Elsevier Editorial System(tm) for
International Journal of Radiation Oncology*Biology*Physics
Manuscript Draft

Manuscript Number: ROB-D-16-00714R1

Title: Functional data analysis applied to modeling of severe acute mucositis and dysphagia resulting from head and neck radiation therapy

Article Type: Full Length Article

Section/Category: Physics Contribution

Corresponding Author: Mr. Jamie Adam Dean, MSci

Corresponding Author's Institution: The Institute of Cancer Research and Royal Marsden NHS Foundation Trust, London

First Author: Jamie Adam Dean

Order of Authors: Jamie Adam Dean; Kee H Wong, MSc; Hiram Gay, MD; Liam C Welsh, PhD; Ann-Britt Jones, MSc; Ulrike Schick, PhD; Jung Hun Oh, PhD; Aditya Apte, PhD; Kate L Newbold, FRCR; Shreerang A Bhide, PhD; Kevin J Harrington, PhD; Jospeh O Deasy, PhD; Christopher M Nutting, PhD; Sarah L Gulliford, PhD

Abstract: Purpose
Current normal tissue complication probability (NTCP) modeling using logistic regression suffers from bias and high uncertainty in the presence of highly correlated radiation therapy (RT) dose data. This hinders robust estimates of dose-response associations and, hence, optimal normal tissue-sparing strategies from being elucidated. Using functional data analysis (FDA) to reduce the dimensionality of the dose data could overcome this limitation.

Methods and Materials
FDA was applied to modeling of severe acute mucositis and dysphagia resulting from head and neck RT. Functional partial least squares regression (FPLS) and functional principal component analysis (FPCA) were used for dimensionality reduction of the dose-volume histogram data. The reduced dose data were input into functional logistic regression models (FPLS-LR and FPC-LR) along with clinical data. This approach was compared with penalized logistic regression (PLR) in terms of predictive performance and the significance of treatment covariate-response associations, assessed using bootstrapping.

Results
The area under the receiver operating characteristic curves (AUC) for the PLR, FPC-LR and FPLS-LR models were 0.65, 0.69 and 0.67 for mucositis (internal validation) and 0.81, 0.83 and 0.83 for dysphagia (external validation), respectively. The calibration slopes/intercepts for the PLR, FPC-LR and FPLS-LR models were 1.6/-0.67, 0.45/0.47 and 0.40/0.49 for mucositis (internal validation) and 2.5/-0.96, 0.79/-0.04 and 0.79/0.00 for dysphagia (external validation). The bootstrapped odds ratios indicated significant associations between RT dose and severe toxicity in the mucositis and dysphagia FDA models. Cisplatin was significantly

associated with severe dysphagia in the FDA models. None of the covariates was significantly associated with severe toxicity in the PLR models. Dose levels greater than approximately 1.0 Gy/fraction were most strongly associated with severe acute mucositis and dysphagia in the FDA models.

Conclusions

FPLS and FPCA marginally improved predictive performance compared with PLR and provided robust dose-response associations. FDA is recommended for use in NTCP modeling.

Suggested Reviewers:

Opposed Reviewers:

Cover Letter

Jamie A Dean
Joint Department of Physics
The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust
London
United Kingdom
SM2 5NG

Tel: +442086616223
Email: jamie.dean@icr.ac.uk

07/29/16

Prof Anthony L Zietman
Clark Center for Radiation Oncology
100 Blossom Street
Boston MA, 02114
United States of America

Dear Prof Zietman,

Please find enclosed the revised version of our manuscript entitled: ‘functional data analysis applied to modeling of severe acute mucositis and dysphagia resulting from head and neck radiation therapy’, for exclusive consideration for publication as an article in the International Journal of Radiation Oncology, Biology, Physics. We have made substantial changes to our original manuscript based on the helpful comments of the editor and reviewers and believe that it now represents a far stronger paper.

The paper demonstrates the successful application of a novel statistical modeling approach, that is robust to the multicollinearity that conventionally used methods suffer from, to improve NTCP modeling of severe acute mucositis and dysphagia. As such this paper should be of interest to a broad readership including radiation oncologists treating head and neck cancer, those interested in head and neck radiation therapy dose-response studies, normal tissue toxicity, treatment planning, statistical modeling and clinical decision-support tools.

Thank you for your consideration of our work. Please address all correspondence regarding this manuscript to me at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust and feel free to correspond with me by email (jamie.dean@icr.ac.uk).

Yours sincerely,

Jamie Dea

Jamie Dean

We wish to thank the editor and reviewers for their helpful comments. We have addressed each one as detailed below.

Editorial Comments:
The manuscript investigated the application of functional data analysis to NTCP modeling in comparison to standard logistic regression in cases of head and neck toxicities . The manuscript is well written and the concept was found to be of interest. However, the reviewers have delineated several issues particularly in the interpretation of the evaluation methods and corresponding results that would benefit from further clarifications as delineated below.

Reviewer #1:
No review

Reviewer #2: This is a generally well-written paper that introduces the use of functional data analysis to NTCP models. While I believe that the paper overstates certain key results, in general it is interesting and will provide a valuable addition to the literature.

I have a number of questions about the presentation.

p. 3, authors state that "logistic regression assume[s] all the covariates are independent." While LR will work better in a number of ways it the predictors are independent, LR does not "assume" that.
We have rephrased this in line with your comment (p.3 par.2).

p. 3, what does "structure of the correlations is often consistent between patients" mean?
We have clarified this (p.3 par.2).

p. 3, there are a number of references to "unstable" regression coefficients. Authors should use the correct statistical term, since I am not sure exactly what they mean by "unstable" in this context.
We have removed this term and described this phenomenon in terms of uncertainty and bias (p.3 par.2 and throughout manuscript).

p. 3, I do not understand "...values for the strength of the correlation between the dosimetric variables and toxicity that are highly sensitive to the training data and so do not generalize well to new patients..." This seems to imply that modeling works better if the predictors are uncorrelated to the outcomes.
We have rewritten this sentence to clarify it (p.4 par.1).

p. 3, "which dose levels are most strongly associated with toxicity" is inexact.
We have rephrased this (p.4 par.1).

p. 4, While it is true that picking the FPCA components with the most variance does not guarantee association with the outcome of interest, I do not agree that there is "no reason" they should be related.

We have changed the wording of this in line with your suggestion (p5. par.1).

p. 5-6 The authors start out with 351 patients, remove some for a number of reasons, and then end up with 183+179=362 patients.
The same patients were used for modeling both toxicities. We have clarified this (p.6 par.1).

p. 6 In the description of the external validation, it appears that a different toxicity criterion was used in the training and validation sets of dysphagia patients. Is this so, if so, why, and what effect does this have on the validation?
A slightly different toxicity endpoint was used due to the data available. We have clarified this (p.6 par. 1) and noted this as a limitation in the discussion (p.21 par.2).

p. 7 Was any effort at dimension reduction (variable selection) attempted for the MLR model? Inclusion of collinear and non-explanatory variables could be a cause of the "instability" the authors observe.
Yes, that is correct. That is the problem we were attempting to overcome using the FDA techniques. We included the MLR model to demonstrate this problem. However, we have repeated the MLR modeling, with the addition of LASSO regularization to reduce the multicollinearity problems by shrinkage of the regression coefficients and variable selection (p.8 par.2). This resulted in more stable (in a way) odds ratios, compared with unpenalized multivariable logistic regression, as most of the odds ratios were set to 1 by the penalization due to the high correlations between the covariates. This led to the PLR models underfitting the data, as evidenced by the calibration slopes being greater than 1.

p. 9 Is equation 5 correct? Does the RHS necessarily resolve to either 0 or 1 for each patient?
Equation 5 is correct (except that there should not be a hat above the y on the left hand side which we have now removed; p.10). The equation describes a linear model so the right hand side does not have to be 0 or 1 in this general case. Functional logistic regression (equation 11) is a functional linear model with a logistic link function and binary (0 or 1) outcome. This is what we used for the FPC-LR and FPLS-LR modeling.

p. 10 More detail is needed on how the BIC was used to select the penalty, especially since the penalty was selected to be 0 (p. 12), which I found remarkable.
We have added more detail about the BIC and penalization (p.11 par.1). Due to a coding error we accidentally left the penalization off (the default setting for this function in fda.usc was to use no penalization). We have now corrected this and rerun all of the analysis.

p. 11 & Table 1 Was bootstrapping used to estimate the AUC (internal validation)? The AUC can be estimated in the bootstrap cycle by classifying the out-of-bag samples. This can also be used to provide confidence intervals for the statistics in Table 1, which need them.

Yes, bootstrapping was used with 2000 replicates and correction for optimism. Bootstrapping with correction for optimism does not provide confidence intervals, but is an alternative technique to estimate performance with correction for overfitting. It is recommended by expert statisticians who provide guidelines for statistical modeling (see, for example, Frank Harrell's textbooks and papers on statistical modeling). We have added additional details on this in the manuscript (p.13 par.3).

p. 12 The Brier Score will be obscure to the readers of the Red Journal and requires some description.
We have added the definition of the Brier score (p.13 par.2).

p. 13 "The decrease in the first FPCA and FPLS component loadings at around 1.8 Gy..." Where is the reader supposed to be seeing this?
This can be seen in figure 1. We have now directed the reader to this figure in the text (p.15 par.2).

p. 14 and Figure 2. What is the reader supposed to conclude from this figure? The text points to it but does not comment on whatever the authors find interesting about it.
We have added an explanation of what the figure indicates (p.17 par.1).

p. 16 "Our results demonstrate that FPC-LR and FPLS-LR produced models with better predictive performance than MLR." This is utterly unsupported. According to Table 1, the AUC of FPC-LR and FPLS-LR is 0.01 better than that of MLR for mucositis, and 0.04 better for dysphagia.
We have toned down the description of the improvement in predictive performance. We have indicated that improvements in discrimination were marginal and the improvements in calibration were larger (p.16 par.3, p.17 par.1).

Table 1. All of the statistics in Table 1 require confidence intervals, which can be generated by bootstrapping
We have already applied bootstrapping with a correction for optimism. We have added additional details on how this was performed (p.13 par.3). Bootstrapping with correction for optimism is an alternative method for estimating performance whilst accounting for overfitting (see, for example, Frank Harrell's textbooks and papers on statistical modeling).

p. 17 "Unlike the FDA models, the MLR models were unable to identify that high doses had higher correlations with toxicity than low doses, as would be intuitively expected." First, the authors need to be specific about the definitions of "high" and "low" doses. Mean? Maximum? This is something of a straw man argument if they did not submit variables to MLR representative of "high dose."
The description of the dose-volume metrics included in each of the models is given in the methods section (p.8 par.2, p.8 par.3). For all of the models the included dose volume metrics went up to V260. We have specified that by high and intermediate doses we mean greater than approximately 1 Gy per fraction (p.19 par.1).

p. 18 All the "unstable" odds ratios observed in MLR models are due to the lack of variable selection prior to estimation. I can't speak for the set of all data analysts, but most of the ones I know who use MLR use it with some sort of variable selection method, or use the lasso.
We have repeated the MLR modeling using LASSO (p.8 par.2).

p. 19 First line of the Conclusion is unsupported by the Results.
We have toned down the description of the improvements in predictive performance (p.22 par.2).

Tables 2-4. It would be more instructive to place the three dysphagia and three mucositis models side-by-side, since the comparison of mucositis with dysphagia is of less interest.
Arranging the tables in the manner suggested would result in lots of blank space in the tables so we prefer to keep them as they are currently arranged.

Reviewer #3: The authors describe the use of Functional Data analysis (FDA) to improve NTCP modeling. Two different methods were applied: functional partial least square regression and functional principal component analysis. The research question addressed from the authors and their findings are important for the readers of Red Journal. The development of this application is important to push research in this field forward. The manuscript is very well written and very readable. The methods are carefully chosen, explained, and evaluated.
I have only few minor comments.

1 "Data from 351 patients....". Please, specify "head and neck RT patients".
We have specified this (p.5 par.3).

In addition, Dysphagia: please clarify if the dysphagia endpoint has been scored using a common definition for both training and external cohort.
We have now specified this and clarified the (slightly different) scoring systems for the training and external validation cohorts (p.6 par.1). We have mentioned this slight difference as a potential limitation (p.21 par.2).

2 The use of fractional DVH offers food for thought. Indeed, the fractional DVH is just a doppelgänger of the total DVH. In order to give a more appropriate dose-volume representation for modeling acute toxicity, the time to acute toxicity event should be considered. A first order more effective alternative approach to total DVH could be to describe acute complication risk as a function of accumulated dose-volume at toxicity appearance. In any way, some radiobiological aspects would be neglected.
We initially considered the time-to-event approach. However, we decided against it as the subjective choice of the treating clinicians of when to initiate a feeding tube intervention would lead to a lot of noise in the cumulative dose delivered up to the time of intervention, which would substantially weaken the study. We have added a note on this to the manuscript (p.8 par.1).

The use of fractional DVH obtained simply as total DVH divided by the number of fractions just focus the attention on the fraction size rather than on the total dose that is a correct approach when different fraction sizes are considered. Because the references are blanked to ensure blind reviewing treatment data are not available to the reviewer, the fraction sizes are not known.
We have added the different fractionation regimens to appendix A. They are quite similar so corrections based on radiobiological models (which were performed in preliminary work) did not make any substantive differences to the results. We have added a note on this to manuscript (p.7 par.3).

This reviewer can suppose a wide variety of fractionation schedules.
The fractionation schedules were similar. They are now listed in appendix A.

The physical dose distribution was converted to the fractional dose distribution (physical dose delivered in each fraction). This, to some extent, accounts for differences in the fractionation schedules. However an EQD2 correction by the Withers formula could have been a better approach. This point should be at least discussed in the Discussion section.
Preliminary work showed that using radiobiological corrections, e.g. BED/EQD2, made negligible differences to the results. We have added a note about this (p.7 par.3).

3. Is the dose level 2.6 Gy the higher dose per fraction?
Yes, there were no doses to any of the OARs higher than 2.6 Gy/fraction. We have added the fractionation regimens used to appendix A to make this clearer.

4. To be fair with the "ridden roughshod over" MVL approach some preprocessing on dosimetric variables should be performed in order to avoid overfitting or at least discuss also this point.
We have repeated the MLR modeling with LASSO regularization to reduce collinearity related problems and overfitting by shrinkage of the regression coefficients and variable selection (p.8 par.2). This resulted in more stable (in a way) odds ratios, compared with unpenalized multivariable logistic regression, as most of the odds ratios were set to 1 by the penalization due to the high correlations between the covariates. This led to the PLR models underfitting the data, as evidenced by the calibration slopes being greater than 1.

5. Page 16. "To the best of our knowledge....the best predictive performance to date". Some numerical comparison and some references would be welcome.
We have added some comparisons (p.18 par.3).

Reviewer #4: The paper investigates the functional data analysis (FDA) as modeling of severe acute mucositis and dysphagia and compared with multivariable logistic regression.FDA models describing the dose-volume histogram as a continuous curve demonstrated better predictive performance and more robust dose- response estimates than MLR. The paper is clearly written and the results sound.

Minor revision
Please provide the definition of 3 or more score of the mucositis and dysphagia grading systems. (page 6, lines 1).
We have added the definitions for these (p.6 par.1).

Authors should clarify te reason to exclude patients with a peak score below 3. (see page 6 lines 7-9)
We have added additional detail on the justification for this approach in appendix B and added a reference (p.6 par.1).

