Statistical-learning strategies generate only modestly-performing predictive models for urinary symptoms following external beam radiotherapy of the prostate: a comparison of conventional and machine-learning methods

Noorazrul Yahya^{1,2}

Martin A Ebert^{1,3}

Max Bulsara⁴

Michael J House¹

Angel Kennedy³

David J Joseph^{3,5}

James W Denham⁶

¹ School of Physics, University of Western Australia, Western Australia, Australia

² School of Health Sciences, National University of Malaysia, Malaysia

³ Department of Radiation Oncology, Sir Charles Gairdner Hospital, Western Australia, Australia

⁴ Institute for Health Research, University of Notre Dame, Fremantle, Western Australia, Australia

⁵ School of Surgery, University of Western Australia, Western Australia, Australia

⁶ School of Medicine and Public Health, University of Newcastle, New South Wales, Australia

Number of pages: 17

Number of figures: 3

Number of tables: 2

Web appendix: 3

Correspondence address:

Noorazrul Yahya

School of Physics

University of Western Australia

35 Stirling Hwy

Crawley, Western Australia 6009

Tel: 08 9346 4931

Email: noorazrul.yahya@research.uwa.edu.au

Running Title: Statistical-learning strategies for urinary symptom models

Keywords: normal tissue complications, neural network, support vector machine, elastic-net, support-vector machine, random forest

CONFLICTS OF INTEREST NOTIFICATION

The TROG 03.04 trial was financially supported by grant funding from Australian and New Zealand government, non-government and institutional sources. Pharmaceutical use and trial logistic support was provided by Abbott Laboratories and Novartis Pharmaceuticals. No financial benefits were paid to trial investigators or listed authors. The authors declare that they have no competing interests.

ABSTRACT

Introduction Given the paucity of available data concerning radiotherapy-induced urinary toxicity, it is important to ensure derivation of the most robust models with superior predictive performance. This work explores multiple statistical-learning strategies for prediction of urinary symptoms following external beam radiotherapy of the prostate.

Method The performance of logistic regression, elastic-net, support-vector machine, random forest, neural network and multivariate adaptive regression splines (MARS) to predict urinary symptoms were analysed using data from 754 participants accrued by TROG03.04-RADAR. Predictive features included dose-surface data, comorbidities and medication intake. Four symptoms were analysed; dysuria, haematuria, incontinence and frequency, each with three definitions (grade $\geq$ 1, grade $\geq$ 2 and longitudinal) with event rate between 2.3-76.1%. Repeated cross-validations producing matched models were implemented. A synthetic minority over-sampling technique was utilized in endpoints with rare events. Parameter optimization was performed on the training data. Area under the receiver operating characteristic curve (AUROC) was used to compare performance using sample size to detect differences of $\geq$ 0.05 at the 95% confidence level.

Results Logistic regression, elastic-net, random forest, MARS and support-vector machine were the highest-performing statistical-learning strategies in 3, 3, 3, 2 and 1 endpoints, respectively. Logistic regression, MARS, elastic-net, random forest, neural network and support-vector machine were the best, or were not significantly-worse than the best, in 7, 7, 5, 5, 3 and 1 endpoints. The best-performing statistical model was for dysuria grade $\geq$ 1 with AUROC $\pm$ standard deviation of 0.649 ± 0.074 using MARS. For longitudinal frequency and dysuria grade $\geq$ 1, all strategies produced AUROC $>$ 0.6 while all haematuria endpoints and longitudinal incontinence models produced AUROC $<$ 0.6.

Conclusion Logistic regression and MARS were most likely to be the best performing strategy for the prediction of urinary symptoms with elastic-net and random forest producing competitive results. The predictive power of the models was modest and endpoint-dependent. New features, including spatial dose maps, may be necessary to achieve better models.

INTRODUCTION

The prediction of radiotherapy treatment-related toxicity has been of significant interest¹⁻¹⁶. Toxicity prediction involves relating the dosimetric factors to the toxicity in question in a model, potentially in combination with clinical information. For urinary symptoms, the availability of such models is limited compared to rectal symptoms following prostate radiotherapy. This is despite the observation that urinary toxicity will likely represent the principal dose-limiting factor, due to the decreasing prevalence of rectal symptoms, in the era of dose-escalated radiotherapy for prostate cancer^{17, 18}. The focus of model derivation should mirror this transitioning prevalence¹⁹.

Along with multivariable logistic regression several more advanced statistical-learning strategies have been used and tested in the development of predictive models. A primary motivation for utilising more advanced strategies is the production of models with more predictive power. Strategies have included the use of regularized regression^{14, 15, 20}, neural networks¹⁻⁶, support vector machines^{5, 7}, decision trees⁸, distance-based measures⁹ and ensemble methods^{10, 11} including random forests¹². In most instances, the studies suggest that the more contemporary strategies show promising results and perform better. Nevertheless, except for the series of studies on radiotherapy-induced pneumonitis^{7-10, 13}, the statistical-learning strategies were rarely extensively compared side-by-side. Furthermore, comparisons in the context of toxicity prediction were mostly performed using only one endpoint, probably chosen based on the clinical relevance. Concluding the superiority of a strategy based on only one endpoint may not be optimal for deriving a solid conclusion; findings have less chance of being fortuitous if they are repeated across endpoints/datasets²¹. Also, to our knowledge, there is no study in the literature comparing different learning strategies using urinary-symptom endpoints. Given the paucity of available urinary toxicity data, it is important to ensure derivation of the most robust statistical model.

The aim of this study was to develop the most accurate predictive model for urinary-symptoms following external beam radiotherapy of the prostate by exploring different statistical-learning strategies. This was achieved by modelling 12 urinary-symptom endpoints from the data of patients treated with external beam radiotherapy in the Trans-Tasman Radiation Oncology Group (TROG) 03.04 trial of Randomised Androgen Deprivation and RT (RADAR-NCT00193856)²²⁻²⁴ using six different strategies - logistic regression, elastic net, support-vector machine, random forest, neural network and multivariate adaptive regression splines (also known as enhanced adaptive regression through hinges). The models were optimized with a robust cross-validation method and subsequently compared in term of their predictive capabilities.

METHODS AND MATERIALS

Patients and treatment

The RADAR trial²²⁻²⁴ database, consisting of data for 754 patients who had received external beam radiotherapy (EBRT) of the prostate to either 66, 70 or 74 Gy, was utilized. Data collection, protocol requirements, treatment technique and quality assurance (QA) have been summarized previously^{22, 25-27}. Patients were routinely followed up in the clinic every 3 months for 18 months, then 6 monthly up to 5 years and then annually. Urinary symptoms reported here were assessed using LENT-SOMA²⁸. Focusing on late effects, only follow-ups at least 1 year after randomization (i.e., 5 months after the end of EBRT) were considered.

