Independent external validation of predictive models for urinary dysfunction following external beam radiotherapy of the prostate: issues in model development and reporting

Noorazrul Yahya[1,2]
Martin A Ebert [1,3]
Max Bulsara [4]
Angel Kennedy[3]
David J Joseph[3,5]
James W Denham[6]

[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
[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 words (Abstract): 195
Number of words (Main text): 2782
Number of tables: 4
Number of figures: 1

Correspondence address:
Noorazrul Yahya
School of Physics
University of Western Australia
Stirling Hwy
Crawley, Western Australia 6009
Tel: +618 9346 4931

Email: noorazrul.yahya@research.uwa.edu.au
azrulyahya@ukm.edu.my

Running Title: Independent validation of predictive models

Keywords: independent external validation, predictive model, prostate radiotherapy, normal tissue complications, urinary symptoms

Conflicts of Interest Notification

Investigators declare no conflict of interest.

ABSTRACT

Background and Purpose: Most predictive models are not sufficiently validated for prospective use. We performed independent external validation of published predictive models for urinary dysfunctions following radiotherapy of the prostate.

Materials/Methods: Multivariable models developed to predict atomised and generalised urinary symptoms, both acute and late, were considered for validation using a dataset representing 754 participants from the TROG 03.04-RADAR trial. Endpoints and features were harmonised to match the predictive models. The overall performance, calibration and discrimination were assessed.

Results: 14 models from four publications were validated. The discrimination of the predictive models in an independent external validation cohort, measured using the receiver operating characteristic (ROC) curve, ranged from 0.473 to 0.695, generally lower than in internal validation. 4 models had ROC >0.6. Shrinkage was required for all predictive models' coefficients ranging from -0.309 (prediction probability was inverse to observed proportion) to 0.823. Predictive models which include baseline symptoms as a feature produced the highest discrimination. Two models produced a predicted probability of 0 and 1 for all patients.

Conclusions: Predictive models vary in performance and transferability illustrating the need for improvements in model development and reporting. Several models showed reasonable potential but efforts should be increased to improve performance. Baseline symptoms should always be considered as potential features for predictive models.

INTRODUCTION

Predictive models can be useful guides in clinical decision making, either diagnostic or prognostic, and have been utilised in many medical domains. For radiotherapy treatment, predictive models can estimate the risk of developing a particular dysfunction. On the basis of such predictions, adjustments can be made to treatment plans to minimise risk, preventive strategies can be optimally selected and patients may have the ability to participate in the decision making process. Recently, there has been a transition from traditional explanatory research to predictive modelling research. Such a transition can provide a clearer route to clinical adaptation including through multifactorial decision support systems [1, 2]. Viswanathan et al. in the Quantitative Analysis of Normal Tissue Effects in the Clinic (QUANTEC) report relevant to urinary dysfunction, have noted a paucity of quantitative models [3]. Since the report in 2010, several predictive models that have been developed.

In many instances, derived models have been internally validated, usually through bootstrapping or cross-validation algorithms. This process helps to provide a more accurate estimate of model performance if used prospectively [4]. Despite the assurance, internal validation is limited by similarities, such as in terms of treatment preferences, in the development cohort which may result in overoptimism of model performance. Validation using datasets external to the one used in the development process would allow the reproducibility and exportability of the models to be evaluated. Often, the external validation was performed by the same group who developed the models and usually the models were developed and externally validated in the same study (e.g. [5-7]). This development-validation sequence has a major advantage in providing a more accurate estimate of the actual performance of the models than by internal validation and in ensuring both the development and validation cohorts are completely harmonised. However, this sequence suggests that the modellers were not blinded to the validation datasets which may lead to certain biases. For example, it is possible for the modellers to overfit the feature selection process by cross-checking the resultant external validation performance. To reduce the potential bias, an independent external validation is needed.

In this analysis, we performed an independent external validation of predictive models available in the literature focusing on urinary dysfunctions following external beam radiotherapy of the prostate. Data from patients accrued to the Trans-Tasman Radiation Oncology Group (TROG) 03.04 trial of Randomised Androgen Deprivation and Radiotherapy (RADAR-NCT00193856) were utilised [8, 9]. The models were critically assessed and potential improvements that could be made in predictive model development and validation were then discussed based on this exercise.