Reviewer #5: There are a number of issues within this paper that require substantial explanation or clarification to warrant publication. Adding a statistician to this paper would greatly benefit it.
We already have multiple authors with expertise in statistical modeling.

* It is surprising that the external validation dataset outperforms the original training dataset. This may be due to overfitting as the external validation set includes only 90 participants with unknown number of events, yet >12 variables are being fit in the model.
To clarify, we did not fit any models using the external validation data. We just used the external validation data to make predictions using the models fit on the training data and test those predictions. There is no reason that the external validation should not outperform the internal validation. Plenty of other papers have external validation results that are better than internal validation results, for example Buettner et al. 2012 Radiother Oncol. We corrected our internal validation results for optimism due to overfitting. Many other studies do not do this so their internal validation results may be overly optimistic and so their external validation results are more likely to be worse than the internal validation results. We have added the severe toxicity incidences in the training and external validation cohorts (p.6 par.1). We have now noted the size of the external validation cohort as a limitation (p.21 par.2).

* How is it possible for the estimate of the odds ratio in the MLR Dysphagia model to be outside of the 95% CI? In fact, many estimates of the odds ratios lie outside the 95% CIs. This indicates some type of coding error or biased estimation.
This is due to bias caused by collinearity. This was one of the problems with MLR modeling with correlated dose metrics that we were trying to demonstrate. We have repeated the MLR modeling using LASSO regularization to reduce bias in the regression coefficients (p.8 par.2). This resulted in more stable (in a way) odds ratios, compared with unpenalized multivariable logistic regression, as most of the odds ratios were set to 1 by the penalization due to the high correlations between the covariates. This led to the PLR models underfitting the data, as evidenced by the calibration slopes being greater than 1.

* Why and how were the categories of dose selected for the MLR model? It appears that perhaps the FPC and FPLS models do better due to poor model construction of the MLR model- we expect unstable coefficient estimates if the discretized dose variables do not include much information- would a different discretization produce different results for MLR? The entire MLR models are unstable due to poor variable selection, yet model selection criteria was used with the FDA models.

The discretization encompasses the entire range of dose levels with plenty of granularity to capture the shape of the DVHs. We have added a line justifying the choice of dose metrics to include (p.8 par.2). We believe that the instability is not related to the discretization, but that it is due to multicollinearity. The FDA methods were employed in order to overcome this limitation. We have now improved our MLR modeling using LASSO regularization, which performs regularization and variable selection to reduce overfitting and the instability problems caused by multicollinearity (p.8 par.2). A different discretization would not be expected to have much influence on the results unless the number of DVH points selected was too small to properly describe the shapes of the DVHs.

* It is argued that it is challenging to interpret results from PCA, but it seems similarly challenging to interpret results from FDA since the sign of the estimates do not indicate direction of effect. How does a clinician actually use one of these models? They see the significance of the effect then must go into the loadings to see where dose has high loadings and then attempt to avoid those doses? What about efficacy?

The statement that PCA is challenging to interpret was from the referenced paper. However, we do not fully agree with this criticism of PCA so have removed this statement from our manuscript. Your interpretation of how to use the models is correct. OAR-sparing should not compromise tumor coverage. We have added a note on this (p.19 par.1).

* Are the same individuals used in both the mucositis and dysphagia models? A CONSORT diagram of inclusion of patients would be useful since the numbers are confusing (351 total patients but after exclusion, 183 and 179 remain for the models, if not the same patients, then 362 patients remain post exclusion which is greater than the initial number).

Yes it is the same patients for mucositis and dysphagia. We have clarified this (p.6 par.1).

What are the consequences of using the same patients in these 2 models- do we expect models to be similar?

There is a correlation between the two different toxicities endpoints, as patients who had large volumes irradiated to high and intermediate doses were more likely to experience severe toxicity, but this does not effect the modeling of either toxicity.

* How does the external validation dataset compare to the training set in terms of characteristics, proportion of toxicities, etc?

We have added details on this (p.6 par.1 and appendix C).

* On page 7, it claims the FDA is penalized, but later in the manuscript it says the penalty was 0 (unpenalized).
Due to a coding error we accidentally left the penalization off (the default setting for this function in fda.usc was to use no penalization). We have now corrected this and rerun all of the analysis.

* Equation 5 denotes an estimated toxicity outcome but the right-hand side includes an error. Either the left-hand side should not contain the hat/estimate or the right-hand side should contain hats and no error
We have removed the hat from the left hand side (p.10 eq.5).

* It is not obvious how equation 7 is found. An appendix of the derivation by substituting equations 4 and 6 into 5 could be beneficial or a reference to the derivation.
We have added the reference to this derivation (p.10, ref.29).

* The bootstrapping process could be further explained- how was it stratified?
We have added more detail to the explanation of the bootstrapping procedure (p.13 par.3).

* The brier scores and AUCs are nearly identical for all three options, so the only discriminating factor is the calibration which is not overly impressive except in the external validation set (which is likely due to overfitting)
We have toned down our description of the benefit in predictive performance using the FDA models. We have indicated that improvements in discrimination were marginal and the improvements in calibration were larger (p.16 par.3, p.17 par.1). There cannot be any overfitting in the external validation as none of the models were fit using that data. None of the external validation data was "seen" during model training.

* There is no discussion how in the FDA models, the clinical covariates are unstable and the tradeoff between clinical and dose estimation.
We have added comments on this to the discussion (p.20 par.2).

* It would be interesting to apply PCA to this study as a comparator
We agree that it would be interesting to compare PCA to the models presented here. It would also be interesting to compare many other models. However, for the sake of clarity we feel that it would be better not to add in extra models, such as PCA. Studies using PCA for NTCP modeling have already been performed (and are referenced in our manuscript; p.4 par.2).

* Minor- results on page 12 are mixed up: 2 components for the mucositis FPLS model and 1 component for the mucositis and dysphagia model for FPCA and for dysphagia FPLS (or table labeling is mixed).
The text was incorrect. We have now corrected it (p.15 par.1).

* Minor- on page 3, should 50 Gy be 50 cGy?
We have generalized this to Vx (p.3 par.2).

We changed the bootstrapping so that it includes the model selection step within each replication to prevent optimistic internal validation. We have noted this in the manuscript (p.14 par.1) and rerun all of the analysis with the improved bootstrapping approach.

Title Page

Title:

Functional data analysis applied to modeling of severe acute mucositis and dysphagia resulting from head and neck radiation therapy

Short title:

Functional data analysis for mucositis and dysphagia modeling

Authors:

Jamie A Dean MSci[a], Kee H Wong MSc[b], Hiram Gay MD[c], Liam C Welsh PhD[b], Ann-Britt Jones MSc[b], Ulrike Schick PhD[b], Jung Hun Oh PhD[d], Aditya Apte PhD[d], Kate L Newbold FRCR[b,e], Shreerang A Bhide PhD[b,e], Kevin J Harrington PhD[b,e], Joseph O Deasy PhD[d], Christopher M Nutting PhD[b,e], Sarah L Gulliford PhD[a]

[a] Joint Department of Physics at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, London, UK, SM2 5NG

[b] Head and Neck Unit, The Royal Marsden NHS Foundation Trust, Fulham Road, London, UK, SW3 6JJ

[c] Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO, USA

[d] Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[e] Division of Radiotherapy and Imaging, The Institute of Cancer Research, Fulham Road, London, UK, SW3 6JJ

Corresponding Author:

Jamie A Dean

Address: Joint Department of Physics at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, London, UK, SM2 5NG

Telephone: +442089156223

Fax: +442086433812

Email: jamie.dean@icr.ac.uk

Conflict of interest:

None.

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). None of the funding sources had any role in any part of the study or manuscript preparation and submission. We wish to thank Hannah Eyles, Emma Wells, James Morden and Dr Emma Hall at The Institute of Cancer Research Clinical Trials and Statistics Unit for data collation and Dr Cornelis Kamerling, Dr Alex Dunlop, Dr Dualta McQuaid, Dr Simeon Nill and Prof Uwe Oelfke for general support.

Summary

Normal tissue complication probability modeling using logistic regression (LR) suffers from bias and uncertainty in the presence of highly correlated radiation therapy dose data. Consequently robust estimates of the dose levels most strongly associated with toxicity and, potentially, predictive performance are limited. To overcome this limitation, functional data analysis (FDA), which describes the dose-volume histogram as a continuous curve, was applied to modeling of severe acute mucositis and dysphagia and compared with LR. FDA models demonstrated slightly better predictive performance and more robust dose-response estimates than LR.

Manuscript

Abstract

Purpose

Current normal tissue complication probability (NTCP) modeling using logistic regression ~~is unstable~~suffers from bias and high uncertainty in the presence of highly correlated radiation therapy (RT) dose data. This hinders robust estimates of dose-response associations and, hence, optimal normal tissue-sparing strategies from being elucidated. Using functional data analysis (FDA) to reduce the dimensionality of the dose data could overcome this limitation.

Methods and Materials

FDA was applied to modeling of severe acute mucositis and dysphagia resulting from head and neck RT. Functional partial least squares regression (FPLS) and functional principal component analysis (FPCA) were used for dimensionality reduction of the dose-volume histogram data. ~~These~~ The reduced dose data were ~~then~~ input into functional logistic regression models (FPLS-LR and FPC-LR) along with ~~non-functional~~ clinical data. This approach was compared with ~~conventional~~ penalized ~~multivariable~~ logistic regression (~~MLR~~PLR) in terms of predictive performance and the significance of treatment ~~the stability of dose~~covariate-response associations, ~~. The stability of regression coefficients was~~ assessed using bootstrapping.

Results

The area under the receiver operating characteristic curves (AUC) for the ~~MLR~~PLR, FPC-LR and FPLS-LR models were 0.~~71~~65, 0.~~72~~ 69 and 0.~~72~~ 67 for mucositis (internal validation) and 0.~~79~~81, 0.83 and 0.83 for dysphagia (external validation), respectively. The calibration slopes/intercepts for the P~~M~~LR, FPC-LR and FPLS-LR models were ~~0.29~~1.6/-0.67~~0~~, 0.45~~1~~/0.4~~3~~7 and 0.~~5~~40/0.4~~4~~9 for mucositis (internal validation) and ~~0.80~~2.5/-0.~~0~~96, 0.79/-0.04 and 0.79/0.00 for dysphagia (external validation). The bootstrapped ~~regression coefficients~~odds ratios indicated significant associations between RT dose and severe toxicity in the mucositis and dysphagia FDA models. Cisplatin was significantly associated with severe dysphagia in the FDA models. None of the covariates was significantly associated with severe toxicity in the PLR models~~were substantially more stable in the FDA models than the MLR models as evidenced by their far narrower confidence intervals~~. ~~High and intermediate d~~Dose levels~~,~~ greater than approximately 1.0 Gy/fraction ~~,~~were most strongly associated with severe acute mucositis and dysphagia in the FDA models.

Conclusions

FPLS and FPCA marginally improved predictive performance compared with PLR and provided robust ~~the stability of estimates of~~ dose-response associations ~~compared with MLR~~. FDA is recommended for use in NTCP modeling.

Introduction

Normal tissue complication probability (NTCP) modeling uses radiation therapy (RT) dose data, often in combination with clinical and biological data, to construct statistical models of RT-induced toxicity. There are two distinct aims of

NTCP modeling: (i) accurate prediction of toxicity outcomes for individual patients and (ii) estimates of associations between treatment covariates and toxicity. Accurate prediction enables clinical decision-support (1), treatment plan comparison, treatment modality selection (2) and personalized dose prescription (3). Robust estimates of associations between covariates and toxicity can inform the design of RT planning interventions aimed at reducing toxicity.

A major weakness of many NTCP models is suboptimal dimensionality reduction of the RT dose distribution (reducing the number of variables used to describe the dose distribution from all of the points on the 3D dose grid to a small number of summary metrics). In order to input dose data into statistical models the 3D dose distribution delivered to an organ at risk (OAR) is reduced to a single or series of scalar metrics, for example maximum dose, mean dose or multiple points sampled from the dose-volume histogram (DVH), such as the volume of an OAR receiving at least ~~*50*~~ *x* cGy (~~*V50*~~*Vx*). Ideally, information from each dose level should be explicitly input into the model to prevent loss of potentially important information. However, due to the nature of the dose deposition within the patient, adjacent dose levels are very highly correlated (4) (appendix D). This is problematic for many statistical modeling methods, such as (the commonly used) logistic regression, which often ~~assume~~ exhibit biased regression coefficients with large standard errors ~~that all covariates are independent~~in the presence of collinearity (5). ~~Since t~~The structure of the correlations is often consistent between patients as the volumes of an OAR receiving adjacent dose

Formatted: Font: Italic

Formatted: Font: Italic

levels are highly correlated for all patients~~,~~. Therefore, if the same or similar treatment techniques are employed, this does not necessarily prevent the models from being able to accurately predict outcomes prospectively for new patients. However, it does result in the ~~unstable~~ regression coefficients of the dosimetric covariates being biased and having large standard errors. The apparent regression coefficients of the dosimetric covariates do not ~~(values for the strength of the correlations between the dosimetric variables and toxicity that are highly sensitive to the training data and so do not~~ generalize well to new patients~~)~~ and, hence, ~~so~~ should not be used to determine ~~which dose levels are most strongly~~the strength of associations between correlated dose metrics ~~ed with~~and toxicity, as is commonly done (6).

A small number of studies have attempted to address this issue through using principal component analysis (PCA) to reduce the dimensionality of the DVH data (7–12). However, PCA ~~was found to be challenging to interpret in the context of NTCP modeling [9,11] and~~ has been shown to perform poorly when the number of predictors (DVH points) is comparable to, or larger than, the number of observations (patients), as is often the case in NTCP modeling (9, 11). Functional data analysis (FDA) is a statistical framework for analyzing continuous curves rather than discrete measurements (13). Treating an entire curve, for example, a DVH curve, as a single entity removes the problem of correlation (14) and explicitly retains the relationship between points on the DVH curve, which standard, non-functional, statistical techniques do not capture. Data are represented as curves through the use of basis functions. There are different types of basis function including *a priori* fixed bases, such as splines or wavelets, and data-driven bases, for example functional principal component analysis (FPCA). Functional logistic regression uses functional data to predict binary outcomes. It is well suited to NTCP modeling due to the ~~functional~~ continuous nature of DVH curves and the binary nature of toxicity endpoints. Functional logistic regression has recently been applied to NTCP modeling by Benadjaoud *et al.* (15), using FPCA (16) for dimensionality reduction of the DVH data. However, FPCA (and non-functional PCA) is unsupervised (does not use outcome data), which may be a limitation for NTCP modeling. ~~There is no reason why t~~The FPCA components with the most variance in the RT dose data ~~should~~ may not be the ones that are most strongly associated with the toxicity outcome of interest. Functional partial least squares regression (FPLS) (17, 18) is a supervised analogue of FPCA. It overcomes this limitation through generating uncorrelated covariates (FPLS components) in the linear space of the predictors, accounting for the correlation between those predictors and ~~toxicity~~outcome, in this case toxicity. As PLS (and FPLS) uses the outcome (toxicity) data in establishing the components it often outperforms PCA (and FPCA) in prediction tasks (19). However, due to the inclusion of outcome data, it is more susceptible to overfitting.

In this study we ~~apply~~ applied FPLS and FPCA to NTCP modeling of severe acute mucositis and dysphagia. We compared our novel application of FDA with non-functional penalized ~~multivariable~~ logistic regression (~~MLR~~PLR) models. The aims of this study were to (i) determine whether using FPLS or FPCA to reduce the DVH data would improve ~~discriminative ability~~predictive performance compared with P~~M~~LR and (ii) assess whether FPLS or FPCA would lead to more robust estimates of associations between DVH data and toxicity than P~~M~~LR.