Dosimetric candidate features

During the RADAR trial, digital treatment plans were independently reviewed and archived to ensure consistency across datasets submitted from different centres²⁹. Physical doses to each voxel in the dose matrix for the bladder for each treatment phase were combined to generate EQD2 (equivalent dose in 2 Gy fractions using $\alpha/\beta=6$ Gy^{30,31}). The relative surface area of tissue receiving more than threshold doses (A1Gy, A2Gy...A75Gy) were calculated from (EQD2) bladder dose-surface histogram (DSH) data. Also, the equivalent uniform doses (EUDs) were derived from the DSHs using the standard formulation³² (Appendix A). Due to the high correlation among dose thresholds and EUDs, an unsupervised feature clustering process, which assesses the collinearity of the features and separates them into clusters, was performed. A similarity measure based on Spearman correlations (cut-point for clustering at $\rho=0.8$) was utilized where only one feature with the largest variance from each cluster was selected. Further analysis was therefore limited to 5 dosimetric features; EUD1 (EUD using $\alpha=1$), EUD4, EUD8, EUD16 and A75 (Appendix A).

Clinical candidate features

Besides the dosimetric information, 23 other features, collected at randomization, were also introduced to improve the predictive power of the models:

- The physical characteristics of the patients (age, body mass index (BMI) and ECOG performance status).
- Clinical co-morbidities (cardiovascular condition, peripheral vascular condition, cerebrovascular condition, hypertension, dyslipidaemia, non-insulin dependent diabetes mellitus (NIDDM), respiratory disorder, bowel disorder, dermatological disorder, bone or calcium metabolism disorder, haematological disorder and thyroid disorder).
- Medication intake (insulin, hypoglycaemic agents, ACE inhibitor, statin, nonsteroidal anti-inflammatory drug and anti-coagulant).
- Lifestyle factors (smoking status and alcohol intake).

The immediate aim in this study was not to study the impact of each feature, but rather to investigate the impact of statistical-learning choices on the predictive power of resulting models. More complete descriptions of the influence of dosimetry, clinical factors and medication intake based on this dataset have been presented previously^{33,34}.

Endpoints

Four atomized symptoms were analysed; dysuria, haematuria, urinary incontinence and frequency. To quantify the urinary symptom events, different definitions have been suggested in the literature. For a complete overview of the symptoms, three endpoint definitions were utilized in the current study. The first two endpoints were peak grade ≥ 1 and grade ≥ 2 which take into account any symptom event at and above the grade cut-point throughout the follow-up period. The third was a longitudinal definition where patients were only considered to have an event when individual symptoms of grade ≥ 1 were reported ≥ 2 times throughout follow-up.

Data pre-processing

Due to the low rate of missing data on most features (Appendix B), simple imputation was performed where the empty data points in continuous and categorical features were replaced by the median and mode of the particular feature, respectively. All continuous candidate features were pre-processed by standardizing to zero mean and unit variance.

Statistical-learning strategies

In this section, the six statistical-learning strategies used to develop the predictive models will be described. The following strategies were considered; logistic regression, elastic net, support-vector machine, random forest, neural network and multivariate adaptive regression splines (also known as enhanced adaptive regression through hinges) (MARS).

Logistic regression

Logistic regression is used in many applications. In this analysis, conventional logistic regression with backward stepwise elimination of the candidate features was applied. The Akaike information criterion (AIC)-based stopping rule was applied where a feature was removed if the chi-square increase was lower than two times the degrees of freedom.

Elastic net

Elastic net is type of penalized multivariate regression method that linearly combines the penalties used in lasso (least absolute shrinkage and selection operator) and ridge regression approaches³⁵. The lasso favours sparse models, removing all features except one in instances of correlated variables. The ridge-regression approach keeps all features as non-zero but shrinks the estimates of correlated features towards each other producing approximately the same value (known as the grouping effect). Compared to pure lasso and ridge regression, elastic-net provides a compromise between the two approaches. This leads to simultaneous model parameter estimation and variable selection by suppressing the estimates for features with no discriminatory power to be exactly zero due to the lasso properties, thus, producing a sparse model. Also, due to the ridge-regression approach, it is able to robustly model correlated variables by encouraging strongly-correlated predictors to be in or out of the model together³⁵. That is an attractive advantage, suitable for datasets with correlated dosimetric information. Thus, elastic-net was selected from the family of regularized logistic approaches. The final values of α (between 0 to 1, $\alpha=1$ leads to application of the lasso penalty and $\alpha=0$ the ridge penalty) and the regularization parameter, λ , were optimized using cross-validation of the training data.

Random forest

Random forest is an ensemble-learning strategy which fuses the ‘votes’ submitted from many low-correlated decision trees to estimate the probability of events³⁶. Randomness is incorporated in the construction of the decision trees by two means. Firstly, for a dataset (S) with N samples and M features, N equivalently-sized samples are taken (by random sampling with replacement) from S to produce a single decision tree. Secondly, for each split, m feature/s (m is a positive integer $\ll M$) is randomly selected from M candidate features to optimally produce a binary partitioning. The process of randomly selecting m features and creating splits are repeated until the trees are fully grown without any pruning. The process is repeated for all trees. The random selection of features for the split decision ensures low correlation between trees. For each tree, m is kept constant. Combining votes from many independently-grown trees correct for overfitting associated with a single unpruned decision tree. In this analysis, the number of trees grown was fixed at 1000 based on the suggestion of Breiman³⁶ while m was optimized using cross-validation of the training data.

Neural network

Neural network models can deduce potentially non-linear patterns in the data after executing the process of learning. The neural network architecture consists of three neuronal layers: input, hidden, and output. An input layer contains neurons which correspond to the number of input features; a hidden layer provides a relationship or pattern between the input and output;

and an output layer contains an output neuron (in this instance, urinary symptom event). In the hidden layer, the network produces an associated output pattern based on transfer functions which correlate inputs with outputs in various ways so complex decisions can be modelled. In this current analysis, a feed-forward single-hidden-layer network method optimized by the BFGS algorithm was utilized³⁷. The number of hidden nodes and weight decay were optimised with cross validation.

Support-vector machine

Support-vector machine searches for the linear hyperplane optimally separating binary classes. The optimal hyperplane (or decision boundary) is the one that produces the maximal margin between the two classes. The support-vector machine can be applied to both linearly and non-linearly separable data. For non-linearly separable data, the support-vector machine first maps the data into a high-dimensional feature space by using non-linear mapping or using a kernel function and then searches for a linear optimally-separating hyperplane in the new space. Prediction is based on which side of the hyperplane the subject lies on. In this current analysis, the support-vector machine was implemented with a radial basis-function kernel, as widely used in many fields including radiotherapy⁷. Two parameters were required and optimized using cross-validation of the training data; the penalty parameter, C , also called the cost or soft margin which controls model overfitting, and σ which controls the degrees of non-linearity.

Multivariate adaptive regression splines

Multivariate adaptive regression splines (MARS) (or enhanced adaptive regression through hinges) is an extension of a linear model that includes nonlinearities, as developed by Friedman³⁸. MARS is capable of automatically producing local models by partitioning the hyperspace of candidate features into separate regions. In each region, a linear relationship is used to characterize the impact of features on the response producing a locally-linear model. The points at which the local model changes are called 'hinges' which indicate the end of one region of the space and the start of another with its own distinct local model. To prevent the model from overfitting the data, a pruning process is implemented which involves searching for the trade-off between complexity and error based on a backwards elimination feature selection procedure that looks at reductions in the cross-validation estimate of error.