MATERIALS AND METHODS

Urinary symptoms predictive models

The Scopus database was searched by use of the text words in the article title, abstract and keywords: bladder AND *urinary AND prostate AND radiotherapy AND predict* AND (toxicity OR symptom) on 5 Feb 2016. The search results were then limited to article only and in the field of medicine. The abstracts were reviewed by NY and MAE to search for predictive models for urinary symptoms following external beam radiotherapy of for prostate cancer.

The predictive models were used to assign the probability of symptoms in the validation cohort through the coefficient estimates provided in the publications. If the coefficient estimates were not provided in the report, authors were contacted to provide the information or the estimates were extracted from provided nomograms. Due to potential errors associated with translating graphical representation of the models, i.e. nomograms, into numbers, coefficient estimates were preferred. In a potentially erroneous report of coefficients, authors were contacted for confirmation.

Patients and treatments for validation cohort

754 participants received 3-dimensional conformal external beam radiotherapy (without a brachytherapy boost) to either 66, 70 or 74 Gy and had complete bladder dose data collected, comprising a digital treatment plan export including axial computed tomography (CT) images and associated planned dose matrix [8, 9]. Extensive dose features, clinical and treatment-related factors were collected during the RADAR trial. Associations of these factors to specific post-treatment symptoms of complications have been reported in previous publications [10-13]. Predictors for atomised urinary symptoms using dose, clinical and medication intake features have been previously discussed [12, 13].

Harmonisation of endpoints

Patients accrued during RADAR were assessed for urinary problems at baseline and at the end of radiotherapy using physician-assessed LENT-SOMA [14] and the International Prostate Symptom Score (IPSS) questionnaire. Patients were routinely followed up every three months for 18 months, then six-monthly up to five years and then annually where urinary symptoms were assessed using LENT-SOMA [14]. Patients were asked to complete the International Prostate Symptom Score (IPSS) questionnaire at 12, 18, 24, 36 and 60-month follow-up post-randomisation. Urinary symptom endpoints were extracted from the RADAR database matching the definition of endpoints found in the report of the predictive models. In instances where there were no similar endpoints collected from RADAR, equivalent endpoints were derived.

Harmonisation of features

The features used in each of the predictive models were matched to fields from the RADAR database. If similar features were not available, the closest equivalent features were selected. In instances where equivalent features were not available, alternative models reported in the studies were used. Only relevant features matching the ones used in the predictive models validated in this study will be reported.

Performance assessment

The overall performance of the predictive models was measured using the Brier score. The Brier score is the mean squared difference between actual and predicted outcome, which captures both discrimination and calibration aspects. The concordance statistic, which is identical to the area under the receiver operating characteristics (ROC) curve in a binary prediction problem, was used to assess the discriminative ability of the predictive models. A calibration plot, with the mean predicted probability on the *x*-axis and observed proportion on the *y*-axis, was plotted for each model. A perfect calibration should give a 45-degree line where the intercept is 0 and the calibration slope is 1. An intercept larger than 0 indicates that predictions are systematically too low and vice versa. A calibration slope of less than 1 indicates that the models were over-fitted and coefficient shrinkage is needed. For a more comprehensive explanation of these measures, Steyerberg et al. is recommended [15]. The validation was performed as implemented in *rms* (version 4.4-1) in R 3.2.3 (The R Foundation for Statistical Computing, Vienna, Austria) [16].

RESULTS

79 articles were found. Four articles [17-20] were selected after excluding other articles for at least one of these reasons; treated using brachytherapy (18) or protons (1), traditional explanatory studies (e.g. finding predictors, dosimetric constraints, comparisons between 3-dimensional conformal radiotherapy to intensity modulated radiotherapy) (42), using non-urinary endpoints (7), non-radiotherapy (5), machine learning study with no access to the final model (1) and our own (1). In total, 14 models were considered. Two of the studies produced predictive models for late urinary symptoms [17, 18] and another two for acute urinary symptoms [19, 20]. The studies and the associated models are listed in Table 1. The event rates were found to be higher in the validation cohort in most endpoints.

The endpoints for models from Mathieu et al. were based on the LENT-SOMA scale while models from Cozzarini et al. and Palorini et al were based on IPSS, both of which were directly comparable to the assessments used in the validation cohort [17, 19, 20]. De Langhe et al. used an in-house developed scoring system. The definition of haematuria was equivalent to the one used in the LENT-SOMA scale while the definition of nocturia was substituted using the increase of more than 2 points from baseline in question 7 of the IPSS questionnaire.