Methods and Materials

Patient data

Data from 351 head and neck RT patients, enrolled in one of the six different clinical trials (20–24)[ISRCTN XXXX], were used to train and internally validate severe acute mucositis and dysphagia NTCP models. Data from the same patients were used for the modeling of both toxicities. This dataset ~~has previously been~~is described in appendix A and ~~detail~~ [XXXX et al. 2016a (in press), XXXX et al. 2016b (under review)]. Mucositis and dysphagia~~Toxicity~~ were~~as~~ both consistently scored for all studies using the Common Terminology Criteria for Adverse Events (CTCAE) versions 2 (25) or 3 (26) instruments. The mucositis and dysphagia grading systems are near equivalent in both versions. Both ~~T~~toxicities were recorded, prospectively, for all patients prior to the start of RT, weekly during RT, and at 1 - 4 and 8 weeks post-RT by head and neck cancer specialists, trained in the use of the scoring systems, using standard trial protocols. The toxicity outcome was defined as the peak grade of toxicity, dichotomized into grade 3 or worse (severe) and less than grade 3 (non-severe). Grade 3 mucositis corresponds to confluent mucositis and grade 3 dysphagia corresponds to feeding tube dependence for more than 24 hours. Patients with baseline toxicity were excluded from the analysis. To reduce bias at the expense of statistical power, patients with any missing toxicity scores and a peak score below 3 were excluded from the analysis. A detailed justification for this approach is provided in appendix B and [XXXX et al. 2016a (in press)]. Of the 351 patients, ~~This left~~ 183 met the inclusion criteria for mucositis modeling (severe mucositis incidence = 73%) and 179 ~~patients for mucositis and~~met the inclusion criteria for dysphagia modeling (severe dysphagia incidence = 66%)~~, respectively~~. Ninety head and neck RT patients treated at XXXX with acute dysphagia data available were used as an external validation cohort for the dysphagia models (severe dysphagia incidence = 48%). In this cohort severe acute dysphagia was defined as the requirement for percutaneous endoscopic gastrostomy (PEG) tube insertion. It should be noted that there was a slight difference in the scoring systems due to the data available. All centers involved in treating patients included in this study employed a reactive approach to PEG-insertion, that is, delaying insertion until deemed clinically necessary.

Induction chemotherapy (yes or no), concurrent chemotherapy regime (cisplatin, carboplatin, one cycle of cisplatin followed by one cycle of carboplatin or none), definitive versus post-operative RT, primary disease site (oropharynx/oral cavity, nasopharynx/nasal cavity, hypopharynx/larynx, parotid gland or unknown primary), age and sex were also included as covariates in the models. Concurrent chemotherapy was administered in two cycles, on day 1 and day 29 of RT. A comparison of the clinical covariate data in the training and external validation data sets is provided in appendix C.

RT dose data

The extended oral cavity [XXXX et al. 2016a (in press)] and pharyngeal mucosa [XXXX et al. 2016b (under review)] were contoured by clinical oncologists and used as OARs in the mucositis and dysphagia models, respectively. The physical dose distribution was converted to the fractional dose distribution (physical dose delivered in each fraction). This has been shown to be appropriate for NTCP modeling of acute toxicity (27) as the toxicities often develop before the total dose is administered. The fractional dose distribution was described by the normalized cumulative dose-volume histogram (DVH). Preliminary work indicated that corrections for different fractionation regimens based on radiobiological models made negligible difference to the results. This is due to the fact that the fractionation regimens employed (appendix A) were similar. An alternative approach would be to use the cumulative dose delivered up to the appearance of the toxicity endpoint. However, subjective choice of the treating clinicians of when to initiate a feeding tube intervention would lead to substantial noise the in the cumulative dose delivered up to the time of intervention.

Penalized ~~Multivariable~~ logistic regression model

For the non-functional model the fractional DVH curves were discretely sampled from 0.2 Gy to 2.6 Gy at 0.2 Gy intervals. This sampling was chosen to encompass the entire range of OAR doses with enough granularity to capture the shapes of the DVHs. These DVH measurements were input into a P~~M~~LR model along with the clinical covariates. Penalization was performed using the least absolute shrinkage and selection operator (LASSO) (28). LASSO reduces the magnitude of the regression coefficients, setting some to 0, in order to prevent overfitting. In the context of correlated variables, it reduces the impact of multicollinearity. The penalization strength was selected by 10-fold cross-validation with the value producing the highest average (over all of the cross-validation folds) area under the receiver operating characteristic curve (AUC) on cross-validation selected.

Functional data analysis

The fractional DVH curves (sampled from 0 Gy to 2.60 Gy in 0.01 Gy intervals) were represented using penalized FPCA (16, 29) and penalized FPLS (17, 30) basis functions. FPCA is a dimensionality reduction technique that represents the functional DVH data as orthonormal vector components explaining the maximum variance between patients in the DVH curves. The orthonormality constraint removes the collinearity in the dose metrics used for subsequent modeling and, hence, overcomes the limitations associated with modeling collinear data. The functional principal components $\{\xi_k(d)\}_{k=1}^{\infty}$ represent the functional DVH data (normalized volume as a function of dose for patient i), $V_i(d)$ as the sum of the eigenfunctions, $\xi_k(d)$ weighted by their coefficients, c_{ik}:

$$V_i(d) - \mu(d) = \sum_{k=1}^{\infty} c_{ik}\xi_k(d) \qquad (1)$$

where $\mu(d)$ is the mean $V(d)$ and c_{ik} describes the score for functional principal component k for the DVH of patient i and is given by

$$c_{ik} = \int (V_i(d) - \mu(d))\xi_k(d)dd \quad (2)$$

The eigenfunctions, $\{\xi_k\}_{k=1}^{\infty}$ and their corresponding eigenvalues (describing the amount of variance explained by each eigenfunction), $\lambda_1 \geq \lambda_2 \geq \cdots$, are determined by eigendecomposition (factorization into eigenvalues and eigenvectors) of the covariance operator, Σ where

$$\Sigma(d_1, d_2) = \mathrm{Cov}[V(d_1), V(d_2)] = \mathrm{E}\left[\big(V(d_1) - \mu(d_1)\big)\big(V(d_2) - \mu(d_2)\big)^{\mathrm{T}}\right] \quad (3)$$

where $\mu(d_1)$ is the mean volume receiving dose d_1. *V(d)* can be approximated by a small number of principal components, k_n, assuming that $c_{ik} = 0$ for $k > k_n$, to achieve dimensionality reduction to a small number of basis functions efficiently describing the variation between patients in the DVH data:

$$V_i(d) \approx \mu(d) + \sum_{k=1}^{k_n} c_{ik}\xi_k(d) \quad (4)$$

The eigenfunctions and their coefficients can then be used in subsequent analyses. The FPCA components can be used to estimate a toxicity outcome for patient *i*, ~~$\hat{y}_i$~~y_i using a functional linear model (29, 31):

$$\cancel{\hat{y}_i} y_i = \alpha + \int \beta(d)V_i(d)dd + \varepsilon_i \quad (5)$$

where α is the intercept and ε_i is a centered random error.

When FPCA is used to describe the DVH data *β(d)* represents a “weighting function” describing the amount of variation between patients at all dose levels on the DVH. It can be approximated by k_n eigenfunctions:

$$\beta(d) = \sum_{k=1}^{\infty} \beta_k \xi_k(d) \approx \sum_{k=1}^{k_n} \beta_k \xi_k(d) \quad (6)$$

~~Substituting equations 4 and 6 into equation 5 allows a~~An estimate of the response, $\hat{y}_i$ ~~to~~ can be made using (derivation in (29)):

$$\hat{y}_i = \alpha + \int \beta(d) V_i(d) dd \approx \alpha + \sum_{k=1}^{k_n} \hat{\beta}_k c_{ik} \quad (7)$$

where

$$\hat{\beta}_{(1:k_n)} = \left(\frac{c_{.1}^{\mathrm{T}} y}{n\lambda_1}, \dots, \frac{c_{.k_n}^{\mathrm{T}} y}{n\lambda_{k_n}} \right) \quad (8)$$

The model was fit to the data, placing penalization on the curvature (second derivative) of the eigenfunctions, by

$$\hat{y}_i = \xi_k \left(\xi_k^{\mathrm{T}} \xi_k + r \xi_k^{\mathrm{T}} \mathbf{P} \xi_k \right)^{-1} \xi_k^{\mathrm{T}} y_i \quad (9)$$

where r is the amount of penalization, **P** is the vector (0, 0, 1) that defines the penalty matrix such that the second derivative (curvature) is penalized and y is the actual response data. The choice of which components to include (within the first 5 components) and the ~~amount of penalization~~magnitude of the roughness penalty, r to apply (selected from a set of values in the range from 0 to 1350), to best estimate the toxicity outcomes, were determined using model selection criteria (MSC) (16) with the Bayesian information criterion:

$$MSC(k_n) = \log \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right] + \frac{\log(n) k_n}{n^2} \quad (10)$$

where n is the number of patients and y_i is the actual outcome (toxicity) data. This penalizes the model complexity to reduce overfitting. Models with different values of r and k_n were generated and the combination of values that minimized MSC was selected. The FPCA or FPLS components included affect the smoothness of the estimate of the $\beta(d)$ function as the dominant mode of variation tends to be smooth and roughness tends to increase for subsequent modes of variation, in part due to the orthogonality constraint.

FPLS is similar to FPCA, but uses the response (toxicity) data in constructing the FPLS components (17), $\{\tilde{\xi}_k\}_{k=1}^{\infty}$ in order to establish orthogonal components that have maximum covariance to the response. This is achieved through maximizing the squared covariance between $V_i(d)$ and the response, y_i with the constraint that all components are mutually orthogonal (30). This takes the place of the eigendecomposition used for FPCA, described in equation 3. The iterative algorithm used to compute the FPLS components is described in (32). When FPLS is used for dimensionality reduction of the DVH data, *β(d)* can be interpreted as a data-driven "weighting function" for the importance of each dose level in causing severe toxicity. It is important to consider that, as this is a data-driven approach, the "weighting function" is an estimate of the "true weighting function" over the range of available data and is influenced by the structure (i.e. distribution in dose-volume space) of the available data. MSC was performed for the FPLS analysis in the same manner as for FPCA. The FPCA and FPLS ~~components were~~analysis was bootstrapped with 2000 ~~stratified~~ replicates to assess the~~ir~~ ~~stability~~uncertainty of the shapes of the components.

The optimal FPCA and FPLS components (those producing the lowest MSC) were used as basis functions as input into functional logistic regression (33, 34) models (FPC-LR and FPLS-LR) along with the (non-functional) clinical covariates. The functional logistic regression model describes the probability of patient *i*, experiencing severe toxicity $P(y_i = 1)$, and is given by

$$\ln\frac{P(y_i = 1)}{P(y_i = 0)} = \alpha + \sum_{j=1}^{p} \beta_j Z^j + \int \beta(d) V_i(d)\, dd \approx \alpha + \sum_{j=1}^{p} \beta_j Z^j + \sum_{k=1}^{k_n} \beta_k c_{ik} \qquad (11)$$

using the substitution for the functional linear model described in equation 7, where α is the intercept and $\{Z^j\}_{j=1}^{p}$ are the non-functional covariates with regression coefficients $\{\beta_j\}_{j=1}^{p}$. Maximum likelihood estimation of the regression coefficients was performed using iteratively weighted least squares.

Model comparisons

The predictive performance and generalizability of the models (addressing aim i) were assessed in terms of discrimination, calibration and overall performance on internal validation, and additionally for the dysphagia models on external validation. The discriminative abilities of the models were assessed using the ~~area under the receiver operating characteristic curve (~~AUC~~)~~. Calibration was evaluated by the slope and intercept of a logistic regression model of the actual toxicity against the predicted probability of severe toxicity (35, 36). Overall model performance was measured using the Brier score (37), BS. It is defined as

$$BS = \frac{1}{N}\sum_{t=1}^{N}(p_t - y_t)^2 \quad (12)$$

where p_t is the predicted probability, y_t is the actual outcome and N is the number of predictions. The score takes a value between 0 and 1 with lower values indicating better performance.

For the internal validation the performance~~se~~ metrics were "corrected for optimism" using bootstrapping with 2000 ~~stratified~~ replicates (38). The optimism-corrected performance metrics, $M_{corrected}$ were calculated by

$$M_{corrected} = M_{apparent} - O \quad (13)$$

where $M_{apparent}$ is the performance metric, for example AUC, calculated using all of the training data to, both, fit the model and evaluate its performance, and the optimism, O is given by

$$O = \frac{1}{B}\sum_{b=1}^{B}(M_{b,boot} - M_{b,orig}) \quad (14)$$

where B is the number of bootstrap replicates, $M_{b,boot}$ is the performance metric calculated using the bootstrap dataset b to, both, fit and evaluate model performance, and $M_{b,orig}$ is the performance of the model fit using the bootstrap dataset b evaluated on the original dataset. This provides an unbiased estimate of internal validity, penalizing for overfitting. Model hyper-parameter tuning, such as the selection of the amount of penalization for the PLR models and the selection of components and penalization for the FDA models, was performed for each bootstrap replicate. This prevents any “data leakage” from the training data into the internal validation data. The dysphagia models were used to predict severe dysphagia probability for the external validation cohort. Those predictions were compared to the actual PEG-dependence data for the cohort and the same performance metrics calculated. The ~~stabilities~~ uncertainties of the ~~regression coefficients~~odds ratios (addressing aim ii) were assessed by calculating ~~the width of their~~ bootstrapp~~ed~~ed ~~-~~95~~%~~ percentile confidence intervals with 2000 replicates. Statistical analysis was performed using the statistical computing R language version 3.2.4 (39) and the fda.usc version 1.2.2 (40), glmnet version 2.0 (41), rms version 4.5 (42) and val.prob.ci.2 (43) packages.

Results

For FPCA, the variances in the DVH data explained by the first ~~three~~ five ~~functional principal~~FPCA components were 80.8%, 12.5% and 3.7%, 1.2% and 0.6% for mucositis and 70.8%, 14.5%, ~~and~~ 5.6%, 4.4% and 1.6% for dysphagia. For FPLS, the ~~co~~variances ~~between the DVH data and severe acute toxicity~~ explained by the first ~~three~~ five ~~functional partial least squares~~FPLS components were ~~80.5~~78.1%, ~~17.5~~16.9%, ~~and~~ 2.0%, 2.5% and 0.6% for mucositis and ~~79.4~~76.2%, ~~9.0~~8.6%, ~~and~~ 11.~~6~~2%, 2.7% and 1.3% for dysphagia. The model selection resulted in ~~only~~ the first two ~~FPCA or FPLS~~ components being selected for the ~~dysphagia~~ FPCA and FPLS mucositis models and only the ~~FPLS mucositis model and the~~ first ~~two~~ component~~s~~ ~~being~~ selected ~~for~~in both of the ~~mucositi~~dysphagia~~s~~ ~~FPCA~~ FDA models. Penalization of 1342 was chosen by the

Formatted: Font: Cambria, 12 pt

Formatted: Font: Cambria, 12 pt

model selection for the mucositis FPCA model, 0 for the mucositis FPLS model and 1350 for both of the dysphagia FDA models.

