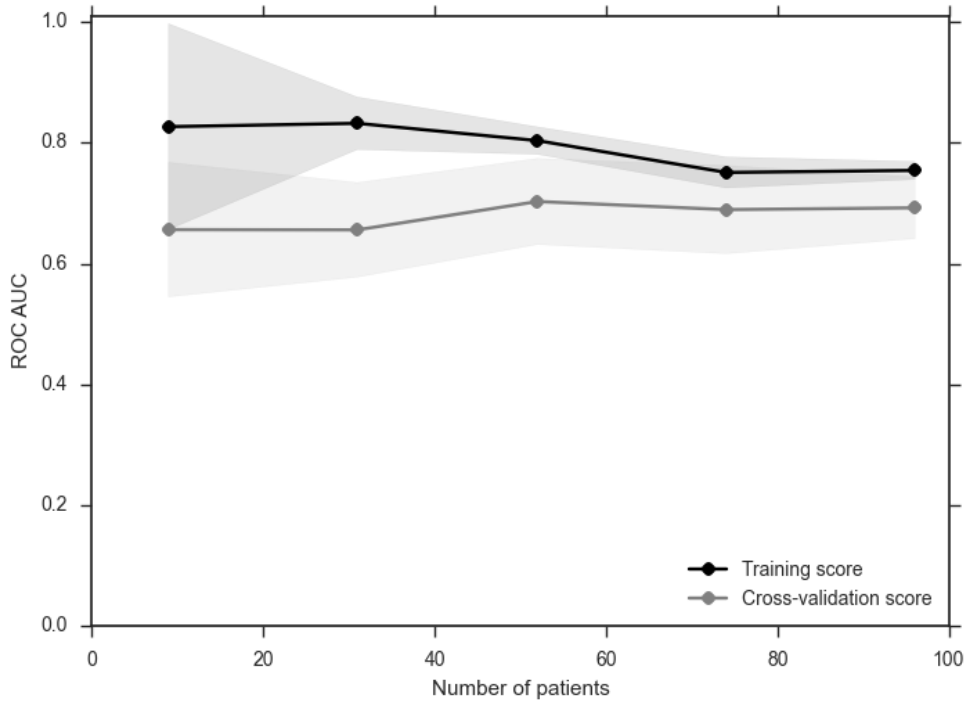
• RFC: number of estimators = 1000; maximum depth = [5, 10, 15, 20]; maximum features = [number of features, number of features/2, square root of number of features].

Internal validation used a 100 iteration shuffled stratified cross-validation, with a train/test split of 80/20. The internal cross-validation used a "nested" design incorporating covariate transformation to standardised scores and hyper-parameter tuning (with 5-fold cross-validation), within each iteration of the internal validation cross-validation. This was done to prevent "data leakage" (unexpected additional information in the training data), which can lead to biased overestimates of how well the model is likely to generalise to other data.

During model fitting the outcome classes were weighted proportionally to the inverse of their class frequencies in the training data to account for the fact that the number of patients experiencing severe and non-severe mucositis was unequal (in part due to the strategy for handling missing data). Higher class weights result in less penalisation for that class and a greater "incentive" for the model to correctly classify it.

Appendix 5: Model diagnostics using learning curves

An important part of statistical modelling is diagnosing why a model does not have perfect predictive ability and, hence, how it can be improved in the future ("model diagnostics"). Model diagnostics were performed using learning curves. The training and cross-validation AUC scores were plotted as a function of the number of patients in the training set. The shapes of the training and cross-validation learning curves were used to infer whether the models displayed underfitting or overfitting and, hence, identify strategies to improve model performance in the future. Appendix figure 2 shows the learning curves for the $PLR_{standard}$ model.



Appendix figure 2: Learning curves to diagnose suboptimal performance the $PLR_{standard}$ model. For small numbers of training cases (patients) the training score is high and decreases as more cases are added and there is less over-fitting. The cross-validation scores, indicating how well the models generalise to unseen data, start low and increase slightly. The training and cross-validation curves converge and level off indicating that adding additional training cases will not improve model performance.

