Published in final edited form as:
Radiother Oncol. 2016 July ; 120(1): 21–27. doi:10.1016/j.radonc.2016.05.015.

Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy

Jamie A Dean[a], **Kee H Wong**[b], **Liam C Welsh**[b], **Ann-Britt Jones**[b], **Ulrike Schick**[b], **Kate L Newbold**[b,c], **Shreerang A Bhide**[b,c], **Kevin J Harrington**[b,c], **Christopher M Nutting**[b,c], and **Sarah L Gulliford**[a]

[a]Joint Department of Physics at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, London, UK

[b]Head and Neck Unit, The Royal Marsden NHS Foundation Trust, London, UK

[c]Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, UK

Abstract

Background and Purpose—Severe acute mucositis commonly results from head and neck (chemo)radiotherapy. A predictive model of mucositis could guide clinical decision-making and inform treatment planning. We aimed to generate such a model using spatial dose metrics and machine learning.

Material and Methods—Predictive models of severe acute mucositis were generated using radiotherapy dose (dose-volume and spatial dose metrics) and clinical data. Penalised logistic regression, support vector classification and random forest classification (RFC) models were generated and compared. Internal validation was performed (with 100-iteration cross-validation), using multiple metrics, including area under the receiver operating characteristic curve (AUC) and calibration slope, to assess performance. Associations between covariates and severe mucositis were explored using the models.

Results—The dose-volume-based models (standard) performed equally to those incorporating spatial information. Discrimination was similar between models, but the $RFC_{standard}$ had the best calibration. The mean AUC and calibration slope for this model were 0.71 (s.d.=0.09) and 3.9 (s.d.=2.2), respectively. The volumes of oral cavity receiving intermediate and high doses were associated with severe mucositis.

Conclusions—The $RFC_{standard}$ model performance is modest-to-good, but should be improved, and requires external validation. Reducing the volumes of oral cavity receiving intermediate and high doses may reduce mucositis incidence.

Corresponding author: Jamie A Dean, Mailing address: Joint Department of Physics at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, Downs Road, Sutton, SM2 5PT, UK, Telephone: +442089156223, Fax: +442086433812, jamie.dean@icr.ac.uk.

Europe PMC Funders Author Manuscripts

Keywords

oral mucositis; NTCP modelling; dose-response modelling; machine learning; spatial dose metrics; head and neck radiotherapy

Introduction

Mucositis is a common acute toxicity of head and neck radiotherapy (RT), which may result in pain, dysphagia [1], weight loss and aspiration, and reduced quality of life [2]. Mucositis may lead to missed treatment fractions [3], potentially compromising locoregional control [4], and is frequently dose-limiting in dose-escalation and accelerated fractionation regimens designed to improve tumour control [5]. Moreover, advances in our understanding of the mechanisms of "late" radiation effects have implicated severe acute reactions in the development of these toxicities [6,7].

There has been a large effort to develop and validate accurate multifactorial normal tissue complication probability (NTCP) models (e.g. [8]) for clinical decision-support [9], treatment modality selection [10] and treatment plan optimisation [11]. However, the prediction of the severity of acute mucositis for individual patients is highly challenging and there are currently no NTCP models that can confidently guide clinical decision-making. Dose objectives, such as those proposed by the Radiation Therapy Oncology Group (RTOG) clinical trials, specify varying limits for the mean dose delivered to the oral cavity in the range of 30 – 50 Gy (RTOG 0912, RTOG 0920, RTOG 1216).

It has been hypothesised that one of the major contributing factors to the suboptimal performance of many NTCP models is an oversimplified description of the dose distribution [12] using dose-volume histograms (DVH). Two assumptions are implicit in this technique. Firstly, each voxel in the organ contributes equally to a toxicity outcome. Secondly, the spatial distribution of that dose has no bearing on toxicity. Our group has previously shown that the spatial distribution of the dose has an impact on toxicity prediction for both rectal toxicity [13] and xerostomia [14]. We, therefore, considered that the spatial distribution of the dose might also play an important role in mucositis. The buccal mucosa is keratinised, whereas other regions of the oral mucosa, such as parts of the soft palate and ventral tongue, are not [15], and hence might be expected to be associated with higher mucositis scores [16].

The two distinct aims of this study were to (i) generate, and validate models for the prediction of the severity of acute oral mucositis for individual patients to guide clinical decision-making; and (ii) use those models to establish RT dose-response associations for severe mucositis that could inform optimal dose-sparing of the oral mucosa in head and neck RT treatment planning protocols.

Materials and Methods

Patient data

Data from 351 head and neck RT patients, enrolled in six different clinical trials [17–19] (with institutional review board approval and signed patient consent; details of the trials in

Europe PMC Funders Author Manuscripts

appendix 1), were available. This builds on a previous study, by our group, based on data from four phase II clinical trials [20] by incorporating more data from two phase III trials, in addition to methodological developments. The patients included in the study encompass a range of head and neck primary disease sites and RT delivery techniques and fractionation schedules, thus ensuring a large variation in the dose distributions across the cohort. Only patients for whom DICOM RT data were available (351 patients) were included.

Toxicity was consistently scored for all studies using the clinician-observed oral mucositis score from the Common Terminology Criteria for Adverse Events (CTCAE) versions 2 (mucositis due to radiation score) [21] or 3 (mucositis/stomatitis (clinical exam) score) [22] instruments, which are near equivalent. Toxicities were recorded prospectively prior to the start of RT, weekly during RT, weekly from 1 - 4 weeks post-RT and at 8 weeks post-RT. No formal quality assurance of these scores (e.g. intra- and/or inter-observer variability) was undertaken, but all data were generated by experienced head and neck cancer specialists working according to standard trial protocols and trained in the use of the scoring systems. The toxicity endpoint of interest chosen for analysis was the maximum reported grade and was dichotomised into severe (maximum toxicity score of grade 3 or worse) and non-severe (maximum toxicity score of less than grade 3) mucositis. Patients with baseline toxicity were excluded from the analysis. Patients with missing data and peak grade below 3 were excluded from the analysis as these patients may have in fact experienced grade 3 mucositis. Following these inclusion criteria 183 patients were available. Maximum toxicity scores of grade 1, 2 and 3 were experienced by 8 (4%), 41 (22%) and 134 (73%) patients, respectively.