~~Penalization of 0 was selected for all four FDA models.~~

Formatted: Font: Cambria, 12 pt

Figure 1 shows the first FPCA and FPLS components for the mucositis and dysphagia models, respectively. Bootstrapping the FPCA and FPLS indicated that the shapes of the first FPCA and FPLS components were very similar irrespective of the random sample of patients selected. There was a general trend that the FPCA and FPLS loadings increased with increasing dose and sharply decreased to 0 at the maximum dose. The FPCA components indicate that the variation, between patients, in the volume of OAR irradiated to a certain dose level increased with increasing dose level. The same trend in the FPLS components indicates that higher doses were more strongly associated with severe toxicity. The decrease in the first FPCA and FPLS component loadings at around 1.8 Gy (figure 1) for the dysphagia training data is indicative of reduced variation in this region of the DVHs between patients. This is likely to be due to the fact that most of the variation in the pharyngeal mucosa dose distribution between patients is related to the variation in volume of overlap of the two different planning target volumes (whose prescription dose levels correspond to the positions of the two peaks in the FPCA and FPLS components) with the pharyngeal mucosa.

For the P~~M~~LR, FPC-LR and FPLS-LR modeling, oropharynx/oral cavity and no concurrent chemotherapy were removed as covariates to prevent perfect collinearity (correlation matrices are shown in ~~the~~ appendix D). Odds ratios for other primary disease sites are thus relative to oropharynx/oral cavity and odds ratios for concurrent chemotherapy are relative to no concurrent chemotherapy.

In regards to aim i, the predictive performance of the three different mucositis and dysphagia models, as assessed by internal and external (for dysphagia) validation, is displayed in table 1. The ~~MLR, FPC-LR and FPLS-LR~~ mucositis models ~~all~~ had modest (PLR and FPLS-LR) or modest-to-good (FPC-LR) discriminative ability (using the interpretation in (44)) on internal validation. The discriminative abilities and overall performances of the FPC-LR and FPLS-LR models were ~~slightly~~ marginally better than the P~~M~~LR model. Calibration was relatively poor for all of the models, with the ~~The~~ FPC-LR and FPLS-LR models overfitting the data (calibration slope less than 1) ~~had substantially better calibration (slope closer to 1 and intercept closer to 0) than~~and the P~~M~~LR model underfitting the data (calibration slope greater than 1). ~~However, none of the mucositis models demonstrated good calibration.~~The underfitting exhibited by the PLR models was likely due to over shrinkage of the regression coefficients by the LASSO penalization caused by high multicollinearity. It should be noted that the "~~However, none of the mucositis models demonstrated good calibration.~~correction for optimism" may have improved the calibration of the PLR models, as they underfit the data.

The discrimination and calibration of the dysphagia models were better than the mucositis models. ~~The MLR~~All three dysphagia models had good discriminative ability on internal validation~~, whilst the FPC-LR and FPLS-LR models showed good-to-excellent discrimination~~. The discriminative abilities of all three models increased on external validation, with the P~~M~~LR model demonstrating good-to-excellent discrimination and the FPC-LR and FPLS-LR models showing excellent discrimination. The overall performance of all of the ~~FPC-LR and FPLS-LR~~ models was ~~slightly better than the MLR model~~ similar, both on internal and external validation. ~~The FPC-LR and FPLS-LR models had substantially better calibration than the MLR model on internal validation.~~ Calibration of all of the models on internal validation was modest, with the PLR model underfitting the data and the FDA models overfitting the data. The FPC-LR and FPLS-LR models had substantially better calibration than the PLR model on external calibration.~~All of the dysphagia models were well calibrated on the external validation data and had similar calibration slopes and intercept.~~ The FPLS-LR model had ~~slightly~~ marginally better calibration than the ~~other~~ FPC-LR model~~s~~ on external calibration. A logistic calibration curve for the external validation of this model is shown in figure 2. The curve lies close to the identity line indicating good model calibration on external validation.

Concerning aim ii, the results of the bootstrapped penalized and functional logistic regression odds ratios are shown in tables 2 – 4. The odds ratios for the ~~dosimetric~~ covariates in the P~~M~~LR models ~~are highly unstable~~were often set to 1 by the LASSO penalization ~~as evidenced by their wide bootstrapped confidence intervals~~. In the mucositis and dysphagia P~~M~~LR models none of the ~~dosimetric~~ covariates ~~was~~ was significantly associated with severe toxicity. ~~and in the dysphagia MLR model only the V140 was significantly associated with severe toxicity (although the odds ratio was outside the 95% confidence interval and the association with severe dysphagia was negative whereas there was a positive~~

~~association between 95% confidence limits and severe dysphagia). The negative association is likely a result of multicollinearity.~~ Conversely, there was a significant association between the first FPLS component and severe toxicity in the mucositis and dysphagia FPLS-LR models. The first FPCA components ~~was~~ were not significantly associated with severe mucositis or dysphagia. Compared with the first FPLS components,~~The difference between this component and the first FPLS component for dysphagia was that~~ slightly less weight was given to the higher doses.~~ The first FPCA component was significantly associated with severe mucositis.~~ It should be noted that the sign of the FPCA component loadings is arbitrary so the fact that the odds ratios are less than 1 does not indicate that there is an inverse correlation between RT dose and severe toxicity.

None of the clinical covariates was significantly associated with toxicity in the mucositis models. ~~Of the clinical covariates, only unknown primary disease site was significantly associated (negative association) with toxicity in the FPC-LR or FPLS-LR mucositis models. In the MLR mucositis model carboplatin was positively associated with severe mucositis and 1 cycle of cisplatin followed by 1 cycle of carboplatin was negatively associated with severe toxicity.~~Concurrent ~~C~~cisplatin ~~and carboplatin were~~was significantly associated with severe acute dysphagia in the FPC-LR and FPLS-LR models, but not the PLR model~~, but not in the MLR model~~. None of the clinical covariates was significantly associated with severe toxicity in either of the PLR models.

Discussion

Our results demonstrate that FPC-LR and FPLS-LR produced models with marginally ~~better~~ better ~~predictive performance~~discrimination and overall performance than ~~M~~PLR and superior calibration (aim i). They also show that FPCA and FPLS are appropriate methods ~~for highly correlated DVH data~~ to provide robust estimates of dose-response associations, to inform RT planning, in the presence of highly correlated DVH data (aim ii). We, therefore, encourage the use of FDA methods in future NTCP modeling studies. We suggest that our externally validated dysphagia FPLS-LR model is suitable for clinical decision-support~~ and~~ . To the best of our knowledge, ~~they~~it represents the severe acute ~~mucositis and~~ dysphagia model~~s~~ with the best predictive performance to date. Previous models of severe dysphagia during or shortly following RT that measured discrimination had AUC values of 0.62 (45) and 0.74 (46). These studies did not perform external validation. ~~that our~~The mucositis FP~~LS~~C-LR model should be externally validated to determine its potential to aid clinical decision-making. ~~To the best of our knowledge, they represent the severe acute mucositis and dysphagia models with the best predictive performance to date.~~ Both models are available at https://github.com/XXXX.

The shapes of the first FPLS components indicate that both severe mucositis and dysphagia are most strongly associated with the volume of the oral cavity or pharyngeal mucosa receiving high and intermediate fractional doses (greater than approximately 1.0 Gy). Therefore, RT planning interventions aiming to minimize the incidence of severe acute mucositis and dysphagia should minimize the volumes of oral cavity and pharyngeal mucosa receiving high and intermediate fractional doses, without compromising other aspects of the plan, such as target coverage. Whilst this is intuitively unsurprising, many RT planning protocols, such as RTOG 0912, RTOG 0920 and RTOG 1216, set planning objectives based on OAR mean doses, which give equal importance to low doses and high doses. This suboptimal approach is likely taken due to the common use of mean dose, to achieve dimensionality reduction, in studies aiming to elucidate dose-response relationships. The first FPCA components, which are unsupervised, had similar shapes to the first FPLS components, which are supervised, suggesting that, for this dataset, the variation in severity of toxicities is related to the variations in the DVHs. This suggestion is further supported by the fact that the MSC for FPCA selected the first FPCA component (the one describing the most variation in the DVH data). This will not necessarily be the case for all datasets. The 95% confidence intervals of the first FPLS components are slightly wider than the first FPCA components (figure 1) due to the presence of patients who did not follow the general dose-response trend (i.e. received lower doses, but experienced severe toxicity and vice versa). The substantial penalization of the ~~instability of the~~ P~~M~~LR odds ratios (many often being set to 1) demonstrate~~s~~ the ~~severe~~ limitations of using ~~MLR~~ PLR models to infer associations between correlated dosimetric covariates and toxicity and, hence, we do not recommend its use in this context. Unlike the FDA models, the P~~M~~LR models were unable to identify that high doses, greater than approximately 1.0 Gy per fraction, had higher correlations with toxicity than low doses, as would be intuitively expected.

The FDA models were also able to identify an association between concurrent cisplatin ~~and carboplatin~~ and severe acute dysphagia. The associations between ~~concurrent chemotherapy~~cisplatin and dysphagia in the ~~MLR~~ PLR model ~~were~~ was not significant ~~(although the size of the odds ratios was similar to that in the FDA models)~~. This may be due to the fact that concurrent chemotherapy was correlated ~~to~~ with the DVH metrics ~~(which have highly unstable odds ratios)~~ due to patients with parotid gland primary tumors (who receive unilateral, rather than bilateral, irradiation and, hence, lower pharyngeal mucosa doses) not receiving concurrent chemotherapy. The number of patients receiving concurrent carboplatin or a combination of cisplatin and carboplatin was low (appendix C), leading to large uncertainties in the odds ratios for those covariates. The FDA models featured large uncertainties for the odds ratios of clinical covariates that were highly correlated with other covariates or which applied to small numbers of patients. It should be noted that the regression coefficients of the clinical covariates were not penalized in the FDA models.

There have been very few previous attempts to apply FDA to NTCP modeling (15, 47, 48). These have used either spline basis functions or FPCA (15). To the best of our knowledge, this study represents the first application of FPLS to NTCP modeling. Many previous NTCP modeling studies have not addressed the problem of ~~instability~~ the high uncertainties of the model regression coefficients caused by multicollinearity. Investigators who have recognized this limitation have avoided the multicollinearity problem by reducing the data describing heterogeneous dose distributions to simple summary metrics, such as mean or maximum dose. However, this leads to suboptimal recommendations for RT planning. For example, using mean dose to optimize or assess RT plans gives equal weight to all dose levels, whereas preferentially minimizing the volume of an OAR receiving high doses rather low doses is, in fact, likely to result in a lower toxicity incidence.

A limitation of our approach is that, as the technique is an empirical data-driven method, there are decreases in the regression coefficient with increasing dose, which does not have a biophysical rationale. This should be carefully considered when interpreting dose-response associations from these components. This limitation could be overcome through adopting a Bayesian approach whereby prior knowledge is provided to the model dictating that with increasing dose level the regression coefficient can only remain constant or increase, and not decrease. Mathematically, this would take the form of a monotonically increasing prior function (47). The slight difference in the dysphagia scoring systems between the training and external validation cohorts may have reduced the performances of the models on external validation. However, the models performed at least as well on external validation as internal validation. The relatively small size of the external validation cohort should also be considered as a potential limitation.

In the future, FPCA or FPLS could be applied to the 3D dose distribution (rather than the DVH) (15), either to a single OAR or the entire dose grid, encompassing multiple OARs. This would allow associations between spatial aspects of the dose distribution and toxicity to be explored. This would require accurate mapping of the 3D dose distributions onto a common reference.

Conclusions

FPC-LR and FPLS-LR models of severe acute mucositis ~~and dysphagia~~ had marginally better ~~predictive performance~~discrimination than P~~M~~LR on internal validation. FDA models of dysphagia had marginally improved discrimination and substantially superior calibration than PLR on external validation indicating potential advantages ~~and should be considered~~ for clinical decision-support. FPCA and FPLS enable robust estimates of dose-response associations in the context of correlated dose data. This permits understanding of the most beneficial dose levels to spare in RT planning. Minimizing the volumes of the oral cavity and pharyngeal mucosa receiving high and intermediate doses is expected to reduce the incidence of severe acute mucositis and dysphagia. We recommend that FDA methods be applied to future NTCP modeling studies.

References

1. Lambin P, Roelofs E, Reymen B, *et al.* "Rapid Learning health care in oncology" - An approach towards decision support systems enabling customised radiotherapy. *Radiother. Oncol.* 2013;109:159–164.

2. Langendijk JA, Lambin P, De Ruysscher D, *et al.* Selection of patients for radiotherapy with protons aiming at reduction of side effects: The model-based approach. *Radiother. Oncol.* 2013;107:267–273.

3. van Baardwijk A, Wanders S, Boersma L, *et al.* Mature results of an individualized radiation dose prescription study based on normal tissue constraints in stages I to III non-small-cell lung cancer. *J. Clin. Oncol.* 2010;28:1380–1386.

4. Yorke ED, Kutcher GJ, Jackson A, *et al.* Probability of radiation-induced complications in normal tissues with parallel architecture under conditions of

uniform whole or partial organ irradiation. *Radiother. Oncol.* 1993;26:226–237.

5. Slinker BK, Glantz SA. Multiple regression for physiological data analysis: the problem of multicollinearity. *Am. J. Physiol. - Regul. Integr. Comp. Physiol.* 1985;249:R1–R12.

6. Bentzen SM, Constine LS, Deasy JO, *et al.* Quantitative analyses of normal tissue effects in the clinic (QUANTEC): An introduction to the scientific issues. *Int. J. Radiat. Oncol. Biol. Phys.* 2010;76:3–9.

7. Söhn M, Alber M, Yan D. Principal component analysis-based pattern analysis of dose-volume histograms and influence on rectal toxicity. *Int. J. Radiat. Oncol. Biol. Phys.* 2007;69:230–239.

8. Dawson LA, Biersack M, Lockwood G, *et al.* Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation. *Int. J. Radiat. Oncol. Biol. Phys.* 2005;62:829–837.

9. Skala M, Rosewall T, Dawson L, *et al.* Patient-assessed late toxicity rates and principal component analysis after image-guided radiation therapy for prostate cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 2007;68:690–698.

10. Liang Y, Messer K, Rose BS, *et al.* Impact of bone marrow radiation dose on acute hematologic toxicity in cervical cancer: Principal component analysis on high dimensional data. *Int. J. Radiat. Oncol. Biol. Phys.* 2010;78:912–919.

11. Vesprini D, Sia M, Lockwood G, *et al.* Role of principal component analysis in predicting toxicity in prostate cancer patients treated with hypofractionated intensity-modulated radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* 2011;81:415–421.

12. Bauer JD, Jackson A, Skwarchuk M, *et al.* Principal component, Varimax rotation and cost analysis of volume effects in rectal bleeding in patients treated with 3D-CRT for prostate cancer. *Phys. Med. Biol.* 2006;51:5105–5123.

13. Ramsay JO. When the data are functions. *Psychometrika*. 1982;47:379–396.

14. Levitin DJ, Nuzzo RL, Vines BW, *et al.* Introduction to functional data analysis. *Can. Psychol. Can.* 2007;48:135–155.

15. Benadjaoud MA, Blanchard P, Schwartz B, *et al.* Functional data analysis in NTCP modeling: a new method to explore the radiation dose-volume effects. *Int. J. Radiat. Oncol. Biol. Phys.* 2014;90:654–663.

16. Hall P, Hosseini-Nasab M. On properties of functional principal components analysis. *Jounal R. Stat. Soc. Ser. B (Statistical Methodol.* 2006;68:109–126.