The distribution of features relevant to the models are listed in Table 2. The distribution of other features can be assessed from the original articles [17-20] and for the RADAR cohort are described in Supplementary Material A and in [10, 12, 13]. Gross target volume (GTV) and the minimum dose to the volume were utilised in place of clinical target volume (CTV) in model E because clinical target volume delineation was not mandatory within the RADAR protocol [21]. CTV and GTV are generally correlated making it a suitable equivalent. However, it is expected that the use of CTV may increase the predicted probability of toxicity in the validation cohort. The use of cardiovascular drugs and anti-hypertensive drugs in model J and M were substituted with cardiovascular disease and hypertension, respectively, due to the lack of details of the type of drugs in the report. The use of anti-hypercholesterolemia was substituted with statin use in the validation cohort. For models from de Langhe et al., the alternative non-SNP models were validated [18]. The use of 5-alpha-reductase inhibitors was not identified in the validation cohort for the validation of IPSS increase of ≥10. As such the alternative model with seven predictors reported in the Supplementary Materials of the publication, was utilised [20]. A comparison of the distribution of features between development and validation cohorts revealed several differences (Table 2). Of note, the validation cohort has all patients treated with hormonal therapy and several dose features consist of all zeros.

Model E predicted certainty of symptoms for all patients in the validation cohort. The clinical target volume (CTV) and minimum dose to the CTV in model E were found to have exceedingly high coefficients (Supplementary Material B) [18]. The predicted probability for a hypothetical patient with CTV and minimum dose to the CTV similar to the minimum in the original publication, expected to be very low, was also found to be 1. Authors did not respond to queries. Model N predicted the complete absence of symptoms in the validation cohort. All features in Model N (except hormonal therapy) have interaction terms to the feature quantifying the surface of the bladder receiving more than 12 Gy per week; a feature with all zeros for conventionally-fractionated treatment. All patients in the validation cohort received hormonal therapy [20].

Brier scores for models C and D were found to be the best (closest to 0). Based on the ROC curve, the performance of the predictive models was found to be not better or worse than random (i.e. ROC $\leq$.5) in several models (Supplementary Material C & Fig. 1). Model L was found to be the most discriminative with an ROC of 0.695. Model F, G, I and L have ROC>0.6. Based on the graphical assessment of calibration, the predicted probability and observed group proportion generally showed direct linearity for most models. Models B, C and D have slopes <0 suggesting prediction probabilities were inverse to the observed proportions. Models F, G, I, L and M have an intercept >0 suggesting that the models underpredict the symptoms in the validation cohort while the rest have an intercept <0 suggesting overprediction. All slopes were $\neq$1 (ranging from -0.309 (prediction probability was inverse to observed proportion) to 0.823) requiring a shrinkage if they are to be used prospectively.

DISCUSSION

This study provides independent external validation for recently-published predictive models and nomograms for urinary symptoms following external beam radiotherapy of the prostate. This is the first independent external validation for multivariate models predictive of radiotherapy-related symptoms. 14 models were considered in this study elucidating the state of predictive performance of these models in an independent dataset and also illustrating improvements that should be considered in future model development processes. The current study focuses on symptoms of complications related to a specific organ-at-risk. The outcome and observations from this study, however, are applicable to predictive modelling of symptoms in other organs.

Relative to the development cohorts, event rates in the validation cohort were found to be higher in most endpoints. Treatment practice has not been standardised across cohorts. Compared to the patients in the development models, all patients in the validation cohort were treated with hormonal therapy which has been repeatedly shown to increase urinary symptoms [22, 23]. Substantial numbers of patients were treated with prostatectomy [18] and with hypofractionation [19, 20] in the development cohorts, neither of which were used in the validation cohort. It may be attributable to cultural and socioeconomic factors which previously shown to impact symptom reporting [24]. Different cohorts from different geographical locations may also have population specific genetic variants that may alter individual sensitivity to radiation [25]. The higher rate of late events for the validation dataset (RADAR) may also be attributed to detailed and frequent follow-up procedures. However, these explanations are rather speculative requiring more extensive investigation, perhaps by pooling the individual patient data in multivariate analysis before definitive dominant factors can be suggested.