The $PLR_{standard}$ training and cross validation learning curves converge at a AUC substantially less than 1.0, indicating that they underfit the data (with a model that does not fully capture the relationship between treatment and toxicity). Therefore, the $PLR_{standard}$ could be improved by adding additional or more appropriate features, using more sophisticated models or decreasing the amount of regularisation. They imply that just adding additional patients (which would be a suitable strategy if the learning curves did not converge, suggesting the models featured was overfitting to noise) will be inadequate to improve

predictive performance. This suggests that having to exclude a large number of patients from the analysis due to missing data will most likely not have had a large detrimental effect on model performance. Since the more complex SVC and RFC models did not improve model performance and the cross-validated grid search allowed for varying amounts of regularisation, it is likely that improving the features used to describe the treatments or including factors not hitherto considered would be the best approach to improving the model. To the best of our knowledge, our study is the first in radiation oncology outcomes modelling to apply learning curves in order to attempt to establish how to enhance model performance. We recommend that this technique be employed in future predictive modelling studies in radiation oncology.

Appendix 6: Results of statistical analysis including patients with non-consecutive missing mucositis measurements

Appendix table 2 shows the discriminative abilities of the models generated by repeating the modelling of peak acute mucositis with the addition of patients with non-consecutive missing mucositis measurements (which led to 245 patients (Grade 0 – 0 patients, Grade 1 – 16 patients, Grade 2 – 95 patients, Grade 3 – 134 patients) being included).

Appendix table 2: Discriminative abilities of models using 245 patients including patients with non-consecutive missing mucositis measurements.

Model	Hyper parameters	Mean AUC (s.d.)	Mean log loss (s.d.)	Mean Brier score (s.d.)	Mean calibration slope (s.d.)	Mean calibration intercept (s.d.)
$PLR_{standard}$	regularisation = LASSO, C = 0.1	0.68 (0.07)	0.66 (0.03)	0.23 (0.01)	8.6 (8.9)	-4.1 (4.4)
$SVC_{standard}$	kernel = radial basis function, gamma = 0.01, C = 0.1	0.68 (0.07)	-	-	-	-
$RFC_{standard}$	max depth = 20, max features = square root	0.72 (0.07)	0.63 (0.07)	0.22 (0.03)	4.1 (1.9)	-2.0 (1.1)
$PLR_{spatial}$	regularisation = LASSO, C = 0.1	0.68 (0.07)	0.66 (0.03)	0.23 (0.01)	8.1 (8.7)	-3.9 (4.4)
$SVC_{spatial}$	kernel = radial basis function, gamma = 0.0001, C = 10	0.67 (0.08)	-	-	-	-
$RFC_{spatial}$	max depth = 5, max features = square root	0.71 (0.06)	0.62 (0.06)	0.22 (0.03)	4.3 (1.8)	-2.1 (1.0)

PLR – penalised logistic regression; SVC - support vector classification; RFC - random forest classification; s.d. – standard deviation; C – inverse regularisation strength.

Including the additional patients did not lead to improved discrimination (it led to a small decrease in discrimination in the PLR and SVC models) supporting the suggestion from the learning curves analysis (appendix 5) that excluding patients due to missing toxicity data is unlikely to have decreased discriminative ability. It is important to note that there is uncertainty in the toxicity outcome

measures for the patients with missing toxicity data, which is a limitation of the evaluation of the discrimination. For these reasons it was decided that models generated excluding the patients with peak toxicity less than grade 3 having missing toxicity measurements (the models in the main manuscript) should be preferred.

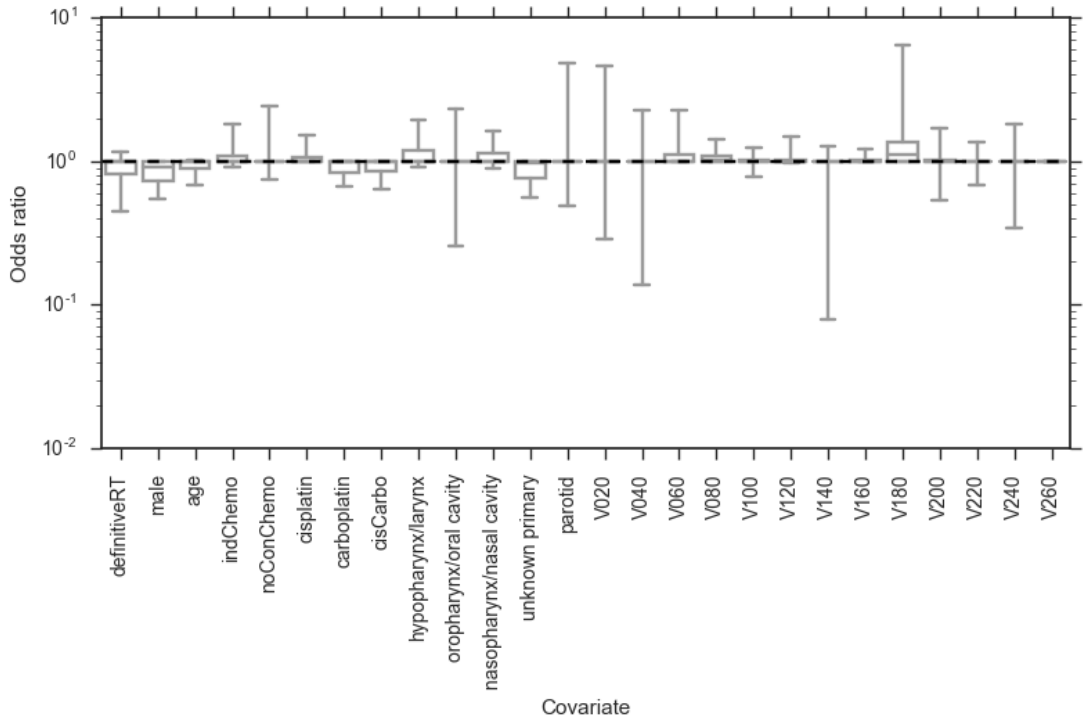
Appendix 7: Statistical modelling of the duration of severe mucositis