17. Preda C, Saporta G. PLS regression on a stochastic process. *Comput. Stat. Data Anal.* 2005;48:149–158.

18. Reiss PT, Ogden RT. Functional principal component regression and functional partial least squares. *J. Am. Stat. Assoc.* 2007;102:984–996.

19. Worley B, Powers R. Multivariate analysis in metabolomics. *Curr. Metabolomics*. 2013;1:92–107.

20. XXXXX

21. XXXXX

22. XXXXX

23. XXXXX

24. XXXXX

25. The National Cancer Institute. Common Toxicity Criteria (CTC) Version 2.0. 1999.

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

27. Tucker SL, Michalski JM, Bosch WR, *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;104:109–113.

28. Tibshirani R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B*. 1996;58:267–288.

29. Cardot H, Ferraty F, Sarda P. Functional linear model. *Stat. Probab. Lett.* 1999;45:11–22.

30. Kraemer N, Sugiyama M. The degrees of freedom of partial least squares regression. *J. Am. Stat. Assoc.* 2011;106:697–705.

31. Tony Cai T, Hall P. Prediction in functional linear regression. *Ann. Stat.* 2006;34:2159–2179.

32. Mevik B-H, Wehrens R. The pls package: Principal component and partial least squares regression in R. *J. Stat. Softw.* 2007;18.

33. Escabias M, Aguilera AM, Valderrama MJ. Modeling environmental data by functional principal component logistic regression. *Environmetrics*. 2005;16:95–107.

34. Müller HG, Stadtmüller U. Generalized functional linear models. *Ann. Stat.* 2005;33:774–805.

35. Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128–138.

36. Pavlou M, Ambler G, Seaman SR, *et al.* How to develop a more accurate risk prediction model when there are few events. *BMJ*. 2015;351:h3868.

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

38. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* 1996;15:361–387.

39. R Development Core Team R. R: A Language and Environment for Statistical Computing Team RDC, ed. *R Found. Stat. Comput.* 2011;1:409.

40. Febrero-Bande M, Oviedo de la Fuente M. Statistical computing in functional data analysis: the R package fda. usc. *J. Stat. Softw.* 2012;51:1–28.

41. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 2010;33:1–22.

42. Harrell FE. *Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis*. Cham, Switzerland: Springer

International Publishing AG; 2015.

43. Van Calster B, Nieboer D, Vergouwe Y, *et al.* A calibration hierarchy for risk models was defined: From utopia to empirical data. *J. Clin. Epidemiol.* 2016;(in press).

44. Hosmer DW, Lemeshow S. *Applied logistic regression*. New York: Wiley; 2000.

45. XXXXX

46. Sanguineti G, Gunn GB, Parker BC, *et al.* Weekly dose-volume parameters of mucosa and constrictor muscles predict the use of percutaneous endoscopic gastrostomy during exclusive intensity-modulated radiotherapy for oropharyngeal cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 2011;79:52–59.

47. Schipper M, Taylor JMG, Lin X. Bayesian generalized monotonic functional mixed models for the effects of radiation dose histograms on normal tissue complications. *Stat. Med.* 2007;26:4643–4656.

48. Schipper M, Taylor JMG, Lin X. Generalized monotonic functional mixed models with application to modelling normal tissue complications. *J. R. Stat. Soc. Ser. C Appl. Stat.* 2008;57:149–163.

Figure and Table Captions

Figure 1: First functional principal component (left column) and first functional partial least squares component (right column) for mucositis training (top row), dysphagia training (middle row) and dysphagia external validation (bottom row) data bootstrapped with 2000 ~~stratified~~ replicates. Each line represents one bootstrap sample. The functional principal components show the variance in the patient DVHs over the range of dose levels. The functional partial least squares components show the covariance between the patient DVHs and toxicity outcomes over the range of dose levels.

Figure 2: Logistic calibration curve of the FPLS-LR dysphagia model predictions against actual toxicity outcome for the external validation data. The relative frequency distribution of the raw predicted probabilities along with the actual outcome (0 = non-severe dysphagia, 1 = severe dysphagia) are displayed at the

bottom of the figure.

Table 1: Predictive performance of the mucositis and dysphagia models on internal validation (corrected for optimism by bootstrapping with 2000 ~~stratified~~ replicates) and external validation (for the dysphagia models). For the dysphagia models the metrics of predictive performance are given as internal validation/external validation. AUC – area under receiver operating characteristic curve; P~~M~~LR – penalized ~~multivariable~~ logistic regression; FPC-LR – functional principal component-logistic regression; FPLS-LR – functional partial least squares-logistic regression.

Table 2: Odds ratios for penalized ~~multivariable~~ logistic regression models. 95% CI – 95 percentile confidence intervals calculated by bootstrapping the model fitting with 2000 ~~stratified~~ replicates; ~~* - statistically significant at the α = 0.05 level;~~ definitiveRT – definitive radiotherapy; indChemo – induction chemotherapy; cisCarbo – one cycle of cisplatin followed by one cycle of carboplatin; Vx – volume of organ receiving x cGy of radiation per fraction.

Table 3: Odds ratios for functional principal component-logistic regression models. 95% CI – 95 percentile confidence intervals calculated by bootstrapping the model fitting with 2000 ~~stratified~~ replicates; * - statistically significant at the α = 0.05 level; definitiveRT – definitive radiotherapy; indChemo – induction chemotherapy; cisCarbo – one cycle of cisplatin followed by one cycle of

carboplatin; ~~dvh.PC1~~DVH FPCx –~~-~~ ~~first~~ functional principal component x of dose-volume histogram data. The sign of the FPC loadings is arbitrary so the fact that the odds ratios are less than 1 does not indicate that there is an inverse correlation between RT dose and severe toxicity.

Table 4: Odds ratios for functional partial least squares-logistic regression models. 95% CI – 95 percentile confidence intervals calculated by bootstrapping the model fitting with 2000 ~~stratified~~ replicates; * - statistically significant at the α = 0.05 level; definitiveRT – definitive radiotherapy; indChemo – induction chemotherapy; cisCarbo – one cycle of cisplatin followed by one cycle of carboplatin; ~~dvh~~DVH~~.~~ FPLSx – functional partial least squares component x of dose-volume histogram data.

Manuscript

Abstract

Purpose

Current normal tissue complication probability (NTCP) modeling using logistic regression suffers from bias and high uncertainty in the presence of highly correlated radiation therapy (RT) dose data. This hinders robust estimates of dose-response associations and, hence, optimal normal tissue-sparing strategies from being elucidated. Using functional data analysis (FDA) to reduce the dimensionality of the dose data could overcome this limitation.

Methods and Materials

FDA was applied to modeling of severe acute mucositis and dysphagia resulting from head and neck RT. Functional partial least squares regression (FPLS) and functional principal component analysis (FPCA) were used for dimensionality reduction of the dose-volume histogram data. The reduced dose data were input into functional logistic regression models (FPLS-LR and FPC-LR) along with clinical data. This approach was compared with penalized logistic regression (PLR) in terms of predictive performance and the significance of treatment covariate-response associations, assessed using bootstrapping.

Results

The area under the receiver operating characteristic curves (AUC) for the PLR, FPC-LR and FPLS-LR models were 0.65, 0.69 and 0.67 for mucositis (internal validation) and 0.81, 0.83 and 0.83 for dysphagia (external validation), respectively. The calibration slopes/intercepts for the PLR, FPC-LR and FPLS-LR models were 1.6/-0.67, 0.45/0.47 and 0.40/0.49 for mucositis (internal validation) and 2.5/-0.96, 0.79/-0.04 and 0.79/0.00 for dysphagia (external validation). The bootstrapped odds ratios indicated significant associations between RT dose and severe toxicity in the mucositis and dysphagia FDA models. Cisplatin was significantly associated with severe dysphagia in the FDA models. None of the covariates was significantly associated with severe toxicity in the PLR models. Dose levels greater than approximately 1.0 Gy/fraction were most strongly associated with severe acute mucositis and dysphagia in the FDA models.

Conclusions

FPLS and FPCA marginally improved predictive performance compared with PLR and provided robust dose-response associations. FDA is recommended for use in NTCP modeling.

Introduction

Normal tissue complication probability (NTCP) modeling uses radiation therapy (RT) dose data, often in combination with clinical and biological data, to construct statistical models of RT-induced toxicity. There are two distinct aims of NTCP modeling: (i) accurate prediction of toxicity outcomes for individual patients and (ii) estimates of associations between treatment covariates and toxicity. Accurate prediction enables clinical decision-support (1), treatment plan comparison, treatment modality selection (2) and personalized dose prescription (3). Robust estimates of associations between covariates and toxicity can inform the design of RT planning interventions aimed at reducing toxicity.

A major weakness of many NTCP models is suboptimal dimensionality reduction of the RT dose distribution (reducing the number of variables used to describe the dose distribution from all of the points on the 3D dose grid to a small number of summary metrics). In order to input dose data into statistical models the 3D dose distribution delivered to an organ at risk (OAR) is reduced to a single or series of scalar metrics, for example maximum dose, mean dose or multiple points sampled from the dose-volume histogram (DVH), such as the volume of an OAR receiving at least *x* cGy (*Vx*). Ideally, information from each dose level should be explicitly input into the model to prevent loss of potentially important information. However, due to the nature of the dose deposition within the patient, adjacent dose levels are very highly correlated (4) (appendix D). This is problematic for many statistical modeling methods, such as (the commonly used) logistic regression, which often exhibit biased regression coefficients with large standard errors in the presence of collinearity (5). The structure of the correlations is often consistent between patients as the volumes of an OAR receiving adjacent dose levels are highly correlated for all patients. Therefore, if the same or similar treatment techniques are employed, this does not necessarily prevent the models from being able to accurately predict outcomes prospectively for new patients. However, it does result in the regression coefficients of the dosimetric covariates being biased and having large standard errors. The apparent regression coefficients of the dosimetric covariates do not generalize well to new patients and, hence, should not be used to determine the strength of associations between correlated dose metrics and toxicity, as is commonly done (6).

A small number of studies have attempted to address this issue through using principal component analysis (PCA) to reduce the dimensionality of the DVH data (7–12). However, PCA has been shown to perform poorly when the number of predictors (DVH points) is comparable to, or larger than, the number of observations (patients), as is often the case in NTCP modeling (9, 11). Functional data analysis (FDA) is a statistical framework for analyzing continuous curves rather than discrete measurements (13). Treating an entire curve, for example, a DVH curve, as a single entity removes the problem of correlation (14) and explicitly retains the relationship between points on the DVH curve, which standard, non-functional, statistical techniques do not capture. Data are represented as curves through the use of basis functions. There are different types of basis function including *a priori* fixed bases, such as splines or wavelets, and data-driven bases, for example functional principal component analysis (FPCA). Functional logistic regression uses functional data to predict binary outcomes. It is well suited to NTCP modeling due to the continuous nature of DVH curves and the binary nature of toxicity endpoints. Functional logistic regression has recently been applied to NTCP modeling by Benadjaoud *et al.* (15), using FPCA (16) for dimensionality reduction of the DVH data. However, FPCA (and non-functional PCA) is unsupervised (does not use outcome data), which may be a limitation for NTCP modeling. The FPCA components with the most variance in the RT dose data may not be the ones that are most strongly associated with the toxicity outcome of interest. Functional partial least squares regression (FPLS) (17, 18) is a supervised analogue of FPCA. It overcomes this limitation through generating uncorrelated covariates (FPLS components) in the linear space of the predictors, accounting for the correlation between those predictors and outcome, in this case toxicity. As PLS (and FPLS) uses the outcome (toxicity) data in establishing the components it often outperforms PCA (and FPCA) in prediction tasks (19). However, due to the inclusion of outcome data, it is more susceptible to overfitting.

In this study we applied FPLS and FPCA to NTCP modeling of severe acute mucositis and dysphagia. We compared our novel application of FDA with non-functional penalized logistic regression (PLR) models. The aims of this study were to (i) determine whether using FPLS or FPCA to reduce the DVH data would improve predictive performance compared with PLR and (ii) assess whether FPLS or FPCA would lead to more robust estimates of associations between DVH data and toxicity than PLR.

Methods and Materials

Patient data

Data from 351 head and neck RT patients, enrolled in one of the six different clinical trials (20–24)[ISRCTN XXXX], were used to train and internally validate severe acute mucositis and dysphagia NTCP models. Data from the same patients were used for the modeling of both toxicities. This dataset is described in appendix A and [XXXX et al. 2016a (in press), XXXX et al. 2016b (under review)]. Mucositis and dysphagia were both consistently scored for all studies using the Common Terminology Criteria for Adverse Events (CTCAE) versions 2 (25) or 3 (26) instruments. The mucositis and dysphagia grading systems are near equivalent in both versions. Both toxicities were recorded, prospectively, for all patients prior to the start of RT, weekly during RT, and at 1 - 4 and 8 weeks post-RT by head and neck cancer specialists, trained in the use of the scoring systems, using standard trial protocols. The toxicity outcome was defined as the peak grade of toxicity, dichotomized into grade 3 or worse (severe) and less than grade 3 (non-severe). Grade 3 mucositis corresponds to confluent mucositis and grade 3 dysphagia corresponds to feeding tube dependence for more than 24 hours. Patients with baseline toxicity were excluded from the analysis. To reduce bias at the expense of statistical power, patients with any missing toxicity scores and a peak score below 3 were excluded from the analysis. A detailed justification for this approach is provided in appendix B and [XXXX et al. 2016a (in press)]. Of the 351 patients, 183 met the inclusion criteria for mucositis modeling (severe mucositis incidence = 73%) and 179 met the inclusion criteria for dysphagia modeling (severe dysphagia incidence = 66%). Ninety head and neck RT patients treated at XXXX with acute dysphagia data available were used as an external validation cohort for the dysphagia models (severe dysphagia incidence = 48%). In this cohort severe acute dysphagia was defined as the requirement for percutaneous endoscopic gastrostomy (PEG) tube insertion. It should be noted that there was a slight difference in the scoring systems due to the data available. All centers involved in treating patients included in this study employed a reactive approach to PEG-insertion, that is, delaying insertion until deemed clinically necessary.

Induction chemotherapy (yes or no), concurrent chemotherapy regime (cisplatin, carboplatin, one cycle of cisplatin followed by one cycle of carboplatin or none), definitive versus post-operative RT, primary disease site (oropharynx/oral cavity, nasopharynx/nasal cavity, hypopharynx/larynx, parotid gland or unknown primary), age and sex were also included as covariates in the models. Concurrent chemotherapy was administered in two cycles, on day 1 and day 29 of RT. A comparison of the clinical covariate data in the training and external validation data sets is provided in appendix C.

RT dose data

The extended oral cavity [XXXX et al. 2016a (in press)] and pharyngeal mucosa [XXXX et al. 2016b (under review)] were contoured by clinical oncologists and used as OARs in the mucositis and dysphagia models, respectively. The physical dose distribution was converted to the fractional dose distribution (physical dose delivered in each fraction). This has been shown to be appropriate for NTCP modeling of acute toxicity (27) as the toxicities often develop before the total dose is administered. The fractional dose distribution was described by the normalized cumulative dose-volume histogram (DVH). Preliminary work indicated that corrections for different fractionation regimens based on radiobiological models made negligible difference to the results. This is due to the fact that the fractionation regimens employed (appendix A) were similar. An alternative approach would be to use the cumulative dose delivered up to the appearance of the toxicity endpoint. However, subjective choice of the treating clinicians of when to initiate a feeding tube intervention would lead to substantial noise the in the cumulative dose delivered up to the time of intervention.

