
Article:
Christophides, D, Appelt, AL orcid.org/0000-0003-2792-9218, Gusnanto, A et al. (2 more authors) (2018) Method for Automatic Selection of Parameters in Normal Tissue Complication Probability Modeling. International Journal of Radiation Oncology Biology Physics, 101 (3). pp. 704-712. ISSN 0360-3016

https://doi.org/10.1016/j.ijrobp.2018.02.152

Title:

A method for automatic selection of parameters in NTCP modelling

Authors:

Damianos Christophides[1, 2], Ane L. Appelt[1,2,3], Arief Gusnanto[4], John Lilley[1] and David Sebag-Montefiore[1,2]

1. Leeds Cancer Centre, St James's University Hospital, Leeds, UK

2. Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, United Kingdom

3. Danish Colorectal Cancer Center South, Vejle Hospital, Vejle, Denmark

4. Department of Statistics, University of Leeds, United Kingdom

Running title:

Automatic NTCP modelling

Corresponding author:

Damianos Christophides

Radiotherapy Physics, Level 1, Bexley Wing, St James's University Hospital, Beckett Street, Leeds LS9 7TF, United Kingdom, email: D.Christophides@leeds.ac.uk, Tel: +44 (0) 113 2067553

Authors responsible for statistical analyses:

Damianos Christophides

Radiotherapy Physics, Level 1, Bexley Wing, St James's University Hospital, Beckett Street, Leeds LS9 7TF, United Kingdom, email: D.Christophides@leeds.ac.uk, Tel: +44 (0) 113 2067553

Arief Gusnanto

Department of Statistics, School of Mathematics, University of Leeds, Leeds LS2 9JT, email: A.Gusnanto@leeds.ac.uk, Tel: +44 (0) 113 3435135

Conflict of interest statement

The authors have no relevant conflicts of interest to disclose.

Acknowledgments

DC was funded by the Cancer Research UK Leeds Centre and AA is supported by Yorkshire Cancer Research Academic Fellowship funding (grant L389AA). The support from Prof. Anders Jakobsen, Vejle Hospital, Denmark, and Prof. Ivan Vogelius, Rigshospitalet, Denmark, for the original data collection and analysis is gratefully acknowledged.

No funders had any involvement in the study design, in the collection, analysis and interpretation of data; in the writing of the manuscript; or in the decision to submit the manuscript for publication.

Title:

A method for automatic selection of parameters in NTCP modelling

Running title:

Automatic NTCP modelling

Summary:

The selection of robust parameters for normal tissue complication probability models poses a challenge due to the high number of parameters and potential collinearity. In this work an automatic method is developed that aims to overcome these challenges using principal component analysis, genetic algorithms and bootstrap methods. The results of the proposed algorithm are compared to a published model, using the same patient cohort, and are found to provide equivalent predictive performance.

Abstract:

Purpose: In this study we present a fully automatic method to generate multiparameter normal tissue complication probability (NTCP) models and compare its results with a published model of the same patient cohort.

Methods and Materials: Data were analysed from 345 rectal cancer patients treated with external radiotherapy to predict the risk of patients developing grade 1 or ≥2 cystitis. In total 23 clinical factors were included in the analysis as candidate predictors of cystitis.

Principal component analysis (PCA) was used to decompose the bladder dose volume histogram (DVHs) into 8 principal components (PCs), explaining more than 95% of the variance. The dataset of clinical factors and PCs was divided into training (70%) and test (30%) datasets, with the training dataset used by the algorithm to compute an NTCP model. The first step of the algorithm was to obtain a bootstrap sample, followed by multicollinearity reduction using the variance inflation factor (VIF) and genetic algorithm optimisation to determine an ordinal logistic regression model that minimises the Bayesian information criterion (BIC). The process was repeated 100 times and the model with the minimum BIC was recorded on each iteration. The most frequent model was selected as the final 'automatically generated model' (AGM). The published model and AGM were fitted on the training datasets and the risk of cystitis was calculated.

Results: The two models had no significant differences in predictive performance both for the training and test datasets (p-value>0.05), and found similar clinical and dosimetric factors as predictors. Both models exhibited good explanatory performance on the training dataset (p-values>0.44) which was reduced on the test datasets (p-values<0.05).