Elastic net regression (linear regression with combined LASSO and ridge priors as the regulariser; EN) and random forest regression (RFR) were used to determine associations between the covariates and the duration of severe mucositis to take advantage of the longitudinal nature of the toxicity data. The outcome variable for the regression modelling was the number of weeks that the patients were scored as having grade 3 mucositis. For patients to be included in this analysis they had to have a complete set of toxicity measurements. This was true of 80 patients. The model hyper-parameters were tuned (using R^2 for scoring) and the odds ratios (for EN) and feature importance (for RFR) bootstrapped in the same manner as detailed for the modelling of the peak grade of mucositis. The possible hyper-parameters over which the cross-validated grid searchers were performed were:

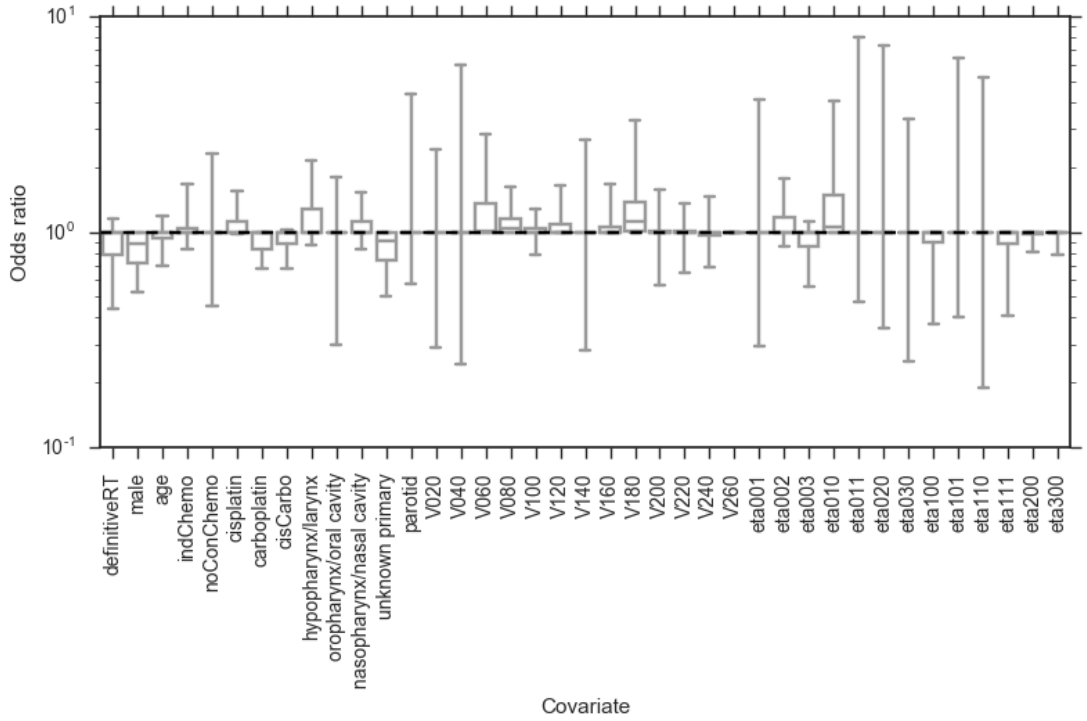
• EN: penalty term multiplier constant (α) = [0.01, 0.1, 1.0, 10.0, 100.0]; elastic net mixing parameter (L1 ratio) = [0.25, 0.5, 0.75].