Penalized logistic regression model

For the non-functional model the fractional DVH curves were discretely sampled from 0.2 Gy to 2.6 Gy at 0.2 Gy intervals. This sampling was chosen to encompass the entire range of OAR doses with enough granularity to capture the shapes of the DVHs. These DVH measurements were input into a PLR model along with the clinical covariates. Penalization was performed using the least absolute shrinkage and selection operator (LASSO) (28). LASSO reduces the magnitude of the regression coefficients, setting some to 0, in order to prevent overfitting. In the context of correlated variables, it reduces the impact of multicollinearity. The penalization strength was selected by 10-fold cross-validation with the value producing the highest average (over all of the cross-validation folds) area under the receiver operating characteristic curve (AUC) on cross-validation selected.

Functional data analysis

The fractional DVH curves (sampled from 0 Gy to 2.60 Gy in 0.01 Gy intervals) were represented using penalized FPCA (16, 29) and penalized FPLS (17, 30) basis functions. FPCA is a dimensionality reduction technique that represents the functional DVH data as orthonormal vector components explaining the maximum variance between patients in the DVH curves. The orthonormality constraint removes the collinearity in the dose metrics used for subsequent modeling and, hence, overcomes the limitations associated with modeling collinear data. The functional principal components $\{\xi_k(d)\}_{k=1}^{\infty}$ represent the functional DVH data (normalized volume as a function of dose for patient *i*), $V_i(d)$ as the sum of the eigenfunctions, $\xi_k(d)$ weighted by their coefficients, c_{ik}:

$$V_i(d) - \mathrm{\mu}(d) = \sum_{k=1}^{\infty} c_{ik}\xi_k(d) \qquad (1)$$

where $\mathrm{\mu}(d)$ is the mean $V(d)$ and c_{ik} describes the score for functional principal component *k* for the DVH of patient *i* and is given by

$$c_{ik} = \int (V_i(d) - \mathrm{\mu}(d))\xi_k(d)dd \qquad (2)$$

The eigenfunctions, $\{\xi_k\}_{k=1}^{\infty}$ and their corresponding eigenvalues (describing the amount of variance explained by each eigenfunction), $\lambda_1 \geq \lambda_2 \geq \cdots$, are determined by eigendecomposition (factorization into eigenvalues and eigenvectors) of the covariance operator, Σ where

$$\Sigma(d_1, d_2) = \mathrm{Cov}[V(d_1), V(d_2)] = \ \mathrm{E}\left[\big(V(d_1) - \mu(d_1)\big)\big(V(d_2) - \mu(d_2)\big)^{\mathrm{T}}\right] \quad (3)$$

where $\mu(d_1)$ is the mean volume receiving dose d_1. *V(d)* can be approximated by a small number of principal components, k_n, assuming that $c_{ik} = 0$ for $k > k_n$, to achieve dimensionality reduction to a small number of basis functions efficiently describing the variation between patients in the DVH data:

$$V_i(d) \approx \mu(d) + \sum_{k=1}^{k_n} c_{ik}\xi_k(d) \quad (4)$$

The eigenfunctions and their coefficients can then be used in subsequent analyses. The FPCA components can be used to estimate a toxicity outcome for patient *i*, y_i using a functional linear model (29, 31):

$$y_i = \alpha + \int \beta(d)V_i(d)dd + \varepsilon_i \quad (5)$$

where α is the intercept and ε_i is a centered random error.

When FPCA is used to describe the DVH data *β(d)* represents a "weighting function" describing the amount of variation between patients at all dose levels on the DVH. It can be approximated by k_n eigenfunctions:

$$\beta(d) = \sum_{k=1}^{\infty} \beta_k\xi_k(d) \approx \sum_{k=1}^{k_n} \beta_k\xi_k(d) \quad (6)$$

An estimate of the response, $\hat{y}_i$ can be made using (derivation in (29)):

$$\hat{y}_i = \alpha + \int \beta(d)V_i(d)dd \approx \alpha + \sum_{k=1}^{k_n} \hat{\beta}_k c_{ik} \quad (7)$$

where

$$\hat{\beta}_{(1:k_n)} = \left(\frac{c_{.1}^{\mathrm{T}}y}{n\lambda_1}, \dots, \frac{c_{.k_n}^{\mathrm{T}}y}{n\lambda_{k_n}}\right) \quad (8)$$

The model was fit to the data, placing penalization on the curvature (second derivative) of the eigenfunctions, by

$$\widehat{y_i} = \xi_k \left(\xi_k^{\mathrm{T}} \xi_k + r \xi_k^{\mathrm{T}} \mathbf{P} \xi_k \right)^{-1} \xi_k^{\mathrm{T}} y_i \quad (9)$$

where r is the amount of penalization, $\mathbf{P}$ is the vector (0, 0, 1) that defines the penalty matrix such that the second derivative (curvature) is penalized and y is the actual response data. The choice of which components to include (within the first 5 components) and the magnitude of the roughness penalty, r to apply (selected from a set of values in the range from 0 to 1350), to best estimate the toxicity outcomes, were determined using model selection criteria (MSC) (16) with the Bayesian information criterion:

$$MSC(k_n) = \log\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2\right] + \frac{\log(n)k_n}{n^2} \quad (10)$$

where n is the number of patients and y_i is the actual outcome (toxicity) data. This penalizes the model complexity to reduce overfitting. Models with different values of r and $k_n$were generated and the combination of values that minimized MSC was selected. The FPCA or FPLS components included affect the smoothness of the estimate of the $\beta(d)$ function as the dominant mode of variation tends to be smooth and roughness tends to increase for subsequent modes of variation, in part due to the orthogonality constraint.

FPLS is similar to FPCA, but uses the response (toxicity) data in constructing the FPLS components (17), $\{\tilde{\xi}_k\}_{k=1}^{\infty}$ in order to establish orthogonal components that have maximum covariance to the response. This is achieved through maximizing the squared covariance between $V_i(d)$ and the response, y_i with the constraint that all components are mutually orthogonal (30). This takes the place of the eigendecomposition used for FPCA, described in equation 3. The iterative algorithm used to compute the FPLS components is described in (32). When FPLS is used for dimensionality reduction of the DVH data, *β(d)* can be interpreted as a data-driven "weighting function" for the importance of each dose level in causing severe toxicity. It is important to consider that, as this is a data-driven approach, the "weighting function" is an estimate of the "true weighting function" over the range of available data and is influenced by the structure (i.e. distribution in dose-volume space) of the available data. MSC was performed for the FPLS analysis in the same manner as for FPCA. The FPCA and FPLS analysis was bootstrapped with 2000 replicates to assess the uncertainty of the shapes of the components.

The optimal FPCA and FPLS components (those producing the lowest MSC) were used as basis functions as input into functional logistic regression (33, 34) models (FPC-LR and FPLS-LR) along with the (non-functional) clinical covariates. The functional logistic regression model describes the probability of patient i, experiencing severe toxicity $P(y_i = 1)$, and is given by

$$\ln\frac{P(y_i = 1)}{P(y_i = 0)} = \alpha + \sum_{j=1}^{p} \beta_j Z^j + \int \beta(d) V_i(d)\, dd \approx \alpha + \sum_{j=1}^{p} \beta_j Z^j + \sum_{k=1}^{k_n} \beta_k c_{ik} \quad (11)$$

using the substitution for the functional linear model described in equation 7, where α is the intercept and $\{Z^j\}_{j=1}^{p}$ are the non-functional covariates with regression coefficients $\{\beta_j\}_{j=1}^{p}$. Maximum likelihood estimation of the regression coefficients was performed using iteratively weighted least squares.

Model comparisons

The predictive performance and generalizability of the models (addressing aim i) were assessed in terms of discrimination, calibration and overall performance on internal validation, and additionally for the dysphagia models on external validation. The discriminative abilities of the models were assessed using the AUC. Calibration was evaluated by the slope and intercept of a logistic regression model of the actual toxicity against the predicted probability of severe toxicity (35, 36). Overall model performance was measured using the Brier score (37), BS. It is defined as

$$BS = \frac{1}{N}\sum_{t=1}^{N}(p_t - y_t)^2 \quad (12)$$

where p_t is the predicted probability, y_t is the actual outcome and N is the number of predictions. The score takes a value between 0 and 1 with lower values indicating better performance.

For the internal validation the performance metrics were "corrected for optimism" using bootstrapping with 2000 replicates (38). The optimism-corrected performance metrics, $M_{corrected}$ were calculated by

$$M_{corrected} = M_{apparent} - O \quad (13)$$

where $M_{apparent}$ is the performance metric, for example AUC, calculated using all of the training data to, both, fit the model and evaluate its performance, and the optimism, O is given by

$$O = \frac{1}{B}\sum_{b=1}^{B}(M_{b,boot} - M_{b,orig}) \quad (14)$$

where B is the number of bootstrap replicates, $M_{b,boot}$ is the performance metric calculated using the bootstrap dataset b to, both, fit and evaluate model performance, and $M_{b,orig}$ is the performance of the model fit using the bootstrap dataset b evaluated on the original dataset. This provides an unbiased estimate of internal validity, penalizing for overfitting. Model hyper-parameter tuning, such as the selection of the amount of penalization for the PLR models and the selection of components and penalization for the FDA models, was performed for each bootstrap replicate. This prevents any "data leakage" from the training data into the internal validation data. The dysphagia models were used to predict severe dysphagia probability for the external validation cohort. Those predictions were compared to the actual PEG-dependence data for the cohort and the same performance metrics calculated. The uncertainties of the odds ratios (addressing aim ii) were assessed by calculating bootstrapped 95 percentile confidence intervals with 2000 replicates. Statistical analysis was performed using the statistical computing R language version 3.2.4 (39) and the fda.usc version 1.2.2 (40), glmnet version 2.0 (41), rms version 4.5 (42) and val.prob.ci.2 (43) packages.

Results

For FPCA, the variances in the DVH data explained by the first five FPCA components were 80.8%, 12.5% and 3.7%, 1.2% and 0.6% for mucositis and 70.8%, 14.5%, 5.6%, 4.4% and 1.6% for dysphagia. For FPLS, the variances explained by the first five FPLS components were 78.1%, 16.9%, 2.0%, 2.5% and 0.6% for mucositis and 76.2%, 8.6%, 11.2%, 2.7% and 1.3% for dysphagia. The model selection resulted in the first two components being selected for the FPCA and FPLS mucositis models and only the first component selected in both of the dysphagia FDA models. Penalization of 1342 was chosen by the model selection for the mucositis FPCA model, 0 for the mucositis FPLS model and 1350 for both of the dysphagia FDA models.

Figure 1 shows the first FPCA and FPLS components for the mucositis and dysphagia models, respectively. Bootstrapping the FPCA and FPLS indicated that the shapes of the first FPCA and FPLS components were very similar irrespective of the random sample of patients selected. There was a general trend that the FPCA and FPLS loadings increased with increasing dose and sharply decreased to 0 at the maximum dose. The FPCA components indicate that the variation, between patients, in the volume of OAR irradiated to a certain dose level increased with increasing dose level. The same trend in the FPLS components indicates that higher doses were more strongly associated with severe toxicity. The decrease in the first FPCA and FPLS component loadings at around 1.8 Gy (figure 1) for the dysphagia training data is indicative of reduced variation in this region of the DVHs between patients. This is likely to be due to the fact that most of the variation in the pharyngeal mucosa dose distribution between patients is related to the variation in volume of overlap of the two different planning target volumes (whose prescription dose levels correspond to the positions of the two peaks in the FPCA and FPLS components) with the pharyngeal mucosa.

For the PLR, FPC-LR and FPLS-LR modeling, oropharynx/oral cavity and no concurrent chemotherapy were removed as covariates to prevent perfect collinearity (correlation matrices are shown in appendix D). Odds ratios for other primary disease sites are thus relative to oropharynx/oral cavity and odds ratios for concurrent chemotherapy are relative to no concurrent chemotherapy.

In regards to aim i, the predictive performance of the three different mucositis and dysphagia models, as assessed by internal and external (for dysphagia) validation, is displayed in table 1. The mucositis models had modest (PLR and FPLS-LR) or modest-to-good (FPC-LR) discriminative ability (using the interpretation in (44)) on internal validation. The discriminative abilities and overall performances of the FPC-LR and FPLS-LR models were marginally better than the PLR model. Calibration was relatively poor for all of the models, with the FPC-LR and FPLS-LR models overfitting the data (calibration slope less than 1) and the PLR model underfitting the data (calibration slope greater than 1). The underfitting exhibited by the PLR models was likely due to over shrinkage of the regression coefficients by the LASSO penalization caused by high multicollinearity. It should be noted that the “correction for optimism” may have improved the calibration of the PLR models, as they underfit the data.

The discrimination and calibration of the dysphagia models were better than the mucositis models. All three dysphagia models had good discriminative ability on internal validation. The discriminative abilities of all three models increased on external validation, with the PLR model demonstrating good-to-excellent discrimination and the FPC-LR and FPLS-LR models showing excellent discrimination. The overall performance of all of the models was similar, both on internal and external validation. Calibration of all of the models on internal validation was modest, with the PLR model underfitting the data and the FDA models overfitting the data. The FPC-LR and FPLS-LR models had substantially better calibration than the PLR model on external calibration. The FPLS-LR model had marginally better calibration than the FPC-LR model on external calibration. A logistic calibration curve for the external validation of this model is shown in figure 2. The curve lies close to the identity line indicating good model calibration on external validation.

Concerning aim ii, the results of the bootstrapped penalized and functional logistic regression odds ratios are shown in tables 2 – 4. The odds ratios for the covariates in the PLR models were often set to 1 by the LASSO penalization. In the mucositis and dysphagia PLR models none of the covariates was significantly associated with severe toxicity. Conversely, there was a significant association between the first FPLS component and severe toxicity in the mucositis and dysphagia FPLS-LR models. The first FPCA components were not significantly associated with severe mucositis or dysphagia. Compared with the first FPLS components, slightly less weight was given to the higher doses. It should be noted that the sign of the FPCA component loadings is arbitrary so the fact that the odds ratios are less than 1 does not indicate that there is an inverse correlation between RT dose and severe toxicity.

None of the clinical covariates was significantly associated with toxicity in the mucositis models. Concurrent cisplatin was significantly associated with severe acute dysphagia in the FPC-LR and FPLS-LR models, but not the PLR model. None of the clinical covariates was significantly associated with severe toxicity in either of the PLR models.

Discussion

Our results demonstrate that FPC-LR and FPLS-LR produced models with marginally better discrimination and overall performance than PLR and superior calibration (aim i). They also show that FPCA and FPLS are appropriate methods to provide robust estimates of dose-response associations, to inform RT planning, in the presence of highly correlated DVH data (aim ii). We, therefore, encourage the use of FDA methods in future NTCP modeling studies. We suggest that our externally validated dysphagia FPLS-LR model is suitable for clinical decision-support. To the best of our knowledge, it represents the severe acute dysphagia model with the best predictive performance to date. Previous models of severe dysphagia during or shortly following RT that measured discrimination had AUC values of 0.62 (45) and 0.74 (46). These studies did not perform external validation. The mucositis FPC-LR model should be externally validated to determine its potential to aid clinical decision-making. Both models are available at https://github.com/XXXX.