Relevant clinical data were included as covariates in the models where available. These were induction chemotherapy (n = 89), concurrent chemotherapy regimen (cisplatin (n = 64), carboplatin (n = 10), one cycle of cisplatin then one cycle of carboplatin (n = 6) or none (n = 103)), definitive (n = 152) versus post-operative RT, primary disease site (nasopharynx (n = 18), oropharynx (n = 100), hypopharynx/larynx (n = 18), parotid gland (n = 39), unknown primary (n = 8)), age (median = 58 years; range = 17 – 88 years) and sex (n_{male} = 116).

Dosimetric data

The oral mucosa was contoured, by clinical oncologists, using our previously described method [23], which represents the mucosa by an approximately spherical volume encompassing the oral cavity (including “the surfaces of the inner table of mandible, tongue, base of tongue, floor of mouth and palate”; see appendix 2 for example). The physical dose distribution was converted to the fractional dose distribution (physical dose delivered in each fraction) [24], to account for differences in the fractionation schedules. The relative cumulative dose-volume histogram in 20 cGy intervals from 20 to 260 cGy per fraction was inserted as covariates in the models. 3D moment invariants [10] were calculated, and used as model covariates, to describe the centre of mass, spread and skewness of the dose distribution in the three orthogonal directions (left-right, anterior-posterior, superior-inferior) within the oral cavity (appendix 3). Differences in treatment technique (unilateral versus bilateral and conformal versus intensity-modulated RT (IMRT)) were captured by the dose distributions.

Europe PMC Funders Author Manuscripts

Statistical analysis

The statistical analysis used machine learning methods and followed the principles suggested by Kang *et al.* [25] for model generation and the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) international consensus guidelines for model evaluation [26]. All RT dose and clinical covariates were transformed to standardised (Z) scores (mean = 0, standard deviation = 1) to avoid scale-related feature dominance. Three different types of classification model: penalised logistic regression (PLR) [27], support vector classification (SVC) [28] and random forest classification (RFC) [29] were developed. Models were generated with (spatial) and without (standard) the addition of the spatial dose metrics.

Removing covariates based on univariable or stepwise methods has been extensively shown to result in overfitting (resulting in models that are not generalisable) and biased parameter estimates (resulting in misleading associations between covariates and outcomes) [30–32]. PLR has previously been found to outperform logistic regression with stepwise variable selection for NTCP modelling studies [33]. Penalisation was performed using ridge regularisation [34] or least absolute shrinkage and selection operator (LASSO) regularisation [27]. These techniques reduce the regression coefficients, setting some of them to zero (removing that covariate) in the case of LASSO. SVC models attempt to find a hyper-plane to separate two outcome classes (in this case severe and non-severe mucositis) and are able to solve non-linearly separable problems (by using non-linear kernels, such as a Gaussian radial basis function). SVC models with non-linear kernels do not have intuitive metrics to describe the strength of associations between covariates and outcome (making them less interpretable than PLR). RFC models construct an ensemble of decision trees. They are non-linear, non-parametric and more robust to correlated covariates than PLR. RFC models provide feature importance measures, which offer information on the relative strength of association between the model covariates and outcome.

The model hyper-parameters were tuned (appendix 4) and the generalisability of the models to predict mucositis severity for individual patients (aim i) was measured through internal validation (methods detailed in appendix 4). The TRIPOD guidelines state that randomly splitting data into development and validation sets is erroneously believed to be external validation, but has been shown to be a “weak and inefficient form of internal validation” [26,35,36]. Therefore, all data were used for model generation and internal validation. Discriminative ability was measured using area under the receiver operating characteristic curve (AUC). To make individual patient predictions of the probability of an outcome good model calibration is important in addition to discrimination [26]. Model calibration was assessed, using the slope and intercept of a logistic regression model (without penalisation) of the actual toxicity outcomes against the predicted probabilities of severe mucositis (perfect calibration would have a slope of 1 and intercept of 0) [37,38]. The Brier score [39] was calculated to evaluate the overall model performance (lower values indicate better performance) and log loss [40] calculated to assess the model probability estimates (lower values indicate better probability estimates). A comprehensive consideration of model discrimination and calibration metrics was used to compare models. This is more appropriate than formal statistical comparison of AUC, which gives equal importance

Europe PMC Funders Author Manuscripts

weighting to sensitivity and specificity, in the context of NTCP modelling. Model diagnostics were performed using learning curves [41] (appendix 5).

The decision to remove patients with any missing data and maximum mucositis score less than grade 3 may be considered overly conservative. For completeness, the modelling of peak mucositis was repeated, but with the inclusion of patients who had non-consecutive missing mucositis measurements (increasing statistical power at the expense of increased potential bias; appendix 6).

To establish dose-response associations (aim ii), the strength of the associations between the covariates and severe mucositis were assessed by bootstrapping (to obtain unbiased confidence intervals) the PLR odds ratios and RFC feature importance measures for the models with 2000 replicates (model hyper-parameters were retuned within each bootstrap replicate). For completeness, the duration of severe mucositis was modelled, using elastic net regression and random forest regression and the associations between the model covariates and outcome assessed (appendix 7). The “conventional” approach to NTCP modelling considering both dosimetric and non-dosimetric covariates is to use univariable and multivariable (unpenalised) logistic regression. Therefore, for completeness, this was also performed (appendix 8).