• RFR: number of estimators = 1000, maximum depth = [5, 10, 15, 20]; maximum features = [number of features, number of features/2, square root of number of features].

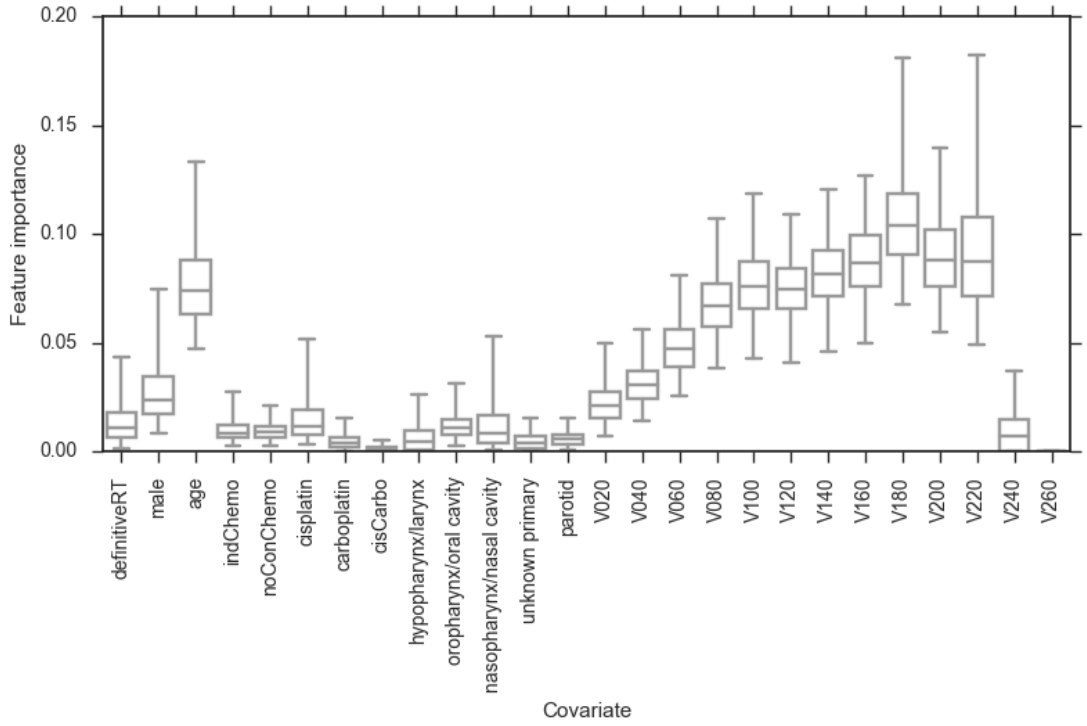
The odds ratios and feature importance measures for the $EN_{standard}$, $EN_{spatial}$, $RFR_{standard}$ and $RFR_{spatial}$ models are shown in appendix figures 3, 4, 5 and 6, respectively.