The shapes of the first FPLS components indicate that both severe mucositis and dysphagia are most strongly associated with the volume of the oral cavity or pharyngeal mucosa receiving high and intermediate fractional doses (greater than approximately 1.0 Gy). Therefore, RT planning interventions aiming to minimize the incidence of severe acute mucositis and dysphagia should minimize the volumes of oral cavity and pharyngeal mucosa receiving high and intermediate fractional doses, without compromising other aspects of the plan, such as target coverage. Whilst this is intuitively unsurprising, many RT planning protocols, such as RTOG 0912, RTOG 0920 and RTOG 1216, set planning objectives based on OAR mean doses, which give equal importance to low doses and high doses. This suboptimal approach is likely taken due to the common use of mean dose, to achieve dimensionality reduction, in studies aiming to elucidate dose-response relationships. The first FPCA components, which are unsupervised, had similar shapes to the first FPLS components, which are supervised, suggesting that, for this dataset, the variation in severity of toxicities is related to the variations in the DVHs. This suggestion is further supported by the fact that the MSC for FPCA selected the first FPCA component (the one describing the most variation in the DVH data). This will not necessarily be the case for all datasets. The 95% confidence intervals of the first FPLS components are slightly wider than the first FPCA components (figure 1) due to the presence of patients who did not follow the general dose-response trend (i.e. received lower doses, but experienced severe toxicity and vice versa). The substantial penalization of the PLR odds ratios (many often being set to 1) demonstrate the limitations of using PLR models to infer associations between correlated dosimetric covariates and toxicity and, hence, we do not recommend its use in this context. Unlike the FDA models, the PLR models were unable to identify that high doses, greater than approximately 1.0 Gy per fraction, had higher correlations with toxicity than low doses, as would be intuitively expected.

The FDA models were also able to identify an association between concurrent cisplatin and severe acute dysphagia. The associations between cisplatin and dysphagia in the PLR model was not significant. This may be due to the fact that concurrent chemotherapy was correlated with the DVH metrics due to patients with parotid gland primary tumors (who receive unilateral, rather than bilateral, irradiation and, hence, lower pharyngeal mucosa doses) not receiving concurrent chemotherapy. The number of patients receiving concurrent carboplatin or a combination of cisplatin and carboplatin was low (appendix C), leading to large uncertainties in the odds ratios for those covariates. The FDA models featured large uncertainties for the odds ratios of clinical covariates that were highly correlated with other covariates or which applied to small numbers of patients. It should be noted that the regression coefficients of the clinical covariates were not penalized in the FDA models.

There have been very few previous attempts to apply FDA to NTCP modeling (15, 47, 48). These have used either spline basis functions or FPCA (15). To the best of our knowledge, this study represents the first application of FPLS to NTCP modeling. Many previous NTCP modeling studies have not addressed the problem of the high uncertainties of the model regression coefficients caused by multicollinearity. Investigators who have recognized this limitation have avoided the multicollinearity problem by reducing the data describing heterogeneous dose distributions to simple summary metrics, such as mean or maximum dose. However, this leads to suboptimal recommendations for RT planning. For example, using mean dose to optimize or assess RT plans gives equal weight to all dose levels, whereas preferentially minimizing the volume of an OAR receiving high doses rather low doses is, in fact, likely to result in a lower toxicity incidence.

A limitation of our approach is that, as the technique is an empirical data-driven method, there are decreases in the regression coefficient with increasing dose, which does not have a biophysical rationale. This should be carefully considered when interpreting dose-response associations from these components. This limitation could be overcome through adopting a Bayesian approach whereby prior knowledge is provided to the model dictating that with increasing dose level the regression coefficient can only remain constant or increase, and not decrease. Mathematically, this would take the form of a monotonically increasing prior function (47). The slight difference in the dysphagia scoring systems between the training and external validation cohorts may have reduced the performances of the models on external validation. However, the models performed at least as well on external validation as internal validation. The relatively small size of the external validation cohort should also be considered as a potential limitation.

In the future, FPCA or FPLS could be applied to the 3D dose distribution (rather than the DVH) (15), either to a single OAR or the entire dose grid, encompassing multiple OARs. This would allow associations between spatial aspects of the dose distribution and toxicity to be explored. This would require accurate mapping of the 3D dose distributions onto a common reference.

Conclusions

FPC-LR and FPLS-LR models of severe acute mucositis had marginally better discrimination than PLR on internal validation. FDA models of dysphagia had marginally improved discrimination and substantially superior calibration than PLR on external validation indicating potential advantages for clinical decision-support. FPCA and FPLS enable robust estimates of dose-response associations in the context of correlated dose data. This permits understanding of the most beneficial dose levels to spare in RT planning. Minimizing the volumes of the oral cavity and pharyngeal mucosa receiving high and intermediate doses is expected to reduce the incidence of severe acute mucositis and dysphagia. We recommend that FDA methods be applied to future NTCP modeling studies.

References

1. Lambin P, Roelofs E, Reymen B, *et al.* "Rapid Learning health care in oncology" - An approach towards decision support systems enabling customised radiotherapy. *Radiother. Oncol.* 2013;109:159–164.

2. Langendijk JA, Lambin P, De Ruysscher D, *et al.* Selection of patients for radiotherapy with protons aiming at reduction of side effects: The model-based approach. *Radiother. Oncol.* 2013;107:267–273.

3. van Baardwijk A, Wanders S, Boersma L, *et al.* Mature results of an individualized radiation dose prescription study based on normal tissue constraints in stages I to III non-small-cell lung cancer. *J. Clin. Oncol.*

2010;28:1380–1386.

4. Yorke ED, Kutcher GJ, Jackson A, *et al.* Probability of radiation-induced complications in normal tissues with parallel architecture under conditions of uniform whole or partial organ irradiation. *Radiother. Oncol.* 1993;26:226–237.

5. Slinker BK, Glantz SA. Multiple regression for physiological data analysis: the problem of multicollinearity. *Am. J. Physiol. - Regul. Integr. Comp. Physiol.* 1985;249:R1–R12.

6. Bentzen SM, Constine LS, Deasy JO, *et al.* Quantitative analyses of normal tissue effects in the clinic (QUANTEC): An introduction to the scientific issues. *Int. J. Radiat. Oncol. Biol. Phys.* 2010;76:3–9.

7. Söhn M, Alber M, Yan D. Principal component analysis-based pattern analysis of dose-volume histograms and influence on rectal toxicity. *Int. J. Radiat. Oncol. Biol. Phys.* 2007;69:230–239.

8. Dawson LA, Biersack M, Lockwood G, *et al.* Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation. *Int. J. Radiat. Oncol. Biol. Phys.* 2005;62:829–837.

9. Skala M, Rosewall T, Dawson L, *et al.* Patient-assessed late toxicity rates and principal component analysis after image-guided radiation therapy for prostate cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 2007;68:690–698.

10. Liang Y, Messer K, Rose BS, *et al.* Impact of bone marrow radiation dose on acute hematologic toxicity in cervical cancer: Principal component analysis on high dimensional data. *Int. J. Radiat. Oncol. Biol. Phys.* 2010;78:912–919.

11. Vesprini D, Sia M, Lockwood G, *et al.* Role of principal component analysis in predicting toxicity in prostate cancer patients treated with hypofractionated intensity-modulated radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* 2011;81:415–421.

12. Bauer JD, Jackson A, Skwarchuk M, *et al.* Principal component, Varimax rotation and cost analysis of volume effects in rectal bleeding in patients treated with 3D-CRT for prostate cancer. *Phys. Med. Biol.* 2006;51:5105–5123.

13. Ramsay JO. When the data are functions. *Psychometrika*. 1982;47:379–396.

14. Levitin DJ, Nuzzo RL, Vines BW, *et al.* Introduction to functional data analysis. *Can. Psychol. Can.* 2007;48:135–155.

15. Benadjaoud MA, Blanchard P, Schwartz B, *et al.* Functional data analysis in NTCP modeling: a new method to explore the radiation dose-volume effects. *Int. J. Radiat. Oncol. Biol. Phys.* 2014;90:654–663.

16. Hall P, Hosseini-Nasab M. On properties of functional principal components analysis. *Jounal R. Stat. Soc. Ser. B (Statistical Methodol.* 2006;68:109–126.

17. Preda C, Saporta G. PLS regression on a stochastic process. *Comput. Stat. Data Anal.* 2005;48:149–158.

18. Reiss PT, Ogden RT. Functional principal component regression and functional partial least squares. *J. Am. Stat. Assoc.* 2007;102:984–996.

19. Worley B, Powers R. Multivariate analysis in metabolomics. *Curr. Metabolomics*. 2013;1:92–107.

20. XXXXX

21. XXXXX

22. XXXXX

23. XXXXX

24. XXXXX

25. The National Cancer Institute. Common Toxicity Criteria (CTC) Version 2.0. 1999.

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

27. Tucker SL, Michalski JM, Bosch WR, *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;104:109–113.

28. Tibshirani R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B*. 1996;58:267–288.

29. Cardot H, Ferraty F, Sarda P. Functional linear model. *Stat. Probab. Lett.* 1999;45:11–22.

30. Kraemer N, Sugiyama M. The degrees of freedom of partial least squares regression. *J. Am. Stat. Assoc.* 2011;106:697–705.

31. Tony Cai T, Hall P. Prediction in functional linear regression. *Ann. Stat.* 2006;34:2159–2179.

32. Mevik B-H, Wehrens R. The pls package: Principal component and partial least squares regression in R. *J. Stat. Softw.* 2007;18.

33. Escabias M, Aguilera AM, Valderrama MJ. Modeling environmental data by functional principal component logistic regression. *Environmetrics*. 2005;16:95–107.

34. Müller HG, Stadtmüller U. Generalized functional linear models. *Ann. Stat.* 2005;33:774–805.

35. Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128–138.

36. Pavlou M, Ambler G, Seaman SR, *et al.* How to develop a more accurate risk prediction model when there are few events. *BMJ*. 2015;351:h3868.

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

38. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* 1996;15:361–387.

39. R Development Core Team R. R: A Language and Environment for Statistical Computing Team RDC, ed. *R Found. Stat. Comput.* 2011;1:409.

40. Febrero-Bande M, Oviedo de la Fuente M. Statistical computing in functional data analysis: the R package fda. usc. *J. Stat. Softw.* 2012;51:1–28.

41. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear

models via coordinate descent. *J. Stat. Softw.* 2010;33:1–22.

42. Harrell FE. *Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis*. Cham, Switzerland: Springer International Publishing AG; 2015.

43. Van Calster B, Nieboer D, Vergouwe Y, *et al.* A calibration hierarchy for risk models was defined: From utopia to empirical data. *J. Clin. Epidemiol.* 2016;(in press).

44. Hosmer DW, Lemeshow S. *Applied logistic regression*. New York: Wiley; 2000.

45. XXXXX

46. Sanguineti G, Gunn GB, Parker BC, *et al.* Weekly dose-volume parameters of mucosa and constrictor muscles predict the use of percutaneous endoscopic gastrostomy during exclusive intensity-modulated radiotherapy for oropharyngeal cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 2011;79:52–59.

47. Schipper M, Taylor JMG, Lin X. Bayesian generalized monotonic functional mixed models for the effects of radiation dose histograms on normal tissue complications. *Stat. Med.* 2007;26:4643–4656.

48. Schipper M, Taylor JMG, Lin X. Generalized monotonic functional mixed models with application to modelling normal tissue complications. *J. R. Stat. Soc. Ser. C Appl. Stat.* 2008;57:149–163.

Figure and Table Captions

Figure 1: First functional principal component (left column) and first functional partial least squares component (right column) for mucositis training (top row), dysphagia training (middle row) and dysphagia external validation (bottom row) data bootstrapped with 2000 replicates. Each line represents one bootstrap sample. The functional principal components show the variance in the patient DVHs over the range of dose levels. The functional partial least squares components show the covariance between the patient DVHs and toxicity outcomes over the range of dose levels.

Figure 2: Logistic calibration curve of the FPLS-LR dysphagia model predictions against actual toxicity outcome for the external validation data. The relative

frequency distribution of the raw predicted probabilities along with the actual outcome (0 = non-severe dysphagia, 1 = severe dysphagia) are displayed at the bottom of the figure.

Table 1: Predictive performance of the mucositis and dysphagia models on internal validation (corrected for optimism by bootstrapping with 2000 replicates) and external validation (for the dysphagia models). For the dysphagia models the metrics of predictive performance are given as internal validation/external validation. AUC – area under receiver operating characteristic curve; PLR – penalized logistic regression; FPC-LR – functional principal component-logistic regression; FPLS-LR – functional partial least squares-logistic regression.

Table 2: Odds ratios for penalized logistic regression models. 95% CI – 95 percentile confidence intervals calculated by bootstrapping the model fitting with 2000 replicates; definitiveRT – definitive radiotherapy; indChemo – induction chemotherapy; cisCarbo – one cycle of cisplatin followed by one cycle of carboplatin; Vx – volume of organ receiving x cGy of radiation per fraction.

Table 3: Odds ratios for functional principal component-logistic regression models. 95% CI – 95 percentile confidence intervals calculated by bootstrapping the model fitting with 2000 replicates; * - statistically significant at the α = 0.05 level; definitiveRT – definitive radiotherapy; indChemo – induction

chemotherapy; cisCarbo – one cycle of cisplatin followed by one cycle of carboplatin; DVH FPCx – functional principal component x of dose-volume histogram data. The sign of the FPC loadings is arbitrary so the fact that the odds ratios are less than 1 does not indicate that there is an inverse correlation between RT dose and severe toxicity.

Table 4: Odds ratios for functional partial least squares-logistic regression models. 95% CI – 95 percentile confidence intervals calculated by bootstrapping the model fitting with 2000 replicates; * - statistically significant at the α = 0.05 level; definitiveRT – definitive radiotherapy; indChemo – induction chemotherapy; cisCarbo – one cycle of cisplatin followed by one cycle of carboplatin; DVH FPLSx – functional partial least squares component x of dose-volume histogram data.

Table 1: Predictive performance of the mucositis and dysphagia models on internal validation (corrected for optimism by bootstrapping with 2000 replicates) and external validation (for the dysphagia models).

Model		AUC	Brier score	Calibration slope	Calibration intercept
Mucositis	PLR	0.65	0.21	1.6	-0.67
	FPC-LR	0.69	0.19	0.45	0.47
	FPLS-LR	0.67	0.20	0.40	0.49
Dysphagia	PLR	0.74/0.81	0.20/0.18	1.2/2.5	-0.15/-0.96
	FPC-LR	0.76/0.83	0.19/0.18	0.59/0.79	0.21/-0.04
	FPLS-LR	0.75/0.83	0.20/0.18	0.56/0.79	0.22/0.00

For the dysphagia models the metrics of predictive performance are given as internal validation/external validation. AUC – area under receiver operating characteristic curve; PLR – penalized multivariable logistic regression; FPC-LR – functional principal component-logistic regression; FPLS-LR – functional partial least squares-logistic regression.