Results

The average DVH for each mucositis grade is shown in appendix 9 and demonstrates a clear relationship between dose and toxicity. The results of the evaluation of the models, using multiple metrics addressing different aspects of predictive performance, (addressing aim i) are shown in table 1 (and appendix 6). SVC models do not provide probability estimates and so only discrimination could be assessed. Attempts were made to convert the SVC model outputs to probability estimates using Platt scaling [42]. However, this led to substantial reductions in AUC (related to the algorithm used; data not shown). The $PLR_{standard}$, $SVC_{standard}$ and $RFC_{standard}$ models had approximately equal discriminative abilities. The addition of 3D moment invariants, describing the spatial distribution of the dose, did not improve the discriminative ability, or other measures of predictive performance, of the models. Therefore, the simpler standard models were favoured. The $RFC_{standard}$ model had better calibration, probability estimates and overall performance than the $PLR_{standard}$ model so was favoured over the other models, for prediction of the severity of mucositis for individual patients (aim i). The $RFC_{standard}$ model is provided at https://github.com/jamiedean/oral-mucositis-model. For completeness, the $PLR_{standard}$ model (accounting for covariate transformations to standardised scores) is given by:

$$NTCP = \frac{e^f}{1+e^f}$$

where

Europe PMC Funders Author Manuscripts

$$f = -0.025\left(\frac{unknownPrimary - 0.044}{0.205}\right) - 0.303\left(\frac{parotid - 0.209}{0.406}\right) + 0.212\left(\frac{V180 - 53.6}{27.5}\right) + 0.194\left(\frac{V220 - 10.5}{11.1}\right)$$

where *unknownPrimary* and *parotid* are binary and *V180* and *V220* are given as percentages.

The odds ratios and feature importance measures of the bootstrapped PLR and RFC model covariates (addressing aim ii), and confidence intervals (95 percentiles of the bootstrapped values; non-normal distributions), are displayed in figures 1 - 4. In the PLR models none of the covariates was significantly associated with severe mucositis (95 percentiles of the odds ratio not crossing 1). The correlation matrix for the data (appendix 10) indicates the highly correlated nature of the dosimetric data. It should be noted that logistic regression assumes covariates are independent so the regression coefficients of correlated covariates are unstable so we do not recommend using logistic regression to infer dose-response associations between correlated dose metrics and toxicity (discussed in appendix 10). RFC models are more robust to correlated covariates and, hence, more appropriate for inferring associations between the correlated dose metrics and severe mucositis. The covariate with the highest RFC feature importance was *V220* in both RFC models. There was a general trend of increasing feature importance with increasing dose and feature importance was also high for RT dose metrics in the range *V80* – *V220*. The high fractional dose-volume parameters, *V240* and *V260*, were either 0% or close to 0% for nearly all patients (appendix 9). Therefore, they did not correlate well with mucositis severity in our dataset. A similar pattern was observed in regression modelling of the duration of severe mucositis (appendix 7). Age was the clinical covariate with the highest feature importance. However, this may be an artefact of the large number of possible values compared with the other clinical covariates [43]. Age was not significantly associated with severe mucositis on univariable logistic regression (appendix 8).

Discussion

We met our first aim of generating and validating predictive models of severe acute mucositis. The discriminative ability of the $RFC_{standard}$ model (and the other models) is modest to good. The $RFC_{standard}$ model was better calibrated to the internal validation data than the $PLR_{standard}$ model, as demonstrated by having a calibration slope closer to 1, calibration intercept closer to 0 and lower log loss, and better overall performance, as indicated by its lower Brier score. We also met our second aim of determining associations between RT dose metrics and severe mucositis that could be used to inform improved RT planning. Regarding aim ii, we determined that the covariate with the strongest association with mucositis outcome (peak grade or duration of grade 3) was the *V220*. In interpreting

the associations, it is important to note that they are data-driven. The fact that, for our dataset, the *V220* had the strongest association with severe mucositis does not mean that this dose level has a greater biological effect than higher dose levels. The variance of the higher dose metrics is lower, as the volumes of oral cavity receiving very high doses is close to 0 for all patients, and so the covariance for these metrics with severe mucositis is lower than for *V220*, which has a higher variance and covariance (appendices 9 and 10). We also found associations between other, intermediate and high, dose levels and severe mucositis. This indicates that constraining the mean dose delivered to the oral cavity, as required in RTOG trials, may not be the optimal treatment planning technique to reduce the incidence of severe mucositis. The mean dose gives equal weighting to all dose levels. However, our findings, suggest that minimising the volume of the oral cavity receiving intermediate and high doses as much as possible would represent a better strategy. We recommend incorporating this approach into RT planning, where possible without compromising other aspects of the plan, such as PTV coverage.

Despite a large number of NTCP models for other toxicities, such as xerostomia [14,44] and dysphagia [45], and the high incidence of severe acute mucositis there is a scarcity of models to allow its prediction for individual patients and inform RT planning protocols. This study improved upon our previous findings [20] due to far greater variation in the RT dose distributions in the patient cohort included (as a result of including patients from the PARSPORT and COSTAR trials) and a more rigorous statistical exploration. The wide range of dose distributions increases the generalisability of the models and reduces the chance of introducing biases, for example, due to the primary tumour location. Only one other model of severe acute mucositis resulting from IMRT has been published [46]. This study made a similar finding to our dose-response association (aim ii) that the volume of oral mucosa (defined as oral cavity, oropharynx and hypopharynx) receiving 10.1 Gy per week (2.0 Gy per daily fraction) was most strongly associated with severe mucositis. The authors also found a positive correlation between concurrent chemotherapy and severe mucositis, unlike our study, which found no significant association. A possible explanation is that, in our dataset, concurrent chemotherapy was positively correlated with the dose-volume metrics (appendix 10) so the RT dose effects may mask the effects of concurrent chemotherapy. This is largely due to the fact that all of the parotid gland tumour patients received unilateral irradiation (less dose delivered to the oral cavity) and did not receive concurrent chemotherapy (appendix 1). It is also likely that the effect of chemotherapy is insufficiently characterised (using binary covariates) in our analysis. The numbers of patients receiving carboplatin or one cycle of cisplatin followed by one cycle of carboplatin are likely too small to be able to detect any significant associations.