Model building

100 iterations of 10-fold cross-validation

To evaluate the power of the predictive models, stratified ten-fold cross-validation was utilized (Fig. 1). Briefly, the patients were randomly divided into ten folds with the number of patients with an event approximately equal in all folds. The development of the model was performed by using 9 folds and validation on the tenth. The development-validation was executed in ten rounds. Each fold was used exactly once as the validation data. Setting aside one fold for validation produced an unbiased estimate of model predictive power as the patients in the validation fold were not used to develop the model. This experiment of partitioning the data into ten and performing ten-fold cross-validation was repeated for a total of 100 iterations using different random seeds. Large iterations prevent selection bias that could occur if only one 10-fold cross-validation was performed and to obtain a representative set of models³⁹. The process was performed for all six statistical-learning strategies using exactly-matched random seeds resulting in 1000 models based on unique development-validation combinations.

Synthetic minority over-sampling technique (SMOTE)

The event rates for grade 2 and longitudinal haematuria, and grade 2 dysuria, were small (<10%) requiring a different modelling approach. First, the data were divided into 3 folds maintaining approximately equal number of events in each fold. For each fold separately, the synthetic minority over-sampling technique (SMOTE) introduced by Chawla et al.⁴⁰ was utilized to generate new instances of the minority class based on 5 randomly chosen nearest neighbours. The generation of synthetic instances were balanced by under-sampling of negative instances to produce a ratio of 1 positive instance to 9 negative instances. 10% positive instances were considered reasonable to raise the weight of the minority class without excessive generation of synthetic instances. 3-fold cross-validation was performed with two folds for development and the third for validation in each run. The process was repeated 333 times to produce 999 models for each endpoint-strategy combination.

Figure 1: Model building process using 100 iterations of 10-fold cross-validation. For endpoints with event rate <10%, a synthetic minority over-sampling technique was also implemented. Please refer to text for details.

Parameter optimization

Each of the statistical-learning strategies requires several parameters to be set up to optimally construct the model (Table 1). Where required, the optimal parameters were estimated via 3 iterations of 10-fold cross-validation, performed using only data from the development fold, through an exhaustive grid search of a predefined parameter search space (Table 1). The optimum parameters were selected based on oneSE method⁴¹. The model building was performed and streamlined using *caret*⁴² in R 3.2.3 (The R Foundation for Statistical Computing, Vienna, Austria)⁴³ based on these packages; *MASS*³⁷, *glmnet*⁴⁴, *nnet*³⁷, *kernlab*⁴⁵, *randomForest*⁴⁶ and *earth* (<http://cran.r-project.org/package=earth>).

Table 1: Modelling parameters

Method	Modelling parameter	Note
Logistic regression	none	
Support Vector Machines	<i>sigma</i>	controls the degrees of non-linearity
	<i>C</i>	cost, controls model overfitting

Neural Network	size	$C = \{0.25, 0.50, 1.00\}$ number of hidden units size = $\{1, 2 \dots 5\}$
	weight decay	regularization parameter weight decay = $\{0, 10^{-4}, 10^{-1}\}$
Random forest	number of trees	fixed to 1000
	m	number of randomly selected features $m = \{1, 2 \dots 16, 33\}$
Elastic-net	α	mixing percentage $\alpha = \{0.1, 0.2 \dots 1\}$
	λ	regularization parameter $\lambda = \{0.1, 0.2 \dots 1\}$
Multivariate adaptive regression splines	degree	interaction terms – fixed to 1

296

297 *Feature selection and feature importance*

298 Logistic regression, elastic-net, MARS and random forest models do not necessarily utilize
299 all predictors and can thus be described as having built-in feature selection. For neural
300 network and support-vector machine feature selection based on recursive feature elimination,
301 a wrapper method which adds and/or removes predictors to find the optimal combination that
302 maximizes the AUROC, was performed. For each model, the top five features (or less in
303 instances where <5 features were included in the final model) were determined using model-
304 based technique for measuring variable importance; random forest using permutation method
305 in randomForest package³⁶, logistic regression and elastic-net using the coefficients, MARS
306 using generalized cross-validation estimate of error reduction when each a feature is added to
307 the model. For neural network and support vector machine with feature selection external to
308 the modelling process, the ranking method used in recursive feature elimination was used to
309 determine feature importance.

310

311 *Model comparison*

312 To compare the power of the statistical-learning strategies outlined above, the area under the
313 receiver operating characteristic curve (AUROC) was evaluated. The AUROC is a common
314 method to assess the power of a binary statistical-learning strategy as its discrimination
315 threshold is varied across all cut-off values. AUROC can take a value between 0 and 1; an
316 AUROC of 1 represents a perfect classification prediction; 0.5 represents a classification with
317 discrimination no better than random; 0 represents a model with all validation instances
318 predicted with a wrong label. In total, 1000/999 models were derived for each learning
319 strategy with a matching subset of patients for both model development and validation from
320 which the mean and standard deviation of the AUROC was derived. However, the utilization
321 of all matched models in a two-tailed matched-pair *t*-test would provide the power to resolve
322 a very small difference between strategies producing statistical significance even for small
323 differences. Therefore, based on the distributions of the resultant models (population mean

and standard deviation), α error probability of 0.05, power of 0.95 and correlation between models, the appropriate sample size for each comparison was calculated to detect a mean AUROC difference between models of 0.05 (arbitrarily chosen as a threshold point for a meaningful difference). The first required model pairs were sampled for comparisons. Correlations between the predictive powers of the models using different statistical-learning strategies were calculated using Spearman correlation. p -Value of 0.05 is considered significant.

RESULTS

Description of study sample

A statistical summary of the features, including the standard deviation and number of missing data for each feature, is available in Appendix B. Event rates for the twelve endpoints range from 2.3% for haematuria grade ≥ 2 to 76.1% urinary frequency grade ≥ 1 (Table 2). Longitudinal haematuria and dysuria grade ≥ 2 have low event-prevalence ($<10\%$).

Table 2: The distribution of events for atomized symptoms (dysuria, haematuria, incontinence and frequency) based on three definitions (peak grade ≥ 1 , peak grade ≥ 2 and longitudinal).