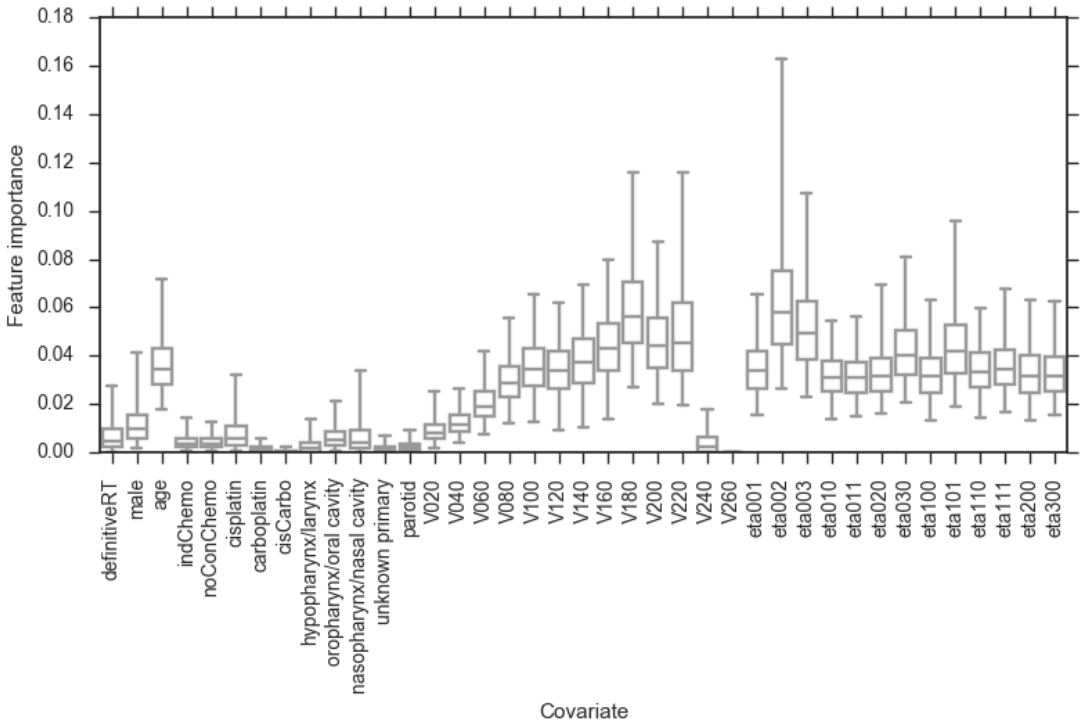
Appendix figure 3: Bootstrapped (2000 replicates) odds ratios for $EN_{standard}$ model. Whiskers show 95 percentiles (non-normal distributions).



Appendix figure 4: Bootstrapped (2000 replicates) odds ratios for $EN_{spatial}$ model. Whiskers show 95 percentiles (non-normal distributions).



Appendix figure 5: Bootstrapped (2000 replicates) feature importance measures for $RFR_{standard}$ model. Whiskers show 95 percentiles (non-normal distributions).



Appendix figure 6: Bootstrapped (2000 replicates) feature importance measures for $RFR_{spatial}$ model. Whiskers show 95 percentiles (non-normal distributions).

None of the covariates were significantly associated with the duration of grade 3 mucositis in the EN models. In the RFR models the covariate most strongly associated with the duration of severe mucositis was *V180*, with the feature importance increasing with increasing dose up to *V180*. This provides further support for aiming to minimise the volume of oral cavity receiving intermediate and high doses rather than mean dose in RT treatment planning. The spatial dose metrics have a relatively high feature importance in the $RFR_{spatial}$ model and age has a higher high feature importance than the other clinical covariates in both RFR models, but this may be due to the tendency of this measure to be biased towards the covariates having a large number of unique values [13], which is the case for the spatial dose metrics and *age*, especially compared to the binary covariates. *eta002*, describing the spread of the dose in the superior-inferior direction, is the spatial dose metric with the highest importance.

Appendix 8: "Conventional" univariate and multivariate logistic regression analysis

Appendix table 3 shows the results of the univariate and multivariate (unpenalised) logistic regression analyses. Covariate data were not transformed to standardised scores for these analyses.

Appendix table 3: Odds ratios and confidence intervals for univariate and multivariate logistic regression analyses.

	Univariate logistic regression		**Multivariate logistic regression**	
Covariate	Odds ratio	95% confidence interval	Odds ratio	95% confidence interval
definitiveRT	0.50	0.36 – 2.30	0.41	2.99×10^{-8} – 7.33
male	1.61	0.49 – 1.98	1.74	0.25 – 21.1
age	0.99	0.97 – 1.03	0.99	0.89 – 1.08
indChemo	1.94	0.51 – 1.99	0.47	4.30×10^{-3} – 7.94
noConChemo	0.47	0.49 – 1.98	0.16	6.22×10^{-7} – 18.0
cisplatin	2.26	0.50 – 2.10	0.76	1.26×10^{-5} – 326
carboplatin	1.49	0.29 – 1.09 x 10^4	0.45	1.10×10^{-5} – 7.40×10^4
cisCarbo	0.72	0.19 – 4.23 x 10^3	0.14	7.70×10^{-7} – 630
hypopharynx /larynx	1.31	0.36 – 5.44	4.94	8.64×10^{-3} – 3.75×10^5
oropharynx	3.47	0.49 – 1.98	0.26	2.62×10^{-6} – 43.9
nasopharynx	1.93	0.36 – 4.00	0.46	2.37×10^{-5} – 2.38×10^4
unknown primary	0.35	0.27 – 1.02 x 10^4	6.74×10^{-3}	5.18×10^{-11} – 21.8