Our study features several limitations. The current delineation technique used to contour the oral mucosa does not provide an anatomically accurate representation of the mucosal surfaces within the oral cavity, but instead an oral cavity volume. A large amount of this volume is the musculature of the tongue and not mucosa. Additionally, the volume does not encompass all of the oral mucosal surfaces, such as the buccal mucosa. This may also have contributed to the lack of increased predicted performance with the addition of spatial dose metrics. The lack of standardised guidelines for accurately delineating the oral mucosa may have contributed to the scarcity of oral mucositis NTCP models. We have recently validated

a method of automatically contouring the oral mucosal surfaces in a more anatomically realistic manner [47,48]. We intend to use this approach in future analyses and determine whether characterising the mucosal dose distributions in this manner improves the predictive power of our NTCP model. It should be noted that the CTCAE clinical mucositis scoring system does not capture the morphological extent of the mucositis. Therefore, the spatial metrics have the potential to be sensitive to regional variations in the radiosensitivity of the oral mucosa, but not the morphological extent of the mucositis. Additionally, there are factors that are likely to contribute to mucositis, but could not be analysed, as insufficient or no data were available. Tobacco and alcohol use were not collected in the PARSPORT or COSTAR trials, so were not included in the analysis. Genetic predispositions to severe (chemo)radiation-induced toxicity are also expected. Finally, our models have not been externally validated. We suggest that their discrimination and calibration are evaluated in different cohorts of patients to better assess their generalisability.

In conclusion, we have (i) generated and validated NTCP models for the prediction of the severity of acute mucositis for individual patients with modest to good discrimination and (ii) established RT dose-response associations for severe mucositis. We found that a RFC model incorporating clinical and DVH data provided equal discriminative ability to, and better calibration than, PLR and SVC models and represents a promising foundation for a clinical decision-support tool for individual patient management. We demonstrated an association between volumes of oral cavity receiving intermediate and high doses and severe mucositis and, hence, recommend that these should be minimised where possible in RT planning.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council, Cancer Research UK Programme Grant A13407 and NHS funding to the NIHR Biomedical Research Centre at The Royal Marsden and ICR. The PARSPORT and COSTAR trials were supported by Cancer Research UK (trial reference numbers CRUK/03/005 and CRUK/08/004). We wish to thank Hannah Eyles, Emma Wells, Shankar Bodla and James Morden and Dr Emma Hall at The Institute of Cancer Research Clinical Trials and Statistics Unit for data collation, Dr Alex Dunlop, Dr Dualta McQuaid, Dr Simeon Nill and Prof Uwe Oelfke for general support and Dr Jung Hun Oh and Prof Joseph Deasy for insightful discussions.

References

[1]. Sanguineti G, Rao N, Gunn B, Ricchetti F, Fiorino C. Predictors of PEG dependence after IMRT ±chemotherapy for oropharyngeal cancer. Radiother Oncol. 2013; 107:300–4. [PubMed: 23773408]

[2]. Trotti A. Toxicity in head and neck cancer: a review of trends and issues. Int J Radiat Oncol Biol Phys. 2000; 47:1–12. [PubMed: 10758302]

[3]. Trotti A, Bellm LA, Epstein JB, Frame D, Fuchs HJ, Gwede CK, et al. Mucositis incidence, severity and associated outcomes in patients with head and neck cancer receiving radiotherapy with or without chemotherapy: a systematic literature review. Radiother Oncol. 2003; 66:253–62. [PubMed: 12742264]

[4]. Robertson AG, Robertson C, Perone C, Clarke K, Dewar J, Elia MH, et al. Effect of gap length and position on results of treatment of cancer of the larynx in Scotland by radiotherapy: a linear quadratic analysis. Radiother Oncol. 1998; 48:165–73. [PubMed: 9783888]

[5]. Bourhis J, Overgaard J, Audry H, Ang KK, Saunders M, Bernier J, et al. Hyperfractionated or accelerated radiotherapy in head and neck cancer: a meta-analysis. Lancet. 2006; 368:843–54. [PubMed: 16950362]

[6]. Bentzen SM. Preventing or reducing late side effects of radiation therapy: radiobiology meets molecular pathology. Nat Rev Cancer. 2006; 6:702–13. [PubMed: 16929324]

[7]. Denham JW, Peters LJ, Johansen J, Poulsen M, Lamb DS, Hindley A, et al. Do acute mucosal reactions lead to consequential late reactions in patients with head and neck cancer? Radiother Oncol. 1999; 52:157–64. [PubMed: 10577701]

[8]. Dehing-Oberije C, De Ruysscher D, Petit S, Van Meerbeeck J, Vandecasteele K, De Neve W, et al. Development, external validation and clinical usefulness of a practical prediction model for radiation-induced dysphagia in lung cancer patients. Radiother Oncol. 2010; 97:455–61. [PubMed: 21084125]

[9]. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, et al. “Rapid Learning health care in oncology” - An approach towards decision support systems enabling customised radiotherapy. Radiother Oncol. 2013; 109:159–64. [PubMed: 23993399]

[10]. Langendijk JA, Lambin P, De Ruysscher D, Widder J, Bos M, Verheij M. Selection of patients for radiotherapy with protons aiming at reduction of side effects: The model-based approach. Radiother Oncol. 2013; 107:267–73. [PubMed: 23759662]

[11]. Kierkels RGJ, Korevaar EW, Steenbakkers RJHM, Janssen T, van’t Veld AA, Langendijk JA, et al. Direct use of multivariable normal tissue complication probability models in treatment plan optimisation for individualised head and neck cancer radiotherapy produces clinically acceptable treatment plans. Radiother Oncol. 2014; 112:430–6. [PubMed: 25220369]

[12]. Marks LB, Yorke ED, Jackson A, Ten Haken RK, Constine LS, Eisbruch A, et al. Use of normal tissue complication probability models in the clinic. Int J Radiat Oncol Biol Phys. 2010; 76:10–9.