		Event	Non-event	% of event
Dysuria				
	Peak grade ≥ 1	153	592	20.5
	Peak grade ≥ 2	53	693	7.1
	Longitudinal	97	649	13.0
Haematuria				
	Peak grade ≥ 1	101	645	13.5
	Peak grade ≥ 2	17	729	2.3
	Longitudinal	26	720	3.5
Incontinence				
	Peak grade ≥ 1	218	528	29.2
	Peak grade ≥ 2	91	655	12.2
	Longitudinal	117	629	15.7
Frequency				
	Peak grade ≥ 1	568	178	76.1
	Peak grade ≥ 2	216	530	29.0
	Longitudinal	434	312	58.2

Comparison of predictive power

Logistic regression, elastic-net, random forest, MARS and SVM were the highest-performing statistical-learning strategies in 3, 3, 3, 2 and 1 endpoints, respectively (Fig. 2). In 7 endpoints, the differences relative to one or more other strategies were not statistically significant. Logistic regression, MARS, elastic-net, random forest, neural network and support-vector machine were the best, or were not significantly-worse than the best, in 7, 7, 5, 5, 3 and 1 endpoints, respectively. The best-performing statistical model was for dysuria

grade $\geq$ 1 with AUROC $\pm$ standard deviation of 0.649 ± 0.074 using MARS, followed by longitudinal frequency using logistic regression (0.647 ± 0.057) and dysuria grade $\geq$ 1 using logistic regression (0.644 ± 0.080).

The predictive power of the statistical models was dependent on the endpoint in question. Two endpoints with the highest average AUROC across all statistical-learning strategies were grade $\geq$ 1 dysuria (0.634 ± 0.081) and longitudinal frequency (0.624 ± 0.069). Endpoints with the lowest average AUROC were haematuria longitudinal and grade $\geq$ 2 (0.493 ± 0.110 and 0.479 ± 0.097). Dysuria (grade $\geq$ 1), incontinence (grade $\geq$ 1 and longitudinal) and all frequency endpoints were less dependent on the model used (i.e. no significant difference between at least 3 models). For longitudinal frequency and dysuria grade $\geq$ 1, all strategies produced AUROC $>$ 0.6 while all haematuria endpoints and longitudinal incontinence models produced AUROC $<$ 0.6.

Figure 2: Predictive performance for the statistical-learning strategies for urinary symptom endpoints and the standard deviations based on 1000 models.

Note: AUROC – area under the receiver operating curve; ^ model with the highest, * not significant worse than the highest AUROC for the endpoint. G1: grade ≥ 1 ; G2: grade ≥ 2 ; Long: longitudinal. The error bars represent the standard deviations.

Correlations of predictive power

The correlation between the predictive power of the models using different statistical-learning strategies revealed a pattern. The correlation was highest between elastic-net and logistic regression; both of which are variations of generalised linear models (Fig. 3). The correlations were dependent on the endpoints, generally higher for dysuria grade ≥ 1 , frequency grade ≥ 2 , and longitudinal frequency. These were also the endpoints that resulted in better predictive power overall (average >0.6) and smaller standard deviation.

Feature importance

The important features for endpoints were found to share many similarities between different learning strategies (Appendix C). Common predictors were found to be less substantial for haematuria endpoints.

Figure 3: Correlation between predictive performance using different statistical-learning strategies (Spearman rho, ρ)

DISCUSSION

The prediction of urinary symptoms is uniquely challenging; patients treated for prostate carcinoma tend to from older portions of the population and subject to myriads of urinary symptoms which may or may not be related to the treatment. This introduces noise to post-treatment urinary assessment. Together with the paucity of available urinary toxicity data, this makes the availability of an optimal learning strategy to improve prediction crucial. This study builds upon a developing body of literature suggesting the improvement of prediction capabilities of different statistical-learning strategies which may be an avenue for improvement in the prediction and alleviation of post-treatment urinary-symptoms.

Compared to most studies related to the use of statistical-learning strategies in the prediction of toxicity outcome in radiotherapy, this work considered a wider range of endpoints. The advantages of using more than one endpoint are at least twofold; first, to provide a selection of datasets that represent the full urinary-symptom modelling problem domain. Readers may appreciate the differences between learning strategies through the use of atomized symptom types and diverse definitions of symptom grade (peak grade ≥ 1 , peak grade ≥ 2 and multiple grade ≥ 1). Peak grade ≥ 1 provides a larger number of events while subjected to more noise while peak grade ≥ 2 provides a significantly-smaller number of events and potentially less noise. The longitudinal grade ≥ 1 tries to compensate for the large potential of noise associated with grade 1 symptoms by incorporating the persistence of symptoms. Second, hypothetically, if only one of the twelve endpoints was reported, the superiority of one learning strategy can be spuriously concluded. Having more endpoints/datasets provides a stronger base for a reasonable conclusion²¹. Utilization of more than one endpoint in studies assessing/implementing different learning strategies in the field of radiotherapy is rare.

It was shown that logistic regression and MARS are most likely to be the best-performing learning strategies for the prediction of urinary symptoms. The conclusion was based on model fitting to a data set comprised of a large number of patients with different event rates for the three definitions of dysuria, haematuria, urinary incontinence and frequency. The superiority of a relatively simple and widely used method like logistic regression is in contrast to many observations from other studies in radiotherapy outcome modelling. There are also other instances where more recently-developed statistical-learning strategies failed to substantially improve prediction accuracy over simpler methods^{47,48}. Logistic regression and MARS along with elastic-net provide clear and reasonably unambiguous models. They use well-established probabilistic frameworks. Estimates describing each of the factors included in the final model provide high interpretability and straightforward inference for the important features which may be used as the basis for decision making.

Random forest, which have been shown in many occasions to be a superior classification strategy^{49, 50} including in the context of predictive modelling in radiotherapy,¹² only performed modestly in this analysis. One of the most recent examples was the use of random forest in the prediction of rectal bleeding. The authors reported significant improvement of random forest model over Lyman-Kutcher-Burman (LKB) model and logistic regression despite a modest cross-validated AUROC difference (validation AUROC: 0.68 vs 0.64 and 0.65, respectively)¹². Due to the relatively small difference between a random forest model and logistic regression, it might be argued that the superiority of random forest could be

endpoint-specific or found by chance. Both support-vector machine and neural network are more commonly used learning strategies in the prediction of radiotherapy toxicity outcome. Neural network is probably one of the first contemporary learning strategy explored for its potential in radiotherapy toxicity-prediction models¹. Since then, many studies have successfully utilized both neural network and support-vector machine in other toxicity-prediction problems^{2-5, 7}. Despite the significantly-lower predictive-power for support-vector machine and neural network, it is acknowledged that only the ‘vanilla’ variants were implemented in this study. There are improvements of the methods suggested in the literature that were not implemented here which may have negatively impacted the result, unfairly discrediting the strategies⁵¹. Thus, it may not be appropriate to conclude the overall inferiority of the strategies but rather only the specific variants used in this study. However, the decision to not include many variants of support-vector machine and neural network was based on the assumption that the basic algorithms should suffice to show improvements, if any, to a more conventional strategy like logistic regression.

Comparisons of different learning strategies have previously been conducted by others using field-specific datasets, many of which have shown significantly-better predictive power than the more conventional alternatives. Boulesteix *et al.*⁵² have argued that researchers were inclined to be over-optimistic with respect to the advantages of the new strategies they are reporting. This can be in the form of over-optimization of the reported strategy’s parameters. Thus, the superiority of contemporary learning-strategies should be considered with caution. As there are hundreds of different strategies available for researchers’ use, either conceptually-unique or variants of established learning-strategies, the use of these strategies without a reasonable justification may potentially cause over-fitting where the strategy in question may perform well in the specific dataset at hand while it may not be generalizable to other datasets.