Conclusions: The predictive value of the AGM is equivalent to the expert-derived published model. It demonstrates potential in saving time, tackling problems with a large number of parameters and standardising variable selection in NTCP modelling.

Introduction

The use of radiotherapy techniques like 3D-conformal radiotherapy (3D-CRT) and intensity modulated radiotherapy (IMRT) makes it possible to modify the dose distribution to accommodate for patient-to-patient anatomical variations and different tumour geometries. To maximize patient benefit and make the most of this capability, the clinician needs to be able to make informed decisions on how the patient outcome can change depending on the dose distribution. The clinical aim would then be to maximise the probability of achieving tumour control while at the same time minimizing the probability of adverse effects. With the development of models that can predict the normal tissue complication probability (NTCP) (1, 2) dosimetric parameters can be used to estimate such risks.

In addition to the metrics derived from the treatment planning dose distribution, clinical factors have also been shown to contribute to NTCP. Smoking is a predictor of treatment-related pneumonitis(3); also previous surgery and bowel disease contributes to the risk of late gastrointestinal toxicity(4). However as the number of clinical and dosimetric parameters increases, to maximize the predictive performance of models (5), there is the danger of overfitting the model onto noise in the data. It is thus important that the model parameters are selected using methods aimed at maximizing predictive performance while avoiding overfit.

A straightforward method would be to evaluate all possible models that can be derived from a set of parameters but this can be prohibitively time consuming since the number of models increases exponentially with the number of parameters. Although there is no widely accepted consensus on how parameters should be selected, some methods have been investigated in the literature using sequential parameter selection(6), the least absolute shrinkage and selection operator (LASSO)(7) and genetic algorithms(8). There are also several methods that can be used to assess the predictive performance of a model for parameter selection, with El Naqa et al(6) using leave-one-out cross-validation. Caution is

needed when using such techniques in which there is an overlap between the datasets used in evaluating performance, as such a test dataset independent of model parameter selection should be used to report final predictive value. Ideally such a test dataset should be external to the institute deriving the model(9, 10). Furthermore it is important to compare newly derived models with already established models to remove any intra-institutional biases(11).

Multicollinearity also imposes a significant challenge in NTCP modelling, particularly when volume data from a dose-volume histogram (DVH) are used as a way to extract useful dose metrics from the 3D dose distribution. A straightforward approach in trying to reduce multicollinearity from the DVH is to resample the volume data considered in the NTCP model, for example use volume data at 5 Gy increments (12). Alternatively principal component analysis (PCA) can be used to reduce the DVH to linearly uncorrelated components(13, 14).

In this work a fully automatic algorithm is developed and evaluated to address the challenges in NTCP modelling, using a combination of PCA, genetic algorithms, bootstrap, and independent model evaluation from a test dataset. Furthermore the generated model is compared with an independently-derived model(15) of the same patient cohort from another institution in an effort to link knowledge-based and machine learning modelling approaches.

Methods and Materials

Automatic modelling algorithm

The algorithm developed was designed to address the challenges of multicollinearity, parameter selection to balance overfit and underfit, and independent assessment of model predictive performance. Ordinal logistic regression was used to calculate the probability of presenting with the different grades of cystitis, to match the previously published model on the same patient cohort(15).

The implementation of the algorithm was performed in Python v2.7.12 (64bit), using the NumPy numerical Python library (v1.11.1), the scientific SciPy library (v0.18.1) and the Pandas data analysis library (v0.18.1). Fitting of the ordinal logistic regression models was performed using the VGAM library of R (R Foundation for Statistical Computing, Vienna, Austria), accessed from Python using the rpy2 module (v2.7.8).

Normalised cumulative DVHs were used in the analysis. PCA was performed as described in Dawson et al(13), to extract the principal components (PCs) explaining >95% of the variance data, resulting in the first 8 PCs being used (Supplementary material: Appendix A). The clinical parameters were combined with the PCs resulting in a total of 31 parameters to be used by the algorithm. The dataset was then split into training (70%, N=241) and test datasets (30%, N=104), maintaining the proportionality between the different cystitis grades (grade 0=48%, grade 1=40%, grade ≥2=12%); with the 70%/30% ratio selected as a general rule of thumb. The training dataset was used to determine the optimal model, whereas the test dataset was used to evaluate the model.