[13]. Buettner F, Gulliford SL, Webb S, Sydes MR, Dearnaley DP, Partridge M. The dose-response of the anal sphincter region - An analysis of data from the MRC RT01 trial. Radiother Oncol. 2012; 103:347–52. [PubMed: 22520267]

[14]. Buettner F, Miah AB, Gulliford SL, Hall E, Harrington KJ, Webb S, et al. Novel approaches to improve the therapeutic index of head and neck radiotherapy: an analysis of data from the PARSPORT randomised phase III trial. Radiother Oncol. 2012; 103:82–7. [PubMed: 22444242]

[15]. Sonis ST. The pathobiology of mucositis. Nat Rev Cancer. 2004; 4:277–84. [PubMed: 15057287]

[16]. Narayan S, Lehmann J, Coleman MA, Vaughan A, Yang CC, Enepekides D, et al. Prospective evaluation to establish a dose response for clinical oral mucositis in patients undergoing head-and-neck conformal radiotherapy. Int J Radiat Oncol Biol Phys. 2008; 72:756–62. [PubMed: 18417299]

[17]. 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. [PubMed: 21236730]

[18]. 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. [PubMed: 21236602]

[19]. 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. [PubMed: 25474250]

[20]. 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 Biol Phys. 2015; 93:507–15. [PubMed: 26460992]

[21]. The National Cancer Institute. Common Toxicity Criteria (CTC) Version 2.0. 1999

[22]. The National Cancer Institute. Common Terminology Criteria for Adverse Events v3.0 (CTCAE). 2006

[23]. Bhide SA, Gulliford S, Fowler J, Rosenfelder N, Newbold K, Harrington KJ, et al. Characteristics of response of oral and pharyngeal mucosa in patients receiving chemo-IMRT for head and neck cancer using hypofractionated accelerated radiotherapy. Radiother Oncol. 2010; 97:86–91. [PubMed: 20826031]

[24]. Tucker SL, Michalski JM, Bosch WR, Mohan R, Dong L, Winter K, et al. Use of fractional dose-volume histograms to model risk of acute rectal toxicity among patients treated on RTOG 94-06. Radiother Oncol. 2012; 104:109–13. [PubMed: 22673726]

[25]. Kang J, Schwartz R, Flickinger J, Beriwal S. Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. Int J Radiat Oncol Biol Phys. 2015; 93:1127–35. [PubMed: 26581149]

[26]. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med. 2015; 162:W1–73. [PubMed: 25560730]

[27]. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B. 1996; 58:267–88.

[28]. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995; 297:273–97.

[29]. Breiman L. Random Forests. Mach Learn. 2001; 45:5–32.

[30]. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. Br J Math Stat Psychol. 1992; 45:265–82.

[31]. Malek M, Berger D, Coburn J. On the inappropriateness of stepwise regression analysis for model building and testing. Eur J Appl Physiol. 2007; 101:263–4. [PubMed: 17520270]

[32]. Fernandes-Taylor S, Hyun JK, Reeder RN, Harris AH. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. BMC Res Notes. 2011; 4:304. [PubMed: 21854631]

[33]. Xu C-J, van der Schaaf A, Schilstra C, Langendijk JA, van't Veld AA. Impact of statistical learning methods on the predictive power of multivariate normal tissue complication probability models. Int J Radiat Oncol Biol Phys. 2012; 82:e677–84. [PubMed: 22245199]

[34]. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970; 12:55–67.

[35]. Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001; 54:774–81. [PubMed: 11470385]

[36]. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM. Internal and external validation of predictive models: A simulation study of bias and precision in small samples. J Clin Epidemiol. 2003; 56:441–7. [PubMed: 12812818]

[37]. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010; 21:128–38. [PubMed: 20010215]

[38]. Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. BMJ. 2015; 351:h3868. [PubMed: 26264962]

[39]. Brier GW. Verification of forecasts expersses in terms of probaility. Mon Weather Rev. 1950; 78:1–3.

[40]. Good IJ. Rational decisions. J R Stat Soc Ser B. 1952; 14:107–14.

[41]. Cortes C, Jackel LD, Solla SA, Vapnik V, Denker JS. Learning curves: asymptotic values and rate of convergence. Adv Neural Inf Process Syst. 1994; 6:327–34.

[42]. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classif. 1999; 10:61–74.

[43]. 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. [PubMed: 17254353]

[44]. Beetz I, Schilstra C, van der Schaaf A, van den Heuvel ER, Doornaert P, Van Luijk P, et al. NTCP models for patient-rated xerostomia and sticky saliva after treatment with intensity modulated radiotherapy for head and neck cancer: The role of dosimetric and clinical factors. Radiother Oncol. 2012; 105:101–6. [PubMed: 22516776]

[45]. Wopken K, Bijl HP, van der Schaaf A, van der Laan HP, Chouvalova O, Steenbakkers RJHM, et al. Development of a multivariable normal tissue complication probability (NTCP) model for tube feeding dependence after curative radiotherapy/chemo-radiotherapy in head and neck cancer. Radiother Oncol. 2014; 113:95–101. [PubMed: 25443500]

[46]. Sanguineti G, Sormani MP, Marur S, Gunn GB, Rao N, Cianchetti M, et al. Effect of radiotherapy and chemotherapy on the risk of mucositis during intensity-modulated radiation therapy for oropharyngeal cancer. Int J Radiat Oncol Biol Phys. 2012; 83:235–42. [PubMed: 22104358]

[47]. Dean JA, Welsh LC, Gulliford SL, Harrington KJ, Nutting CM. A novel method for delineation of oral mucosa for radiotherapy dose–response studies. Radiother Oncol. 2015:10–3. [PubMed: 26026485]