From the above results, there is a compulsion to discourage the use of more complex modelling strategies. However, specific problems may benefit from specific statistical-learning strategies on the basis of the properties of the associated information – such as better treatment of missing data, computational efficiency, and the presence of sparse events or sparse features which may benefit from one specific strategy. For example, compared to logistic regression, random forest does not require feature selection and requires very minimal tuning/optimization.

Even though the improvements afforded from the implementation of different statistical-learning strategies were seen in certain endpoints, the overall predictive power remained modest which may limit the applicability of the models. Several endpoints studied in this analysis showed a substantially better model performance regardless of the statistical learning strategies used. These endpoints have higher performance correlation between models and more common important features suggesting that these endpoints have less complex relationships with strong features in the feature pools. These common important features may be used as a basis for integration into clinical decisions. Some endpoints have poor performance regardless of the strategy used suggesting that features included in the feature pools did not produce satisfactory predictive power even with extensive dosimetric, comorbidities and medication-intake information available. There are several explanations for the low predictive-power achieved for the derived models; first, predictors for urinary

toxicity following radiotherapy are known to be elusive partially due to bladder volume variability and other unresolved issues⁵³. In addition to the difficulty in assessing urinary symptoms specifically related to radiotherapy in the context of elderly patients with otherwise-increasing rates of urinary symptoms and symptoms related to prostatic hyperplasia, the pathophysiological processes involved are still not well understood. Second, despite the attempt to generate synthetic instances in this analysis, the paucity of events for certain endpoints may require a more specific learning strategy specifically designed for that purpose including RUSboost⁵⁴.

The challenge to improve predictive models for urinary symptoms remains open. The availability of strong features is key in the construction of a predictive model. There are suggestions that specific dose-surface maps, which take into consideration the spatial information of the dose, have stronger relationships to post-treatment effects than dosimetric indices derived from the dose-surface histogram alone⁵⁵. The inclusion of the maps in the potential predictor pool may provide the predictive power necessary for a clinically-useful model.

Conclusion

In the context of urinary-symptom predictions, logistic regression and MARS were most likely to be the best-performing strategies for the prediction of urinary symptoms. The predictive power was modest – models are in need of new features including spatial descriptions of the dose distribution to achieve better predictive capability.

ACKNOWLEDGEMENTS

We acknowledge funding from Cancer Australia and the Diagnostics and Technology Branch of the Australian Government Department of Health and Ageing (grant 501106), the National Health and Medical Research Council (grants 300705, 455521, 1006447, 1077788), the Hunter Medical Research Institute, the Health Research Council (New Zealand), the University of Newcastle and the Calvary Mater Newcastle, Abbott Laboratories, Novartis Pharmaceuticals. We gratefully acknowledge the support of the participating RADAR centres and the Trans-Tasman Radiation Oncology Group. We are grateful to Michelle Krawiec and Sharon Richardson for delineating bladder structures on trial datasets.

REFERENCES

- ¹ S.L. Gulliford, S. Webb, C.G. Rowbottom, D.W. Corne, D.P. Dearnaley, "Use of artificial neural networks to predict biological outcomes for patients receiving radical radiotherapy of the prostate," *Radiother Oncol* **71**, 3-12 (2004).
- ² A. Pella, R. Cambria, M. Riboldi, B.A. Jereczek-Fossa, C. Fodor, D. Zerini, A.E. Torshabi, F. Cattani, C. Garibaldi, G. Pedroli, G. Baroni, R. Orecchia, "Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy," *Med Phys* **38**, 2859-2867 (2011).
- ³ S. Tomatis, T. Rancati, C. Fiorino, V. Vavassori, G. Fellin, E. Cagna, F.A. Mauro, G. Girelli, A. Monti, M. Baccolini, G. Naldi, C. Bianchi, L. Menegotti, M. Pasquino, M.

Stasi, R. Valdagni, "Late rectal bleeding after 3D-CRT for prostate cancer: development of a neural-network-based predictive model," *Phys Med Biol* **57**, 1399-1412 (2012).

⁴ F. Buettner, S.L. Gulliford, S. Webb, M. Partridge, "Using dose-surface maps to predict radiation-induced rectal bleeding: a neural network approach," *Phys Med Biol* **54**, 5139-5153 (2009).

⁵ I. El Naqa, J.D. Bradley, P.E. Lindsay, A.J. Hope, J.O. Deasy, "Predicting radiotherapy outcomes using statistical learning techniques," *Phys Med Biol* **54**, S9-S30 (2009).

⁶ S. Chen, S. Zhou, J. Zhang, F.F. Yin, L.B. Marks, S.K. Das, "A neural network model to predict lung radiation-induced pneumonitis," *Med Phys* **34**, 3420-3427 (2007).

⁷ S. Chen, S. Zhou, F.F. Yin, L.B. Marks, S.K. Das, "Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis," *Med Phys* **34**, 3808-3814 (2007).

⁸ S.K. Das, S.M. Zhou, J. Zhang, F.F. Yin, M.W. Dewhirst, L.B. Marks, "Predicting lung radiotherapy-induced pneumonitis using a model combining parametric lyman probit with nonparametric decision trees," *Int J Radiat Oncol Biol Phys* **68**, 1212-1221 (2007).

⁹ S.F. Chen, S.M. Zhou, F.F. Yin, L.B. Marks, S.K. Das, "Using patient data similarities to predict radiation pneumonitis via a self-organizing map," *Phys Med Biol* **53**, 203-216 (2008).

¹⁰ S.K. Das, S.F. Chen, J.O. Deasy, S.M. Zhou, F.F. Yin, L.B. Marks, "Combining multiple models to generate consensus: Application to radiation-induced pneumonitis prediction," *Med Phys* **35**, 5098-5109 (2008).

¹¹ T.W. Schiller, Y.X. Chen, I. El Naga, J.O. Deasy, "Modeling radiation-induced lung injury risk with an ensemble of support vector machines," *Neurocomputing* **73**, 1861-1867 (2010).

¹² J.D. Ospina, J. Zhu, C. Chira, A. Bossi, J.B. Delobel, V. Beckendorf, B. Dubray, J.L. Lagrange, J.C. Correa, A. Simon, O. Acosta, R. de Crevoisier, "Random forests to predict rectal toxicity following prostate cancer radiation therapy," *Int J Radiat Oncol Biol Phys* **89**, 1024-1031 (2014).

¹³ O. Gayou, S.K. Das, S.M. Zhou, L.B. Marks, D.S. Parda, M. Miften, "A genetic algorithm for variable selection in logistic regression analysis of radiotherapy treatment outcomes," *Med Phys* **35**, 5426-5433 (2008).