Bootstrap was used to obtain random samples (N=241) with replacement from the training dataset, this was done to be able to analyse the stability of the model selection process in the presence of random fluctuations in the data. The issue of multicollinearity was addressed using the variance inflation factor (VIF) calculated by (1)

$$VIF_i = \frac{1}{1-R^2{}_i}, \quad (1)$$

where $R^2{}_i$ is the coefficient of determination of fitting a linear regression model with the ith parameter as the independent variable and the rest of the parameters as the dependent variables. As a general rule of thumb(16) a value of 5 was set to indicate if a variable was excessively collinear with the other parameters included in the analysis, as such if $\max\{VIF_i\}$

≥5 then the parameter with the maximum VIF was removed. The multicollinearity reduction process was repeated until max$\{VIF_i\}$ <5.

Ideally all possible models would be considered in the next step to determine the model with the best predictive performance but since there were 31 parameters overall this would lead to ~2 billion potential models. Instead a genetic algorithm was used to optimise the parameter selection process, since genetic algorithms have the advantage to be capable of converging to the global minimum in complex optimisation problems with a large number of parameters(17). For this work the genetic algorithm was implemented to select a model that minimises the Bayesian information criterion (BIC) (Fig. 1).

In total 100 bootstrap samples were used in the algorithm, for each sample the VIF was used to remove collinear parameters and an optimised model was selected that minimised the BIC. At each step of the bootstrap process the model parameters were recorded and the model with the highest recorded frequency was chosen to be fitted on the training dataset without bootstrap sampling. The resulting model was denoted as the 'automatically generated model' (AGM) and was compared with the published model(15). The full algorithm flowchart is shown in Fig. 2. The code used in this work can be found at online at https://github.com/blindedauthor/AutoRegression; including the eigenvectors, mean DVH and example code necessary to transform new bladder DVHs into principal components that can be related directly to this work.

Final model evaluation

The final model evaluation was performed by fitting both the AGM and the model published in (15) onto the training dataset (N=241). The predicted probability of presenting with a toxicity grade ≥1 and ≥2 could then be calculated using these models and compared with the observed values, both on the training (N=241) and test (N=104) datasets. By grouping the data into 6 groups of increasing probability of toxicity the two-sided Fagerland-

Hosmer test(18) was performed to determine the goodness-of-fit between the predicted and observed data. In addition the discriminative ability of the models fitted on the training dataset was assessed using receiver operating characteristic (ROC) curves with any differences between the ROC areas under the curve (AUCs) determined using the one-sided test proposed by Delong et al(19).

Fisher's exact test was used to evaluate whether the selection frequency of the final model was significantly higher compared to the second most frequently selected model. The level of significance for all tests was set at a p-value of 0.05.

Patient data

This study included 345 patients treated for rectal cancer with long course chemotherapy and external radiotherapy between January 2007 and May 2012. The data has previously been published in a paper on dose response modelling of acute bladder toxicity(15).

Radiotherapy treatments were planned using the Oncentra Masterplan (Elekta, Stockholm, Sweden) treatment planning system with either three-field 3D-CRT or with IMRT. Radiotherapy prescriptions to the tumour and elective volumes were 50.4 Gy in 28 fractions (N=219), 60 Gy in 30 fractions (N=117) with the remaining 9 patients receiving between 27 Gy in 15 fractions to 62 Gy in 31 fractions. A small minority of patients (11%) received an external boost to the tumour using 3D-CRT (up to a total dose of 60-62 Gy) and 40% of the patients received additional brachytherapy tumour boosts (5-10 Gy in 1-2 fractions), as part of a clinical trial (20, 21). The brachytherapy tumour boost dose was handled as an independent continuous variable in the outcome modelling, whereas the brachytherapy status as a binary variable of 0/1 values. Potentially relevant clinical factors were retrospectively extracted from the available data including gender, disease stage, age, chemotherapy status; as well as brachytherapy status and dose. The Computational Environment for Radiotherapy

Research (CERR)(22) software was used to extract the DVHs from the DICOM files of the plans.