[48]. Dean JA, Welsh LC, McQuaid D, Wong KH, Aleksic A, Dunne E, et al. Assessment of fully-automated atlas-based segmentation of novel oral mucosal surface organ-at-risk. Radiother Oncol. 2016

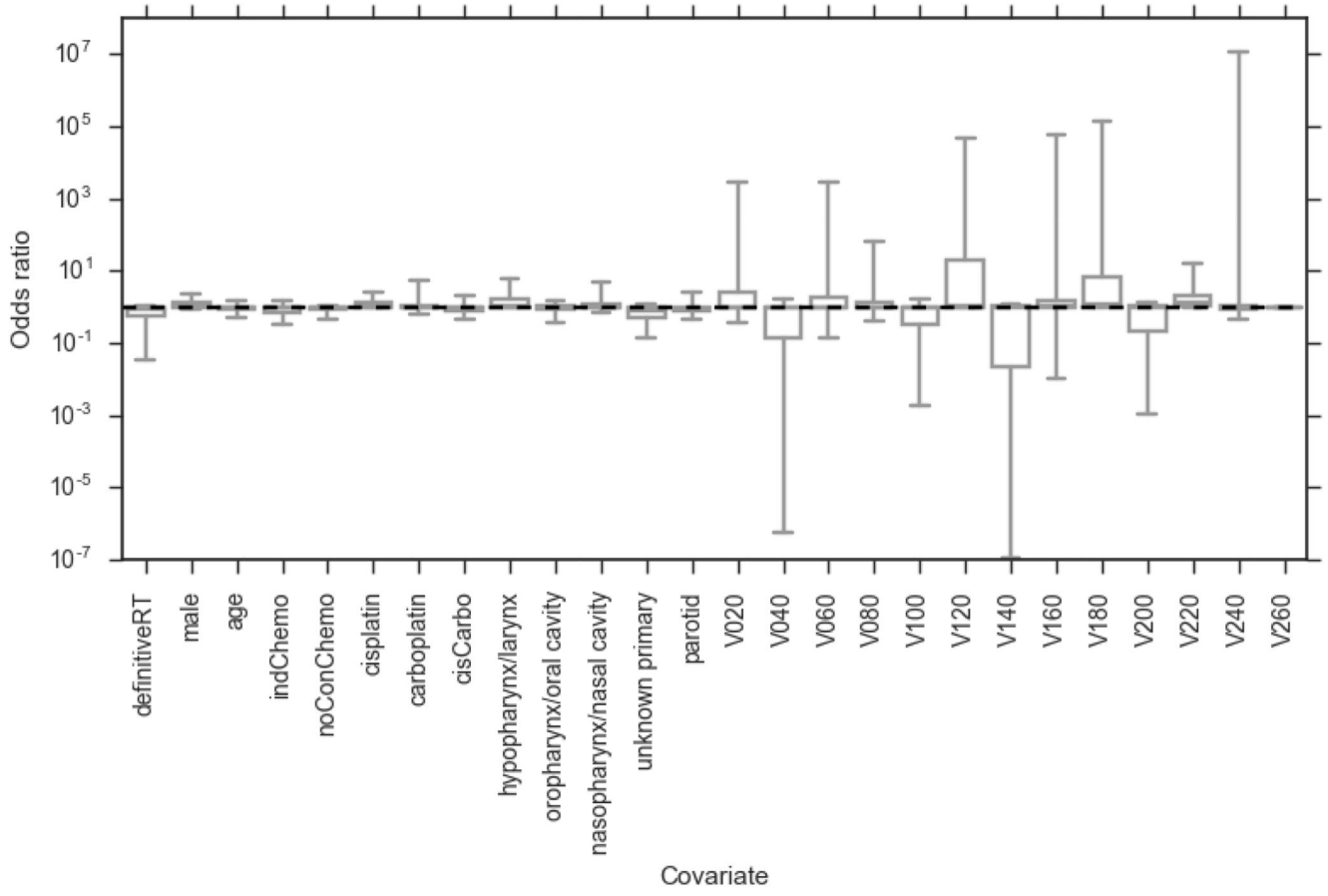

Figure 1.

Bootstrapped (2000 replicates) odds ratios for $PLR_{standard}$ model. Whiskers show 95 percentiles (non-normal distributions). 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. None of the covariates are significantly associated with severe mucositis.