¹⁴ K. Wopken, H.P. Bijl, A. van der Schaaf, M.E. Christianen, O. Chouvalova, S.F. Oosting, B.F. van der Laan, J.L. Roodenburg, C.R. Leemans, B.J. Slotman, P. Doornaert, R.J. Steenbakkers, I.M. Verdonck-de Leeuw, J.A. Langendijk, "Development and validation of a prediction model for tube feeding dependence after curative (chemo-) radiation in head and neck cancer," *PLoS One* **9**, e94879 (2014).

¹⁵ T.F. Lee, M.H. Liou, Y.J. Huang, P.J. Chao, H.M. Ting, H.Y. Lee, F.M. Fang, "LASSO NTCP predictors for the incidence of xerostomia in patients with head and neck squamous cell carcinoma and nasopharyngeal carcinoma," *Scientific reports* **4**, 6217 (2014).

¹⁶ P. Lambin, R.G.P.M. van Stiphout, M.H.W. Starmans, E. Rios-Velazquez, G. Nalbantov, H.J.W.L. Aerts, E. Roelofs, W. van Elmpt, P.C. Boutros, P. Granone, V. Valentini, A.C. Begg, D. De Ruyscher, A. Dekker, "Predicting outcomes in radiation

oncology-multifactorial decision support systems," *Nat Rev Clin Oncol* **10**, 27-40 (2013).

17 M.J. Zelefsky, E.J. Levin, M. Hunt, Y. Yamada, A.M. Shippy, A. Jackson, H.I. Amols, "Incidence of late rectal and urinary toxicities after three-dimensional conformal radiotherapy and intensity-modulated radiotherapy for localized prostate cancer," *Int J Radiat Oncol Biol Phys* **70**, 1124-1129 (2008).

18 M.J. Zelefsky, M. Kollmeier, B. Cox, A. Fidaleo, D. Sperling, X. Pei, B. Carver, J. Coleman, M. Lovelock, M. Hunt, "Improved Clinical Outcomes With High-Dose Image Guided Radiotherapy Compared With Non-IGRT for the Treatment of Clinically Localized Prostate Cancer," *Int J Radiat Oncol Biol Phys* **84**, 125-129 (2012).

19 H. Yamazaki, S. Nakamura, T. Nishimura, K. Yoshida, Y. Yoshioka, M. Koizumi, K. Ogawa, "Transitioning from conventional radiotherapy to intensity-modulated radiotherapy for localized prostate cancer: changing focus from rectal bleeding to detailed quality of life analysis," *J Radiat Res* **55**, 1033-1047 (2014).

20 T.F. Lee, P.J. Chao, H.M. Ting, L.Y. Chang, Y.J. Huang, J.M. Wu, H.Y. Wang, M.F. Horng, C.M. Chang, J.H. Lan, Y.Y. Huang, F.M. Fang, S.W. Leung, "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," *PLoS One* **9**(2014).

21 A.L. Boulesteix, "Ten simple rules for reducing overoptimistic reporting in methodological computational research," *PLoS Comput Biol* **11**, e1004191 (2015).

22 J.W. Denham, C. Wilcox, D. Joseph, N.A. Spry, D.S. Lamb, K.-H. Tai, J. Matthews, C. Atkinson, S. Turner, D. Christie, N.K. Gogna, L. Kenny, G. Duchesne, B. Delahunt, P. McElduff, "Quality of life in men with locally advanced prostate cancer treated with leuporelin and radiotherapy with or without zoledronic acid (TROG 03.04 RADAR): secondary endpoints from a randomised phase 3 factorial trial," *Lancet Oncol* **13**, 1260-1270 (2012).

23 J.W. Denham, D. Joseph, D.S. Lamb, N.A. Spry, G. Duchesne, J. Matthews, C. Atkinson, K.H. Tai, D. Christie, L. Kenny, S. Turner, N.K. Gogna, T. Diamond, B. Delahunt, C. Oldmeadow, J. Attia, A. Steigler, "Short-term androgen suppression and radiotherapy versus intermediate-term androgen suppression and radiotherapy, with or without zoledronic acid, in men with locally advanced prostate cancer (TROG 03.04 RADAR): an open-label, randomised, phase 3 factorial trial," *Lancet Oncol* **15**, 1076-1089 (2014).

24 J.W. Denham, A. Steigler, D. Joseph, D.S. Lamb, N.A. Spry, G. Duchesne, C. Atkinson, J. Matthews, S. Turner, L. Kenny, K.H. Tai, N.K. Gogna, S. Gill, H. Tan, R. Kearvell, J. Murray, M. Ebert, A. Haworth, A. Kennedy, B. Delahunt, C. Oldmeadow, E.G. Holliday, J. Attia, "Radiation dose escalation or longer androgen suppression for locally advanced prostate cancer? Data from the TROG 03.04 RADAR trial," *Radiother Oncol* **115**, 301-307 (2015).

25 J.W. Denham, C. Wilcox, D.S. Lamb, N.A. Spry, G. Duchesne, C. Atkinson, J. Matthews, S. Turner, L. Kenny, K.H. Tai, N.K. Gogna, M. Ebert, B. Delahunt, P. McElduff, D. Joseph, "Rectal and urinary dysfunction in the TROG 03.04 RADAR trial for locally advanced prostate cancer," *Radiother Oncol* **105**, 184-192 (2012).

26 A. Haworth, R. Kearvell, P.B. Greer, B. Hooton, J.W. Denham, D. Lamb, G. Duchesne, J. Murray, D. Joseph, "Assuring high quality treatment delivery in clinical