During the course of the treatment, cystitis was scored weekly by trained nurses using CTCAE v3.0; the weekly scores were concatenated and the highest score was used for NTCP modelling. In total there were 166 patient presented with grade 0 toxicity, 138 with grade 1, 39 with grade 2 and two with grade 3. It was decided to concatenate the patients that had grade 2 and grade 3 into one group of grade 2/3, since only 2 patients presented with CTCAE grade 3 cystitis. Patients with missing data values were filled based on the predictive mean matching technique via the R (R Foundation for Statistical Computing, Vienna, Austria) package 'MICE' (23); these included diabetes status ($N_{missing}$=4), partner ($N_{missing}$=5), prior operation ($N_{missing}$=8), and metastatic stage ($N_{missing}$=1).

Results

The execution time of the proposed algorithm (Fig. 2) was approximately 4 hours running in parallel on an Intel i7-6700 CPU with 16 GB of RAM. The algorithm was run four times to investigate the repeatability of model selection and the same model was chosen in all calculations with a mean selection frequency of 16.5% (range=13%-19%) (Supplementary material: Appendix B), with the second most frequent model varying and having a selection frequency of less than 5%. The lowest selection frequency of the final model of 13% was significantly higher compared to the highest of the second most selected model of 5% (p-value<0.05). Furthermore the robustness of the AGM model is demonstrated by considering the mean selection frequency over the four repeated runs of the algorithm, with both the AGM and its parameters selected with statistically significant higher mean frequency compared to other models and parameters (Supplementary material: Appendix C).

The model derived in (15) identified the $V_{35.4Gy}$ as a significant dosimetric predictor of cystitis along with gender and the brachytherapy dose. The model derived from the proposed method also established three parameters as important predictors of cystitis, including gender with the coefficient values calculated from both models overlapping over their 95% confidence intervals (Table 1). The dosimetric parameter identified by AGM was PC1 which was found to be highly correlated with the $V_{35.4Gy}$ with a Pearson's r of -0.98. The remaining parameters differed with the external boost selected for the AGM and brachytherapy dose for the published model(15).

The calibration of the two models in predicting the observed risk of presenting with grade≥1 or grade≥2 cystitis was visualised using scatter plots with good agreement for both models on the training dataset and worsening results on the test dataset (Fig. 3). The Fagerland-Hosmer test(18) found no significant deviations on the goodness-of-fit for the training datasets of both models (p-values>0.44) but with significant disagreement for the test datasets (p-values<0.05), confirming the significant worsening of the calibration observed visually (Fig. 3).

The discriminative ability of the models was further tested by calculating ROC curves for the probabilities of having grades ≥1 and ≥2 cystitis for both the training and test datasets (Fig. 4). The results show that the AGM has marginally improved discriminative ability for cystitis risk grade ≥2 with an AUC of 0.67 compared to 0.62 for the published model when used on the test dataset however the AUC differences between the AGM and the published model were not statistically significant, for all grades and datasets used (p-values>0.05). Both models had a significant reduction in ROC AUC between training and test datasets for risk grade ≥1 (p-value<0.05) however this reduction was not statistically significant for risk grade ≥2 for both the AGM (p-value=0.16) and the published model (p-value=0.12) (Fig. 4).

The dose response relationship of the $V_{35.4Gy}$ and PC1 was visualised to assess the agreement of the observed probability of complication with the calculated NTCP from the models (Table 1), across the range of the dosimetric values found in the total patient sample (N=345) (Fig. 5).

Discussion

There is a clinical need for robust NTCP modelling to guide clinical decisions in the context of personalised radiotherapy, especially for sites with a limited number of available studies investigating toxicity such as rectal cancer. In this work a fully automated algorithm is presented that can remove operator bias in NTCP modelling parameter selection. The performance of the algorithm is evaluated by comparing the results obtained to a published model(15) on the same patient cohort. The published model and the AGM are shown to have equivalent discriminative ability, both quantitatively using statistical metrics of predictive performance and qualitatively by comparative plots (Fig. 3-5). The agreement shows the potential of machine learning methods in complementing classical knowledge-based NTCP modelling.