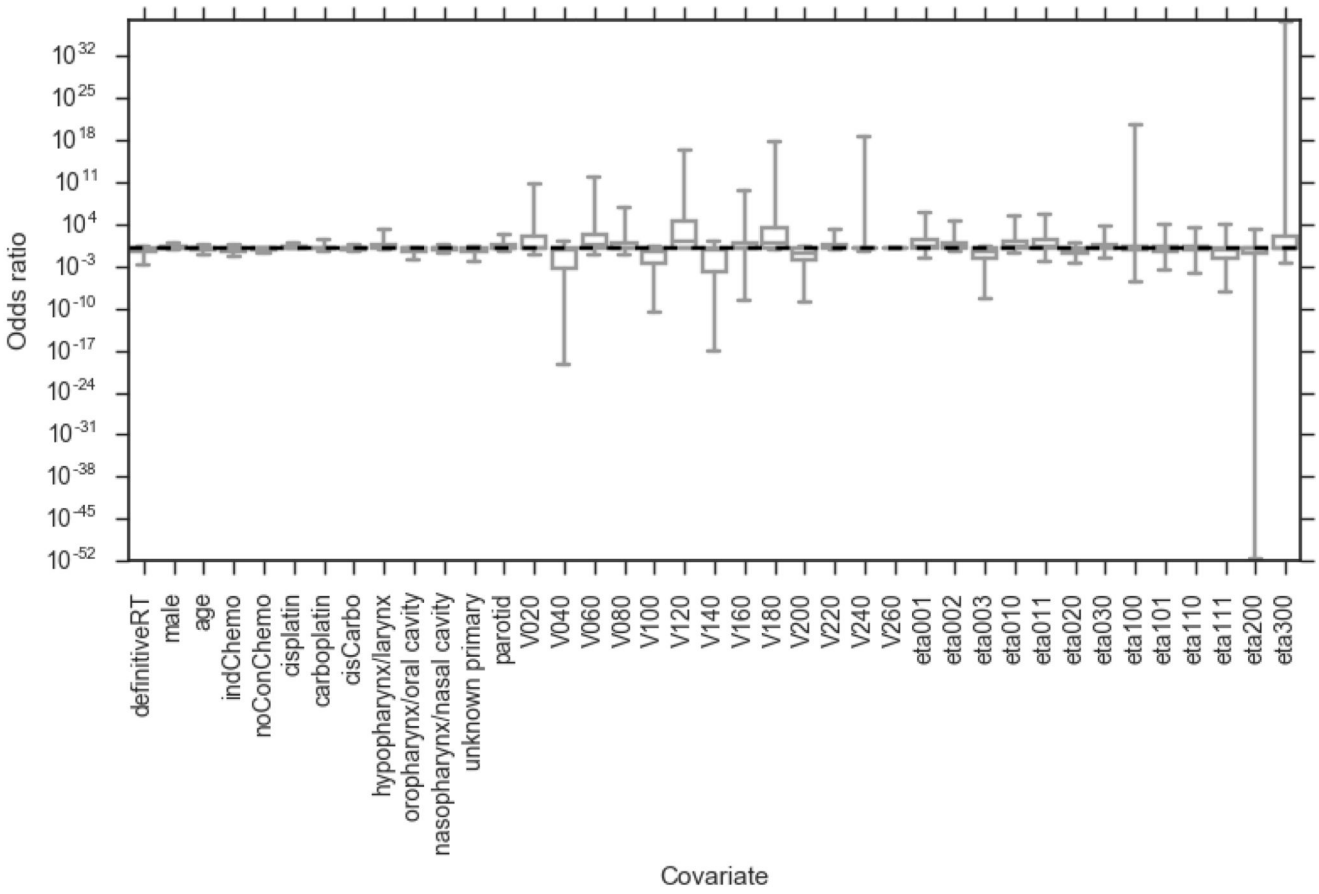

Figure 2.

Bootstrapped (2000 replicates) odds ratios for $PLR_{spatial}$ model. Whiskers show 95 percentiles (non-normal distributions). 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. None of the covariates are significantly associated with severe mucositis.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

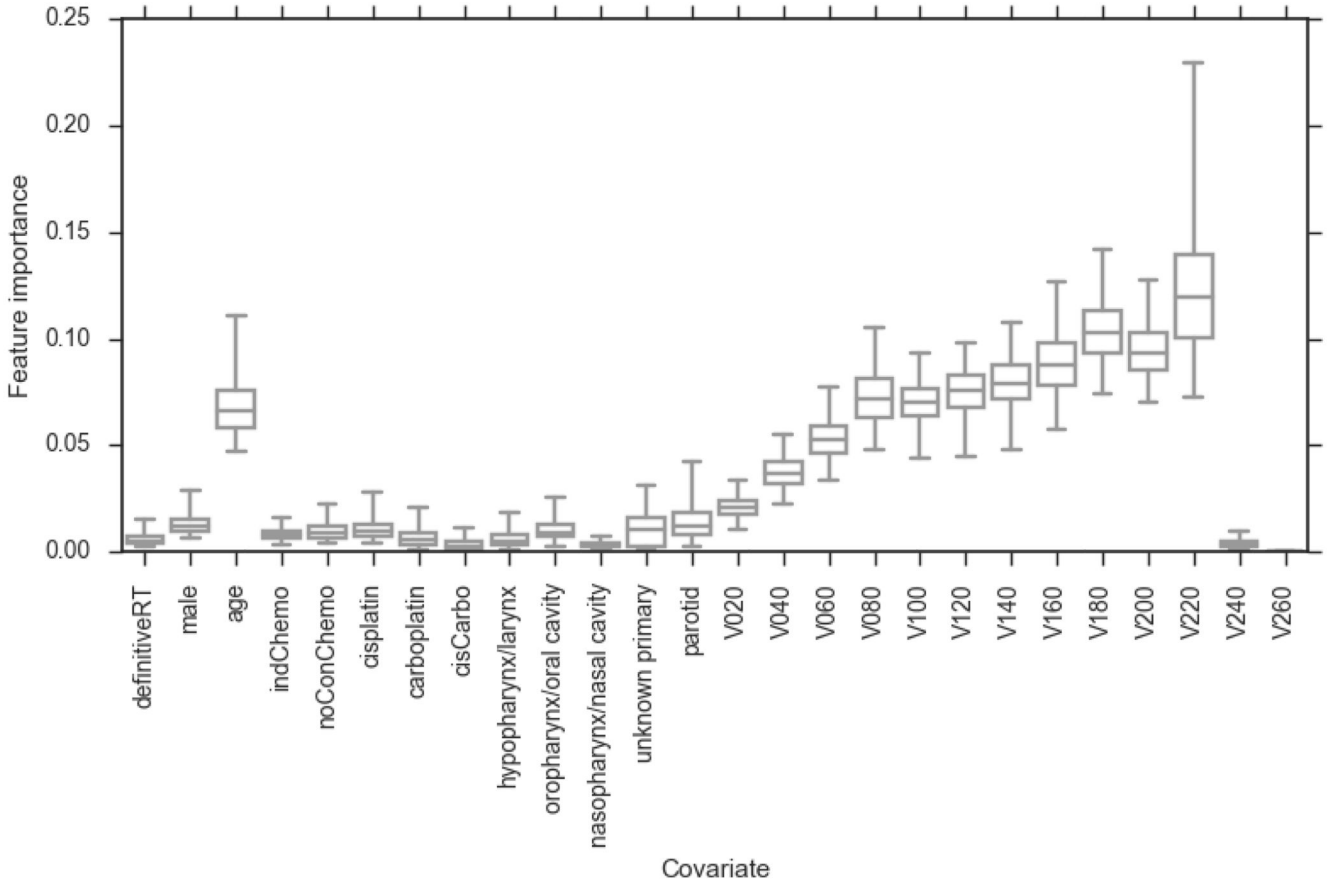

Figure 3.