- 621 trials - Results from the Trans-Tasman Radiation Oncology Group (TROG) study
622 03.04 "RADAR" set-up accuracy study," *Radiother Oncol* **90**, 299-306 (2009).
- 623 27 R. Kearvell, A. Haworth, M.A. Ebert, J. Murray, B. Hooton, S. Richardson, D.J.
624 Joseph, D. Lamb, N.A. Spry, G. Duchesne, J.W. Denham, "Quality improvements in
625 prostate radiotherapy: Outcomes and impact of comprehensive quality assurance
626 during the TROG 03.04 'RADAR' trial," *J Med Imag Radiat On* **57**, 247-257 (2013).
- 627 28 "Lent soma scales for all anatomic sites," *Int J Radiat Oncol Biol Phys* **31**, 1049-1091
628 (1995).
- 629 29 M.A. Ebert, A. Haworth, R. Kearvell, B. Hooton, R. Coleman, N. Spry, S. Bydder, D.
630 Joseph, "Detailed review and analysis of complex radiotherapy clinical trial planning
631 data: evaluation and initial experience with the SWAN software system," *Radiother*
632 *Oncol* **86**, 200-210 (2008).
- 633 30 A.N. Viswanathan, E.D. Yorke, L.B. Marks, P.J. Eifel, W.U. Shipley, "Radiation
634 dose-volume effects of the urinary bladder," *Int J Radiat Oncol Biol Phys* **76**, S116-
635 122 (2010).
- 636 31 S.M. Bentzen, W. Dorr, R. Gahbauer, R.W. Howell, M.C. Joiner, B. Jones, D.T.
637 Jones, A.J. van der Kogel, A. Wambersie, G. Whitmore, "Bioeffect modeling and
638 equieffective dose concepts in radiation oncology--terminology, quantities and units,"
639 *Radiother Oncol* **105**, 266-268 (2012).
- 640 32 A. Niemierko, "Reporting and analyzing dose distributions: A concept of equivalent
641 uniform dose," *Med Phys* **24**, 103-110 (1997).
- 642 33 N. Yahya, M.A. Ebert, M. Bulsara, A. Haworth, A. Kennedy, D.J. Joseph, J.W.
643 Denham, "Dosimetry, clinical factors and medication intake influencing urinary
644 symptoms after prostate radiotherapy: An analysis of data from the RADAR prostate
645 radiotherapy trial," *Radiother Oncol* **116**, 112-118 (2015).
- 646 34 N. Yahya, M.A. Ebert, M. Bulsara, M.J. House, A. Kennedy, D.J. Joseph, J.W.
647 Denham, "Urinary symptoms following external beam radiotherapy of the prostate:
648 Dose-symptom correlates with multiple-event and event-count models," *Radiother*
649 *Oncol* **117**, 277-282 (2015).
- 650 35 H. Zou, T. Hastie, "Regularization and variable selection via the elastic net," *J Roy*
651 *Stat Soc B* **67**, 301-320 (2005).
- 652 36 L. Breiman, "Random Forests," *Mach Learn* **45**, 5-32 (2001).
- 653 37 W.N. Venables, B.D. Ripley, *Modern applied statistics with S*, 4th ed. (Springer, New
654 York, 2002).
- 655 38 J.H. Friedman, "Multivariate Adaptive Regression Splines," *Ann Stat* **19**, 1-67 (1991).
- 656 39 C.J. Xu, A. van der Schaaf, A.A. Van't Veld, J.A. Langendijk, C. Schilstra,
657 "Statistical validation of normal tissue complication probability models," *Int J Radiat*
658 *Oncol Biol Phys* **84**, e123-129 (2012).
- 659 40 N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, "SMOTE: Synthetic
660 minority over-sampling technique," *J Artif Intell Res* **16**, 321-357 (2002).
- 661 41 W.N. Venables, B.D. Ripley, W.N. Venables, *Modern applied statistics with S*, 4th
662 ed. (Springer, New York, 2002).
- 663 42 M. Kuhn, "Building Predictive Models in R Using the caret Package," *J Stat Softw*
664 **28**(2008).
- 665 43 R Development Core Team, "R: a language and environment for statistical
666 computing. 2013. R Foundation for Statistical Computing, Vienna, Austria," (ISBN
667 3-900051-07-0, 2013).

- 44 J. Friedman, T. Hastie, R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," J Stat Softw **33**, 1-22 (2010).
- 45 A. Karatzoglou, K. Hornik, A. Smola, A. Zeileis, "kernlab - An S4 package for kernel methods in R," J Stat Softw **11**, 1-20 (2004).
- 46 A. Liaw, M. Wiener, "Classification and Regression by randomForest," R News **2**, 18-22 (2002).
- 47 J.N. Cooper, L. Wei, S.A. Fernandez, P.C. Minneci, K.J. Deans, "Pre-operative prediction of surgical morbidity in children: comparison of five statistical models," Comput Biol Med **57**, 54-65 (2015).
- 48 P. Gao, X. Zhou, Z.N. Wang, Y.X. Song, L.L. Tong, Y.Y. Xu, Z.Y. Yue, H.M. Xu, "Which is a more accurate predictor in colorectal survival analysis? Nine data mining algorithms vs. the TNM staging system," PLoS One **7**, e42015 (2012).
- 49 T.G. Dietterich, "Ensemble methods in machine learning," Lect Notes Comput Sc **1857**, 1-15 (2000).
- 50 M. Fernandez-Delgado, E. Cernadas, S. Barro, D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," J Mach Learn Res **15**, 3133-3181 (2014).
- 51 T. Hastie, R. Tibshirani, J.H. Friedman, *The elements of statistical learning : data mining, inference, and prediction*, 2nd ed. (Springer, New York, NY, 2009).
- 52 A.L. Boulesteix, S. Lauer, M.J. Eugster, "A plea for neutral comparison studies in computational sciences," PLoS One **8**, e61562 (2013).
- 53 T. Rosewall, C. Catton, G. Currie, A. Bayley, P. Chung, J. Wheat, M. Milosevic, "The relationship between external beam radiotherapy dose and chronic urinary dysfunction--a methodological critique," Radiother Oncol **97**, 40-47 (2010).
- 54 C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano, "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on **40**, 185-197 (2010).
- 55 F. Palorini, C. Cozzarini, S. Gianolini, T. Rancati, V. Carillo, A. Botti, C. Iotti, R. Valdagni, C. Fiorino, "Bladder Dose-Surface Maps show Evidence of Spatial Effects for the Risk of Acute Urinary Toxicity after Moderate Hypofractionated Radiation for Prostate Cancer," International Journal of Radiation Oncology Biology Physics **90**, S42-S43 (2014).

Appendix A: EUD and clustering

$$\text{EUD}_a = \left(\sum_i s_i D_i^a \right)^{a^{-1}}$$

where s_i is the fractional area of the dose bin (dose bin = 0.1 Gy) corresponding to dose D_i in the differential DSH of the bladder and the exponent a (range: $a \in [1 \dots 100]$) is a unitless parameter associated with the specific behaviour of the endpoint. $a=1$ denotes the mean dose where associated endpoints are strongly dependent on dose received by the whole organ while large exponent a denotes a stronger maximum dose impact to a smaller portion of the organ. Variable clustering is used for assessing collinearity, redundancy, and for separating variables into clusters that can be scored as a single variable, thus resulting in data reduction. Similarity measure based on Spearman correlations ($\rho = 0.8$).

714

715

Cluster	EUD	Representative EUD
1	1-3	1
2	4-7	4
3	8-15	8
4	16-23	16
5	24-100	24

716

Appendix B: Distribution of clinical features. Continuous distributions are specified as mean $\pm$ standard deviation (range), categorical variables are specified as number of patients (%). Missing data were replaced with the median for continuous and mod for categorical features.