parotid	0.18	0.48 – 2.62	1.88	4.00×10^{-4} – 4.37×10^{3}
V20	1.00	0.96 – 1.03	1.18	0.69 – 4.07
V40	1.00	0.97 – 1.02	0.76	0.15 – 1.39
V60	1.00	0.97 – 1.02	1.18	0.75 – 2.83
V80	1.02	0.98 – 1.02	1.06	0.86 – 1.67
V100	1.02	0.98 – 1.01	0.90	0.55 – 1.08
V120	1.02	0.99 – 1.01	1.22	1.00 – 2.37
V140	1.02	0.99 – 1.01	0.77	0.35 – 1.11
V160	1.03	0.99 – 1.01	1.20	0.67 – 2.27
V180	1.03	0.99 – 1.01	1.07	0.85 – 2.60
V200	1.03	0.99 – 1.01	0.92	0.60 – 1.03
V220	1.09	0.97 – 1.04	1.06	0.95 – 1.47
V240	1.07	0.57 – 3.54×10^{100}	0.84	0.04 – 2.92×10^{14}
V260	1.00	1.00 – 1.00	1.00	1.00 – 1.00
η_{001}	1.28	0.01 – 24.8	1.75×10^{9}	4.20×10^{-7} – 2.93×10^{32}
η_{002}	2.04	0.14 -21.9	0.77	2.00×10^{-5} – 2.10×10^{16}
η_{003}	0.11	1.52×10^{-4} – 200	7.45×10^{-15}	4.79×10^{-58} – 6.10×10^{4}
η_{010}	8.24	2.63×10^{-3} – 777	2.25×10^{8}	0.03 – 2.12×10^{42}
η_{011}	$4.54 \times$	9.21×10^{-7}	$1.77 \times$	1.07×10^{-13} –

	10^4	– 3.49×10^5	10^5	3.18×10^{37}
η_{020}	0.03	9.18×10^{-3} – 676	5.19×10^{-4}	7.51×10^{-25} – 1.18×10^{10}
η_{030}	4.41×10^4	1.99×10^{-5} – 1.13×10^6	3.38×10^{10}	3.00×10^{-5} – 1.57×10^{52}
η_{100}	0.36	0.02 – 363	3.75×10^{-4}	1.96×10^{-29} – 1.35×10^{17}
η_{101}	0.18	3.55×10^{-6} – 5.94×10^3	8.50×10^{-12}	3.06×10^{-49} – 1.42×10^{18}
η_{110}	58.5	4.53×10^{-6} – 1.64×10^7	200	1.73×10^{-28} – 6.62×10^{31}
η_{111}	1.26×10^5	2.92×10^{-18} – 2.51×10^{14}	5.27×10^{-7}	5.52×10^{-41} – 5.45×10^9
η_{200}	0.66	0.01 – 636	1.21	7.56×10^{-9} – 1.08×10^{10}
η_{300}	1.12	0.01 – 2.56×10^3	3.53×10^3	6.98×10^{-18} – 1.16×10^{28}
intercept	-	-	0.01	5.12×10^{-20} – 1.11×10^6

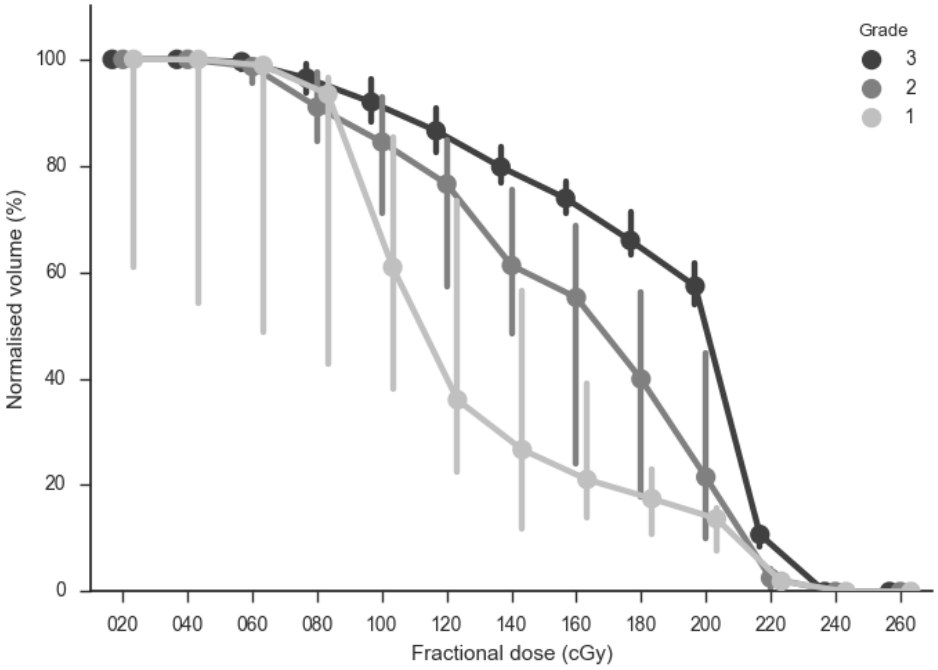
The 95% confidence intervals are the 95 percentiles of the bootstrapped (with 2000 replicates) odds ratios.