Bootstrapped (2000 replicates) feature importance measures for $RFC_{standard}$ model. Whiskers show 95 percentiles (non-normal distributions). 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. The feature importance of the dose metrics increases with increasing dose up to V220, which has the highest feature importance of any covariate.

Europe PMC Funders Author Manuscripts

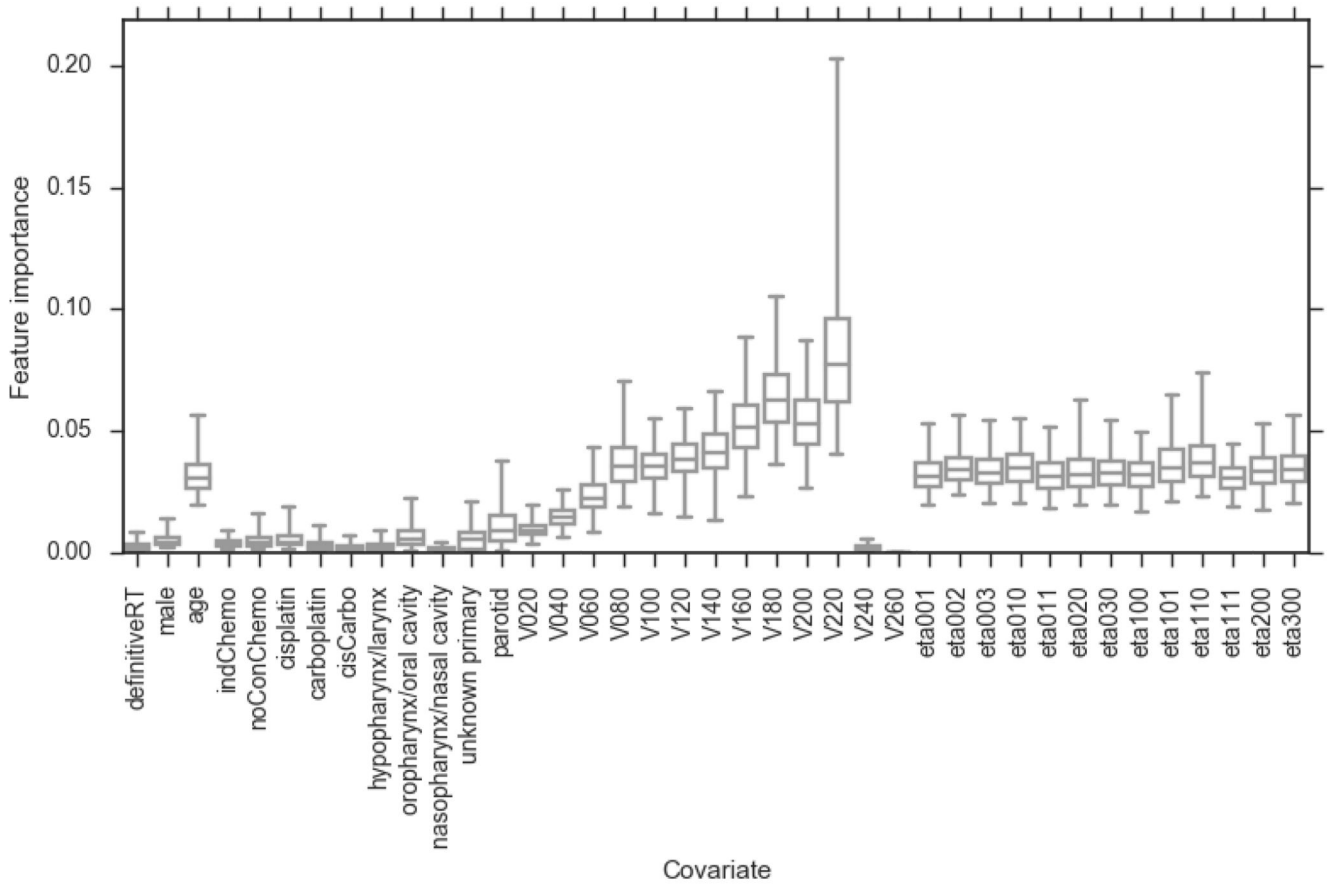

Figure 4.

Bootstrapped (2000 replicates) feature importance measures for $RFC_{spatial}$ model. Whiskers show 95 percentiles (non-normal distributions). 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. The feature importance of the dose metrics increases with increasing dose up to V220, which has the highest feature importance of any covariate.

Europe PMC Funders Author Manuscripts

Table 1
Performance of models on internal validation.

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	0.72 (0.09)	0.66 (0.03)	0.23 (0.02)	12.4 (10.9)	-5.0 (5.2)
$SVC_{standard}$	kernel = radial basis function, C = 0.1, gamma = 0.01	0.72 (0.09)	-	-	-	-
$RFC_{standard}$	max depth = 5, max features = square root	0.71 (0.09)	0.56 (0.08)	0.19 (0.03)	3.9 (2.2)	-1.5 (1.4)
$PLR_{spatial}$	regularisation = LASSO, C = 0.1	0.72 (0.09)	0.66 (0.04)	0.23 (0.02)	11.9 (10.9)	-4.8 (5.2)
$SVC_{spatial}$	kernel = radial basis function, C = 1.0, gamma = 0.001	0.71 (0.09)	-	-	-	-
$RFC_{spatial}$	max depth = 5, max features = square root	0.70 (0.09)	0.56 (0.07)	0.18 (0.03)	4.2 (2.3)	-1.9 (1.6)

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