Factors			Missing
Physical & Trial factors			
	Age	69 $\pm$ 7(49-85) years	3
	BMI	27.98 $\pm$ 4.12 (17.17-45.77) kg/m ²	22
	ECOG Performance Status (=1)	123 (16%)	1
Comorbidities			
	Cardiovascular condition	217 (29)	0
	Peripheral vascular condition	44 (6)	0
	Cerebrovascular condition	37 (5)	0
	Hypertension	353 (49)	1
	Dyslipidaemia	248 (33)	2
	NIDDM	92 (12)	2
	IDDM	14 (2)	0
	Respiratory disorder	99 (13)	0
	Bowel disorder	91 (12)	1
	Dermatological disorder	52 (7)	1
	Collagen disorder	15 (2)	1
	Bone or calcium metabolism disorder	66 (9)	1
	Haematological disorder	11 (1)	1
	Thyroid disorder	24 (3)	1
Medication intake			
	Insulin	14 (2)	6
	Hypoglycaemic agents	55 (7)	7
	ACE Inhibitor	240 (32.1)	8
	Statin	221 (29.6)	8
	Steroids	24 (3)	8
	NSAID	136 (18.2)	6
	Anti-coagulant	120 (16.0)	6
	Antioxidants, flavonoids, phyto-oestrogens or selenium	25 (3)	17
Lifestyle factors			
	Smoking status	Never 274 (36); Previous 380 (50); Current 99 (13)	1
	Alcohol intake	None 100 (13); Occasional 279 (37); Regular 370 (49)	5

Abbreviations; OR- Odds ratio; BMI - body mass index; ECOG - ECOG Performance Status; NIDDM – non-insulin dependent diabetes mellitus; IDDM – insulin dependent diabetes

724 mellitus; ACE - angiotensin-converting-enzyme; NSAID – non-steroidal anti-inflammatory
725 drugs
726
727 Appendix C: Feature importance based on statistical learning strategies. Features consistently
728 selected in the top 5 for all statistical learning strategies are bold.
729

	Logistic	MARS	Elastic-net	Neural network	Random forest	Support vector machine
Dysuria						
Grade 1						
1	Bowel dis.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.
2	Smoking status	Bowel dis.	Bowel dis.	EUD4	Bowel dis.	EUD4
3	BMI	BMI	NSAIDs	EUD8	ECOG	EUD8
4	Pre-treat. symp.	NSAIDs	ECOG	Bowel dis.	EUD1	Bowel dis.
5	NSAIDs	Smoking status	EUD4	BMI	NSAIDs	BMI
Grade 2						
1	Pre-treat. symp.	Pre-treat. symp.	Bowel dis.	Pre-treat. symp.	BMI	Pre-treat. symp.
2	dyslipidaemia	Smoking status	Pre-treat. symp.	Bowel dis.	Age	dyslipidaemia
3	Bowel dis.	Bowel dis.	Smoking status	Smoking status	EUD1	BMI
4	BMI	BMI	dyslipidaemia	D75 Gy	PTV	Bowel dis.
5	Smoking status	EUD1	BMI	BMI	EUD4	Smoking status
Longitudinal						
1	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.
2	BMI	Cerebrovas. cond	Cerebrovas. cond.	BMI	Cerebrovas. cond.	BMI
3	ACE-inhibitor	EUD4	Bowel dis.	Cerebrovas. cond.	EUD1	EUD8
4	EUD8	EUD1	Smoking status	Smoking status	EUD16	EUD16
5	EUD16	Bowel dis.	ACE-inhibitor	Cardiovas. cond.	EUD8	ACE-inhibitor
Haematuria						
Grade 1						
1	PTV	EUD1	Bowel dis.	Age	EUD4	PTV
2	Statin intake	Bowel dis.	Statin intake	Bowel dis.	EUD1	Statin intake
3	BMI	Statin intake	EUD4	medcoag	EUD8	Age
4	EUD1	Age	EUD1	Cerebrovas. cond.	EUD16	BMI
5	Age	EUD24	PTV	PTV	NIDDM	dyslipidaemia
Grade 2						
1	EUD24	EUD24	BMI	BMI	BMI	EUD24
2	BMI	EUD16	PTV	PTV	PTV	BMI
3	PTV	EUD8	Statin intake	Statin intake	EUD24	Statin intake
4	Statin intake	EUD4	Bowel dis.	Age	EUD16	PTV
5	EUD16	BMI	periyn	EUD24	EUD4	EUD16
Longitudinal						
1	dyslipidaemia	EUD16	dyslipidaemia	D75 Gy	BMI	dyslipidaemia
2	Statin intake	EUD24	BMI	Smoking status	EUD24	Statin intake
3	BMI	BMI	Statin intake	dyslipidaemia	Age	BMI
4	EUD16	dyslipidaemia	Respirat. cond.	Hypertension	EUD16	Age
5	Age	EUD8	Smoking status	ECOG	PTV	EUD16
Incontinence						
Grade 1						
1	Age	Age	Pre-treat. symp.	Pre-treat. symp.	Age	ECOG
2	ECOG	Pre-treat. symp.	Cerebrovas. cond.	Cardiovas. cond.	Pre-treat. symp.	Age
3	BMI	Cerebrovas. cond.	ECOG	Cerebrovas. cond.	Cerebrovas. cond.	BMI
4	Pre-treat. symp.	Hypertension	BMI	EUD1	Hypertension	Pre-treat. symp.
5	dyslipidaemia	ACE-inhibitor	Age	BMI	BMI	dyslipidaemia
Grade 2						
1	Age	Pre-treat. symp.	Pre-treat. symp.	Cerebrovas. cond.	Pre-treat. symp.	Age
2	Pre-treat. symp.	Cerebrovas. cond.	Cerebrovas. cond.	Pre-treat. symp.	EUD24	Pre-treat. symp.
3	Cerebrovas. cond.	EUD24	Alcohol intake	EUD24	Cerebrovas. cond.	Cerebrovas. cond.
4	Hypertension	Age	Age	Age	EUD16	Hypertension
5	ACE-inhibitor	EUD1	Hypertension	D75 Gy	PTV	ACE-inhibitor
Longitudinal						
1	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	BMI
2	BMI	BMI	BMI	Cerebrovas. cond.	EUD8	Pre-treat. symp.
3	Age	Age	Bone dis.	Age	BMI	Age
4	ACE-inhibitor	EUD8	Cerebrovas. cond.	BMI	Cerebrovas. cond.	ACE-inhibitor
5	Respirat. cond.	Cerebrovas. cond.	ACE-inhibitor	EUD1	EUD1	Respirat. cond.
Frequency						
Grade 1						
1	Risk category	Age	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.
2	Age	Pre-treat. symp.	Age	Age	ECOG	Age
3	Pre-treat. symp.	EUD4	Risk category	Risk category	Age	Risk category
4	Bone dis.	EUD8	Dermat. dis.	Bone dis.	Risk category	PTV
5	Statin intake	EUD1	Bone dis.	Statin intake	EUD1	NSAIDs
Grade 2						
1	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.
2	Age	Age	ECOG	Age	Age	Age
3	EUD4	BMI	Age	ECOG	BMI	EUD4

	4	EUD1	EUD1	EUD4	EUD1	ECOG	EUD1
	5	EUD8	EUD16	EUD1	Hypoglycaemic agents	Hypoglycaemic agents	EUD8
Longitudinal							
	1	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.	Pre-treat. symp.
	2	EUD1	Age	Risk category	Risk category	EUD4	EUD1
	3	EUD4	EUD4	Age	Age	EUD1	Age
	4	Age	EUD1	EUD1	EUD4	PTV	Risk category
	5	Risk category	EUD8	ECOG	dyslipidaemia	Age	EUD4