On univariate analysis none of the covariates were significantly associated with severe mucositis. On multivariate analysis *V120* was significantly associated with severe mucositis.

The AUC for the multivariate logistic regression model was 0.62 (standard deviation = 0.10) on stratified shuffle split internal cross-validation with 100 iterations.

Appendix 9: Dose-volume data

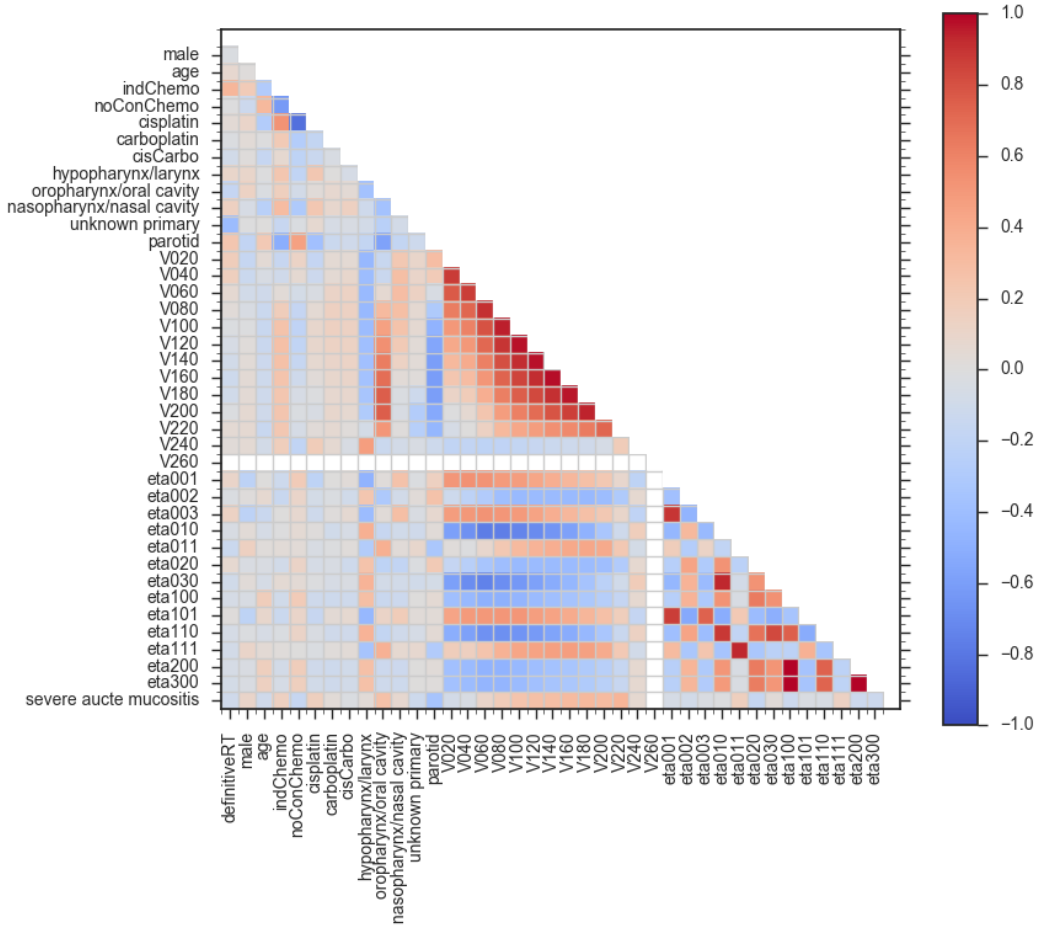
Appendix figure 7 shows the average DVH for each mucositis Grade.