Covariate	Mucositis model		Dysphagia model	
	Odds ratio	95% CI	Odds ratio	95% CI
intercept	2.512	0.016 – 12.43	0.360	0.007 – 2.583
male	1.000	1.000 – 2.554	1.000	1.000 – 1.945
age	1.000	0.971 – 1.006	1.000	0.980 – 1.000
definitiveRT	1.000	0.110 – 1.000	1.000	0.544 – 1.000
indChemo	1.000	0.410 – 1.166	1.000	1.000 – 2.089
cisplatin	1.000	1.000 – 3.464	1.277	1.000 – 3.230
carboplatin	1.000	0.361 – 4.015	1.000	1.000 – 4.278
cisCarbo	1.000	0.136 – 1.769	1.000	0.989 – 2.930
hypopharynxLarynx	1.000	1.000 – 14.71	1.000	1.000 – 2.203
nasopharynxNasalCavity	1.000	0.905 – 6.190	1.000	0.247 – 1.000
unknownPrimary	1.000	0.022 – 1.000	1.000	0.945 – 1.210
parotid	0.814	0.231– 2.546	0.600	0.208 – 1.000
V020	1.000	1.000 – 1.119	1.000	1.000 – 1.031
V040	1.000	0.891 – 1.000	1.000	1.000 – 1.014
V060	1.000	1.000 – 1.032	1.000	1.000 – 1.003
V080	1.000	1.000 – 1.050	1.000	1.000 – 1.023
V100	1.000	0.934 – 1.000	1.000	1.000 – 1.029
V120	1.000	1.000 – 1.084	1.019	1.000 – 1.044
V140	1.000	0.917 – 1.000	1.000	1.000 – 1.019
V160	1.000	1.000 – 1.038	1.000	1.000 – 1.011
V180	1.002	1.000 – 1.085	1.000	0.997 – 1.009
V200	1.000	0.949 – 1.007	1.000	1.000 – 1.019
V220	1.000	1.000 – 1.098	1.008	1.000 – 1.031
V240	1.000	0.616 – 1.154	1.000	1.000 – 1.025
V260	1.000	1.000 – 1.000	1.000	1.000 – 1.000

Table 2: Odds ratios for the penalized multivariable logistic regression models.

95% CI – 95 percentile confidence intervals calculated by bootstrapping the model fitting with 2000 replicates; definitiveRT – definitive radiotherapy; indChemo – induction chemotherapy; cisCarbo – one cycle of cisplatin followed by one cycle of carboplatin; Vx – volume of organ receiving x cGy of radiation per fraction.

Table 3: Odds ratios for the functional principal component-logistic regression models.

Covariate	Mucositis model		Dysphagia model	
	Odds ratio	95% CI	Odds ratio	95% CI
intercept	12.89	1.035 – 1.734 x 10^{9}*	1.616	0.142 – 77.46
male	1.535	0.637 – 4.088	1.675	0.533 – 4.880
age	0.991	0.951 – 1.029	0.988	0.943 – 1.027
definitiveRT	0.254	2.679 x 10^{-9} – 1.773	0.997	0.080 – 7.541
indChemo	0.487	0.070 – 1.960	1.100	0.210 – 7.670
cisplatin	2.251	0.745 – 9.540	4.255	1.077 – 19.86*
carboplatin	1.320	0.142 – 7.314 x 10^{7}	4.429	0.685 – 8.332 x 10^{7}
cisCarbo	0.311	7.815 x 10^{-9}– 2.531 x 10^{7}	2.238	0.319 – 4.587 x 10^{7}
hypopharynxLarynx	4.371	0.512 – 143.9	1.723	0.193 – 1.881 x 10^{7}
nasopharynxNasalCavity	2.370	0.308 – 1.096 x 10^{8}	0.263	0.026 – 1.223
unknownPrimary	0.136	3.042 x 10^{-9} – 3.707	0.859	0.077 – 3.876 x 10^{6}
parotid	1.387	0.103 – 40.37	1.135	0.068 – 18.72
DVH FPC1	0.997	0.993 – 1.007	0.996	0.990 – 1.008
DVH FPC2	1.003	0.992 – 1.009	-	0.992 – 1.003
DVH FPC3	-	0.996 – 1.003	-	0.995 – 1.000
DVH FPC4	-	0.987 – 1.010	-	0.991 – 1.006
DVH FPC5	-	0.971 – 1.033	-	0.991 – 1.006

95% CI – 95 percentile confidence intervals calculated by bootstrapping the model fitting with 2000 replicates; * - statistically significant at the $\alpha = 0.05$ level; definitiveRT – definitive radiotherapy; indChemo – induction chemotherapy; cisCarbo – one cycle of cisplatin followed by one cycle of carboplatin; DVH FPCx – functional principal component x of the dose-volume histogram data. The sign of the FPC loadings is arbitrary so the fact that the odds ratios are less than 1 does not indicate that there is an inverse correlation between RT dose and severe toxicity.

Table 4: Odds ratios for the functional partial least squares-logistic regression models.

Covariate	Mucositis model		Dysphagia model	
	Odds ratio	95% CI	Odds ratio	95% CI
intercept	12.90	0.961 – 2.424 x 10^{10}	1.634	0.128 – 104.4
male	1.539	0.620 – 4.757	1.661	0.472 – 4.719
age	0.991	0.947 – 1.033	0.988	0.942 – 1.029
definitiveRT	0.260	7.707 x 10^{-11} – 1.245	0.975	0.046 – 7.831
indChemo	0.484	0.064 – 2.442	1.100	0.222 – 7.866
cisplatin	2.246	0.728 – 11.33	4.235	1.083 – 20.88*
carboplatin	1.315	0.110 – 1.051 x 10^{8}	4.393	0.580 – 8.424 x 10^{7}
cisCarbo	0.313	8.668 x 10^{-9} – 3.303 x 10^{7}	2.245	0.324 – 4.247 x 10^{7}
hypopharynxLarynx	4.169	0.506 – 484.8	1.677	0.168 – 1.998 x 10^{7}
nasopharynxNasalCavity	2.336	0.350 – 1.457 x 10^{8}	0.266	0.028 – 1.250
unknownPrimary	0.132	2.020 x 10^{-9} – 95.47	0.903	0.092 – 2.895 x 10^{6}
parotid	1.408	0.097 – 56.81	1.196	0.071 – 27.80
DVH FPLS1	1.004	1.002 – 1.017*	1.005	1.001 – 1.016*
DVH FPLS2	1.002	1.000 – 1.047	-	1.000 – 1.041
DVH FPLS3	-	1.000 – 1.110	-	1.000 – 1.009
DVH FPLS4	-	1.000 – 1.107	-	1.000 – 1.009
DVH FPLS5	-	1.000 – 1.085	-	1.000 – 1.009

95% CI – 95 percentile confidence intervals calculated by bootstrapping the model fitting with 2000 replicates; * - statistically significant at the $\alpha = 0.05$ level; definitiveRT – definitive radiotherapy; indChemo – induction chemotherapy; cisCarbo – one cycle of cisplatin followed by one cycle of carboplatin; DVH FPLSx – functional partial least squares component x of the dose-volume histogram data.

Figure 1

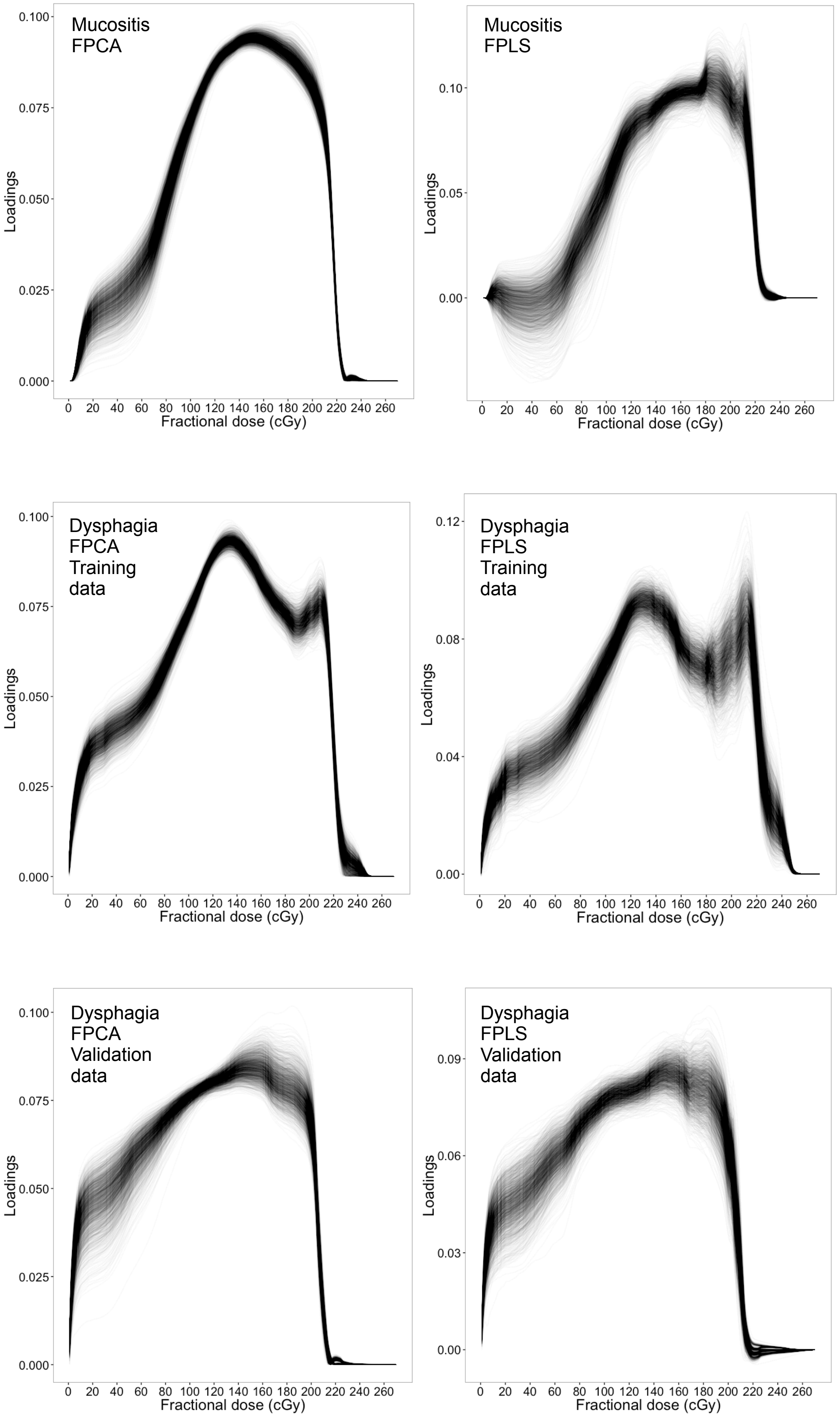
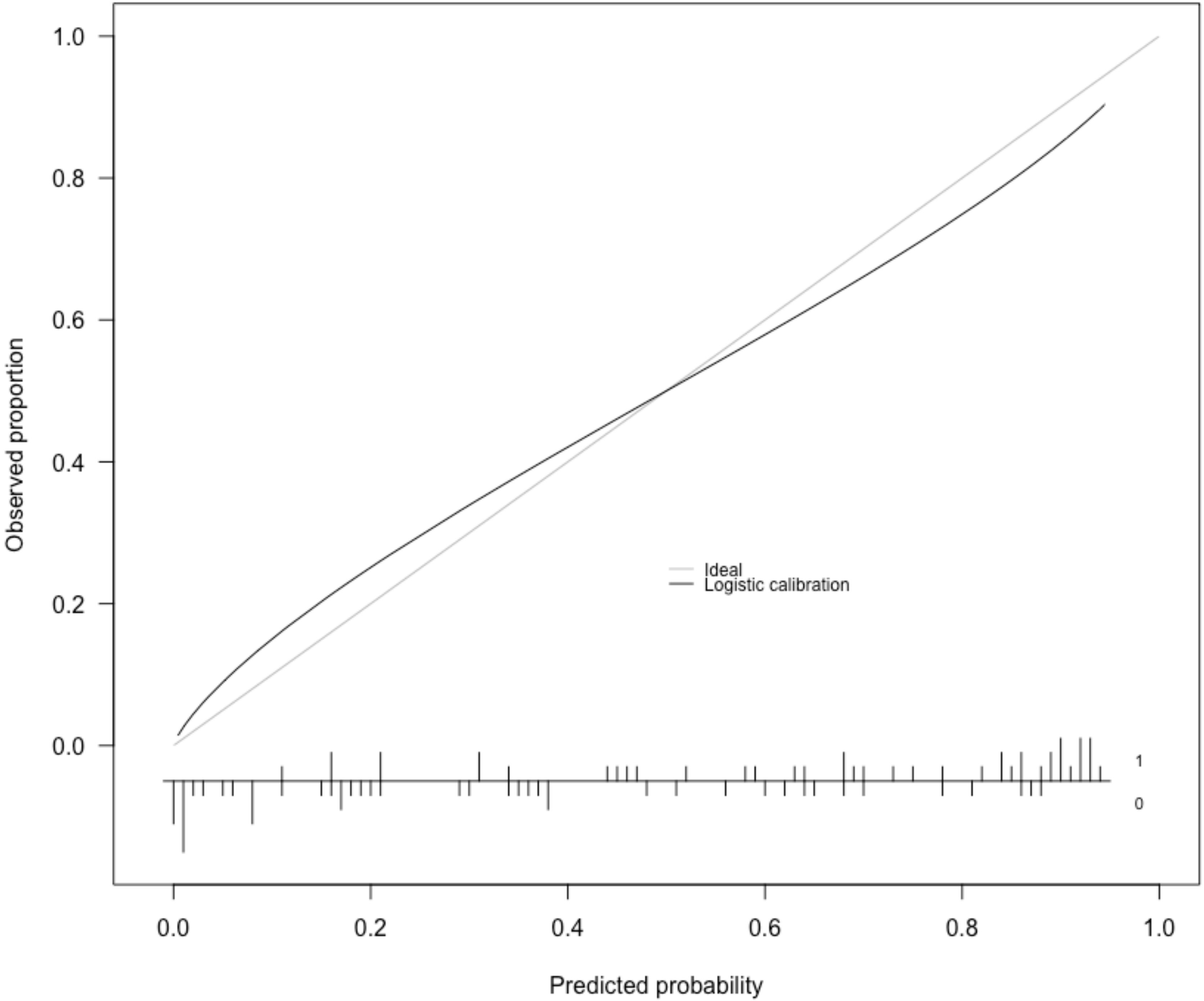

Mucositis
FPCA
0.100
0.075
0.050
0.025
0.000
Loadings
0 20 40 60 80 100 120 140 160 180 200 220 240 260
Fractional dose (cGy)
Mucositis
FPLS
0.10
0.05
0.00
Loadings
0 20 40 60 80 100 120 140 160 180 200 220 240 260
Fractional dose (cGy)
Dysphagia
FPCA
Training
data
0.100
0.075
0.050
0.025
0.000
Loadings
0 20 40 60 80 100 120 140 160 180 200 220 240 260
Fractional dose (cGy)
Dysphagia
FPLS
Training
data
0.12
0.08
0.04
0.00
Loadings
0 20 40 60 80 100 120 140 160 180 200 220 240 260
Fractional dose (cGy)
Dysphagia
FPCA
Validation
data
0.100
0.075
0.050
0.025
0.000
Loadings
0 20 40 60 80 100 120 140 160 180 200 220 240 260
Fractional dose (cGy)
Dysphagia
FPLS
Validation
data
0.09
0.06
0.03
0.00
Loadings
0 20 40 60 80 100 120 140 160 180 200 220 240 260
Fractional dose (cGy)

Figure 2

Click here to download Figure: Figure 2 Submitted.pdf



***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement JAD.pdf

***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement KHW.pdf

*Uniform Disclosures Form

Click here to download Uniform Disclosures Form: Conflict of Interest Statement HG.pdf

***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement LCW.pdf

***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement ABJ.pdf

***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement US.pdf

***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement JHO.pdf

***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement AA.pdf

***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement KLN.pdf

***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement SAB.pdf

*Uniform Disclosures Form

Click here to download Uniform Disclosures Form: Conflict of Interest Statement KJH.pdf

***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement JOD.pdf

***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement CMN.pdf

***Uniform Disclosures Form**

Click here to download Uniform Disclosures Form: Conflict of Interest Statement SLG.pdf