The selection of the variables to include in a multivariate NTCP model is a challenging task which should lead to a desirable balance between overfit and underfit, otherwise the model's predictive performance could suffer. In this work the optimal parameters to include in the NTCP model were calculated based on a genetic algorithm used to minimise the BIC. The use of genetic algorithms allows for simultaneous search for both the model order and parameters compared to forward variable selection used by El Naqa et al(6) in multivariate NTCP logistic regression models of esophagitis and xerostomia. Gayou et al(8) also used a genetic algorithm to determine an optimal multivariate logistic regression models that explained the incidence of radiotherapy-induced lung injury. The application of genetic algorithms by Gayou et al is markedly different from this work since they run optimisations

with constant model order having individuals being represented by strings of integers rather than bits, this enabled them to compare their findings with the DREES software tool that uses sequential forward variable selection(24). One of the aims of this work is to compare the algorithm proposed to published results, as such the model order was not predefined since an expert with radiobiology-specific knowledge would not start from a pre-defined order to derive the model.

A limitation of this study is the poor predictive performance of both the expert-derived model and AGM on the test dataset compared to the training dataset (Fig. 3, Fig. 4). Mbah et al (25) has demonstrated that good predictive performance of models on the training dataset does not necessarily translate to good performance on the test data. In general statistical models like ordinal logistic regression, although they do provide for better interpretation of the results (parameter coefficients, p-values, confidence intervals), they do not have as good predictive performance as machine learning models(25). However the models presented do have an explanatory value, rather than predictive, by highlighting statistically significant parameters that explain the risk of cystitis (Table 1). Modelling approaches like LASSO, ridge regression and elastic net allow handling of overfit by tuning of their hyper-parameters. However using this approach would not allow for direct comparison of the resulting model to the published results (15), an important validation step in this work.

The final output model of the proposed algorithm can potentially vary depending on the selection of the variable values. To test the robustness of the method algorithm calculations were performed a total of twelve times, with different parameter values, demonstrating the stability of the model selection (Supplementary material: Appendix B). The only instance in which the algorithm did not select the AGM presented (Table 1) was when the VIF was equal to 3; instead a model including only PC1 and male was selected, with a frequency of 13%. This can be explained from the VIF threshold value of 3 corresponding to a relatively small

R^2 of 0.67 (eq. (1)) that resulted in the removal of a greater number of collinear parameters before the genetic algorithm stage (Fig. 2); this caused the overall algorithm to favour models with fewer parameters. The robustness analysis also highlights the long execution time of the proposed algorithm of approximately 4 hours, running in parallel on an Intel i7-6700 CPU with 16 GB of RAM. Although the execution time is not excessively long, considering that the algorithm is automatic and can be run unsupervised, it can be limiting in how many tests can be performed to investigate its robustness, with a total calculation time in this study of 52 hours (Supplementary material: Appendix B).

Technical and statistical considerations need to be taken into account when applying the proposed method to other datasets. One consideration is the minimum number of patients needed for the analysis. There needs to be enough patients to ensure that the bootstrap process provides a rich enough set of distinct bootstrap samples to perform the analysis. We would recommend at least 20 patients for the training dataset resulting in ~6.9×10^{10} distinct bootstrap samples, enough to reliably perform model selection. Furthermore it is important that the user performs repeatability measurements, similar to this work (Supplementary material: Appendices B and C), to ensure that the model selection is robust. This might require increasing the number of bootstrap iterations in the presence of varying population sizes, number of events, effect size and level of significance of the features; we have investigated the effect of population size and number of bootstrap iterations in Supplementary material: Appendix D. It also needs to be emphasized that the algorithm might not be able to provide a robust model because of excessive noise in the data and very little or no dependence of the outcome on the features under consideration. It is important that the user is familiar with the statistical and machine learning techniques used by the algorithm to diagnose such problems and to already have performed descriptive statistical analysis to investigate the characteristics of the dataset. The issues mentioned make evaluating the final

model on a test dataset crucial and in all cases this must be performed with a test dataset size that provides sufficient confidence on the model. The run time of the method might also be a concern for larger datasets. In such cases we would recommend running the algorithm using cloud computing on high performance servers or clusters of servers.

Ideally NTCP models should be validated against datasets external to the institution that produced the model to determine the generalizability (9, 10). This type of external validation has been performed in the literature, with Jayasurya et al(26) reporting very good predictive performance of Bayesian network and support vector machine (SVM) models on datasets collected from three different cancer centres. In this study the predictive performances of the published model(15) and AGM were reported on a test dataset that was not used in the training and fit of the models, since an external validation was not available. Although this is a limitation the main aim of the study is to provide a general framework for generating NTCP regression models, and compare it against an expert-derived model already published(15).