Appendix figure 7: Summary of dose-volume data for oral cavity grouped by peak mucositis Grade. The lines represent the group medians and the error bars represent the 95 percentiles.

Appendix 10: Correlated RT dose data

Appendix figure 8 shows the spearman correlation matrix of the model variables.



Appendix figure 8: Spearman correlation matrix of the model variables. The colour scale shows the spearman correlation coefficients between the model covariates. definitiveRT – definitive radiotherapy (versus post-operative radiotherapy); indChemo – induction chemotherapy; noConChemo – no concurrent chemotherapy; cisCarbo – one cycle of cisplatin followed by one cycle of carboplatin; Vx – volume of organ receiving x cGy of radiation per fraction; etaxyz – 3D moment invariant of order x in the left-right direction, y in the anterior-posterior direction and z in superior-inferior direction; severe acute mucositis – peak grade of acute mucositis (severe = 1, non-severe = 0).

The dose-volume-based metrics are highly correlated. Logistic regression assumes covariates are independent so the regression coefficients (and odds ratios) of correlated covariates are unstable. This can lead to erroneous inference of the relationship between the covariates (dose-volume metrics) and outcome (mucositis severity) [14]. The fact that the odds ratios for some of the

dose covariates are less than 1.0, indicating a negative correlation between dose and toxicity, and that neighbouring dose-volume metrics do not have similar odds ratios (which are counterintuitive), highlights the limitations of attempting inference from logistic regression models incorporating correlated covariates. We, therefore, stress that logistic regression models using correlated RT dose metrics, whilst potentially suitable for predicting patient toxicity outcomes (aim i), should not be used to infer which dose levels are driving that toxicity (aim ii). RFC models are robust to correlated covariates. Therefore, they are more suitable than PLR for inferring associations between correlated RT dose metrics and toxicity. However, it should be noted that, the RFC feature importance measures tend to be biased towards the covariates having the greatest number of unique values [13]. This could explain the relatively high feature importance given to the spatial dose metrics in the $RFC_{spatial}$ model.

Appendix references

[1] Nutting CM, Morden JP, Harrington KJ, Urbano TG, Bhide SA, Clark C, et al. Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial. Lancet Oncol 2011;12:127–36.

[2] Miah AB, Bhide SA, Guerrero-Urbano MT, Clark C, Bidmead AM, St Rose S, et al. Dose-escalated intensity-modulated radiotherapy is feasible and may improve locoregional control and laryngeal preservation in laryngo-hypopharyngeal cancers. Int J Radiat Oncol Biol Phys 2012;82:539–47.

[3] Miah AB, Schick U, Bhide SA, Guerrero-Urbano M-T, Clark CH, Bidmead AM, et al. A phase II trial of induction chemotherapy and chemo-IMRT for head and neck squamous cell cancers at risk of bilateral nodal spread: the application of a bilateral superficial lobe parotid-sparing IMRT technique and treatment outcomes. Br J Cancer 2015;112:32–8.

[4] Bhide S, Clark C, Harrington K, Nutting CM. Intensity modulated radiotherapy improves target coverage and parotid gland sparing when delivering total mucosal irradiation in patients with squamous cell carcinoma of head and neck of unknown primary site. Med Dosim 2007;32:188–95.

[5] Otter S, Schick U, Gulliford S, Lal P, Franceschini D, Newbold K, et al. Evaluation of the Risk of Grade 3 Oral and Pharyngeal Dysphagia Using Atlas-Based Method and Multivariate Analyses of Individual Patient Dose Distributions. Int J Radiat Oncol 2015;93:507–15.

[6] Python Software Foundation. http://www.python.org (accessed July 1, 2015).

[7] van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. Comput Sci Eng 2011;13:22–30.

[8] Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng 2007;9:90–5.

[9] Waskom M. https://github.com/mwaskom/seaborn/tree/v0.6.0 (accessed July 1, 2015).

[10] Mason D. https://code.google.com/p/pydicom/ (accessed July 1, 2015).

[11] McKinney W. Data Structures for Statistical Computing in Python. Proc 9th Python Sci Conf 2010;1697900:51–6.

[12] Pedregosa F, Weiss R, Brucher M. Scikit-learn: Machine Learning in Python 2011;12:2825–30.

[13] Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007;8:25.

[14] Farrar DE, Glauber RR. Multicollinearity in regression analysis: the problem revisited. Rev Econ Stat 1967;49:92–107.