In this study PCA was used to decompose the DVH data into linearly uncorrelated PCs. This has the advantage of removing multicollinearity and enabling better feature selection, since the standard errors of the features are not inflated due to collinearity. However the disadvantage of using PCs is that it is difficult to interpret them and thus use them practically in treatment planning and evaluation. A method to overcome this limitation is to find correlations with established dose metrics and use them as surrogates in treatment planning optimisation and evaluation, for example Sohn et al (14) found correlations of PC1 with V_{60Gy} and D_{mean} in their investigation of modelling rectal toxicity following prostate radiotherapy. Another approach would be to use the corresponding eigenvector of a PC to derive theoretical DVHs that can result in a reduction of the predicted risk and use them as a guide in treatment plan optimisation (27).

The good agreement of the parameters of the AGM and the published model should be noted (Table 1). Both models selected gender as a significant parameter with its significance indicated from the percentage increase of male patients from 54%, 65% and 76% with increasing cystitis grade. Also the dosimetric parameter of PC1, included in the AGM has a high correlation with $V_{35.4Gy}$ (r=-0.98), explaining similar dosimetric differences in the patient sample. The disagreement between the models is in the AGM having the external boost as a third parameter instead of brachytherapy dose found in the published model (Table 1). However there is an inter-dependence between the brachytherapy dose and external boost since patients that had brachytherapy did not have an external boost, with the external boost parameter having a negative regression coefficient (Table 1) expressing the reduced risk of cystitis for patients that did not have brachytherapy. This point highlights the importance of the clinical explanation of any NTCP modelling to avoid erroneous clinical conclusions.

Conclusions

A method is presented to automatically generate NTCP models addressing the challenges of variable selection, collinearity and model validation. The automatically-generated multivariate NTCP model was validated against a knowledge-based published model from the same patient cohort, and was found to have equivalent explanatory and predictive performance. The algorithm presented can complement knowledge-based approaches to produce NTCP models providing additional confidence in the derived model. In addition it can potentially save time, tackle problems with a large number of parameters and standardise variable selection in NTCP modelling.

References

1. Lawrence TS, Ten Haken RK, Kessler ML, et al. The use of 3-D dose volume analysis to predict radiation hepatitis. Int. J. Radiat. Oncol. 1992;23:781–788.

2. Lyman JT, Wolbarst AB. Optimization of radiation therapy, IV: A dose-volume histogram reduction algorithm. Int. J. Radiat. Oncol. 1989;17:433–436.

3. Jin H, Tucker SL, Liu HH, et al. Dose–volume thresholds and smoking status for the risk of treatment-related pneumonitis in inoperable non-small cell lung cancer treated with definitive radiotherapy. Radiother. Oncol. 2009;91:427–432.

4. Rancati T, Fiorino C, Fellin G, et al. Inclusion of clinical risk factors into NTCP modelling of late rectal toxicity after high dose radiotherapy for prostate cancer. Radiother. Oncol. 2011;100:124–130.

5. Lind PA, Marks LB, Hollis D, et al. Receiver operating characteristic curves to assess predictors of radiation-induced symptomatic lung injury. Int. J. Radiat. Oncol. 2002;54:340–347.

6. El Naqa I, Bradley J, Blanco AI, et al. Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors. Int. J. Radiat. Oncol. 2006;64:1275–1286.

7. Lee T-F, Chao P-J, Ting H-M, et al. Using Multivariate Regression Model with Least Absolute Shrinkage and Selection Operator (LASSO) to Predict the Incidence of Xerostomia after Intensity-Modulated Radiotherapy for Head and Neck Cancer Cordes N, ed. PLoS One. 2014;9:e89700.

8. Gayou O, Das SK, Zhou S-M, et al. A genetic algorithm for variable selection in logistic regression analysis of radiotherapy treatment outcomes. Med. Phys. 2008;35:5426–33.

9. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann. Intern. Med. 2015;162:55.

10. Meldolesi E, van Soest J, Damiani A, et al. Standardized data collection to build

prediction models in oncology: a prototype for rectal cancer. Futur. Oncol. 2016;12:119–136.

11. Kang J, Schwartz R, Flickinger J, et al. Machine Learning Approaches for Predicting Radiation Therapy Outcomes: A Clinician's Perspective. Int. J. Radiat. Oncol. 2015;93:1127–1135.

12. Robinson M, Sabbagh A, Muirhead R, et al. Modeling early haematologic adverse events in conformal and intensity-modulated pelvic radiotherapy in anal cancer. Radiother. Oncol. 2015;117:246–51.

13. 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;62:829–37.

14. 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. 2007;69:230–239.

15. [blinded]

16. Craney TA, Surles JG. Model-Dependent Variance Inflation Factor Cutoff Values. Qual. Eng. 2002;14:391–403.

17. Sivanandam SN, Deepa SN. Introduction to Genetic Algorithms. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008.

18. Fagerland MW, Hosmer DW. Tests for goodness of fit in ordinal logistic regression models. J. Stat. Comput. Simul. 2016;86:3398–3418.

19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45.

20. Appelt AL, Vogelius IR, Pløen J, et al. Long-term results of a randomized trial in

locally advanced rectal cancer: no benefit from adding a brachytherapy boost. Int. J. Radiat. Oncol. Biol. Phys. 2014;90:110–8.

21. Jakobsen A, Ploen J, Vuong T, et al. Dose-effect relationship in chemoradiotherapy for locally advanced rectal cancer: a randomized trial comparing two radiation doses. Int. J. Radiat. Oncol. Biol. Phys. 2012;84:949–54.

22. Deasy JO, Blanco AI, Clark VH. CERR: A computational environment for radiotherapy research. Med. Phys. 2003;30:979–985.

23. Buuren S van, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in R. J. Stat. Softw. 2011;45.

24. Naqa I El, Suneja G, Lindsay PE, et al. Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose–volume outcome relationships. Phys. Med. Biol. 2006;51:5719–5735.

25. Mbah C, Thierens H, Thas O, et al. Pitfalls in Prediction Modeling for Normal Tissue Toxicity in Radiation Therapy: An Illustration With the Individual Radiation Sensitivity and Mammary Carcinoma Risk Factor Investigation Cohorts. Int. J. Radiat. Oncol. 2016;95:1466–1476.

26. Jayasurya K, Fung G, Yu S, et al. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. Med. Phys. 2010;37:1401–1407.

27. Christophides D, Gilbert A, Appelt AL, et al. OC-0255: Practical use of principal component analysis in radiotherapy planning. Radiother. Oncol. 2017;123:S129–S130.

Figure Captions

Fig. 1. Illustration of the genetic algorithm optimisation method used to select the parameters in the model that minimize the Bayesian information criterion (BIC). The algorithm

repeatedly modifies an initial population of 200 models, defined as 'individuals', based on principles of biological evolution. The convergence criteria used for the optimisation were 20 consecutive generations without an improvement to the BIC of the best model; with a minimum number of generations of 30. Note that in the illustration shown there are 9 individuals in the population with a number of parameters of 10, whereas in the actual implementation there were 200 individuals and the number of parameters was 31.

Fig. 2. Flowchart of the method for the automatic generation of ordinal logistic regression models. The variance inflation factor (VIF) was used to remove collinear parameters before the genetic algorithm optimisation (Fig. 1) minimised the Bayesian information criterion (BIC).

Fig. 3. Calibration plots comparing a) the published model(15) and b) the automatically generated model (AGM) showing the predicted versus observed grade ≥1 and grade ≥2 risk for the training and test datasets (with 68% confidence intervals).

Fig. 4. Receiver operating characteristic (ROC) curves of a) the published model(15) and b) the automatically generated model (AGM) calculated on the training and test datasets.

Fig. 5. NTCP curves for the model derived from the parameters in a) (15) and b) the automatically generated model (AGM) using PCA. The curves were calculated across the range of the $V_{35.4Gy}$ and PC1 values found in the whole patient dataset (N=345) with the remaining parameters kept constant using their median values. Data points represent observed

toxicity for the training and test datasets, shown for the different grades ≥1 and ≥2 with 68% confidence intervals.