The models from Mathieu et al. were not internally validated. For other models, the resultant performance was lower in the independent external validation cohort than in internal validation for almost all models. This is not unexpected given the inherent underlying variations between the cohorts that were not accounted for in the development process including those attributable to cultural and socioeconomic features which, previously shown found to have impact symptom reporting [24]. Different cohorts from different geographical locations may also have population specific genetic variants that may alter individual sensitivity to radiation [25]. Treatment practice has not been standardised across the cohorts. For example, all patients in the validation cohort were treated with hormonal therapy. Substantial numbers of patients were treated with prostatectomy [18] and with hypofractionation [19, 20] in the development cohorts, neither of which were used in the validation cohort.

Fortunately, the performance measured using the ROC was above 0.600 in several of the models with a maximum of 0.695. These values are better than most models based on dose features alone with ROCs of 0.51-0.64 in cross-validation as recently found by Thor et al. [26]. However, it can be argued that the ROCs were overoptimistic given the preselection of dose features with the lowest *p*-values in univariate analysis of the whole population before including them in the multivariable cross-validation from which the ROCs were estimated [26]. Although none exceeded the common standard expected from a prognostic model (i.e. ROC>0.8), the models are expected to perform better than human assessment alone as shown in a study involving lung cancer patients [27].

The models from Cozzarini et al. produced the highest predictive performance in the validation cohort [19]. This can be attributed to the inclusion of baseline symptoms as one of the features, a characteristic that distinguishes the models from Cozzarini et al. to other models validated in this study. The impact of baseline symptoms on the formation of post-treatment symptoms havs been repeatedly shown [12, 28, 29]. Models from Palorini et al. and de Langhe et al. used relative symptoms by looking at the difference between the grade before and after the treatment [18, 20]. This can be suboptimal given the difference between grades is not always linear and patients with higher baseline grades are less likely to reach a given grade difference threshold due to symptom saturation [30]. These models may be improved by including the symptom baseline. Models from Mathieu et al. did not include the baseline symptoms possibly due to the lack of such information in the cohort [17].

In order for the models to be used as a tool for decision making, they need to be validated externally using an independent cohort and in prospective cohorts. To aid the process, modellers need to ensure the models can be easily applied to other cohorts through the use of reasonable features and unambiguous model descriptions [31]. The predictive model for nocturia from de Langhe et al. has unexpectedly high coefficients causing all patients in the validation cohort to have predicted probability of 1 [18]. It is suspected that the coefficients were erroneously reported. Cozzarini et al. and Palorini et al. produced models with dose factors described by “absolute bladder surface receiving more than (a certain dose)/week” [19, 20]. This was designed to take into account different dose fractionation for the patients in the cohort. The validation cohort and probably cohorts which may use the nomograms prospectively were/will be treated with conventional fractionation, thus, dose features above 10Gy/week are essentially zero. This may partially explain the discrepancies between the prediction probability based on the models of Cozzarini et al. and the observed proportion found in this validation study. Future studies aiming to develop predictive models should be encouraged to use a more standard form of dose description by converting it into equieffective dose in 2-Gy fractions [32]. Alternatively, models using features within the range of conventional fractionation correlated to the one used in the published models can be developed as a substitute. The applicability of the models can be further reduced with the use of interaction terms to these dose features, as seen in IPSS increase of ≥15 developed by Palorini et al., resulting in all zeros in the predicted probability [20]. It is common and advisable for models with interaction terms to also include the main effects which may remedy this problem of non-discriminative models [33].

In some instances, the clinical features were not identical in the development and validation cohorts, requiring harmonisation. Meldolesi et al. have described an enticing proposition of the use of an umbrella protocol for standardized data collection which may result in more consistent datasets [34]. The use of in-house grading systems, potentially not optimally translatable to the standard grading systems, may also hamper validation and prospective application of the models.

Many studies were not included in this analysis due to the explanatory nature of the studies making the translation into predictive probability in other datasets impossible (e.g. [28-30, 35-38]. These studies, however, carry substantial important information on the impact of clinical, treatment and dosimetric features on the formation of urinary symptoms. A literature-based meta-analysis synthesising studies related to urinary symptoms to determine features consistently predictive can be performed as successfully done for radiation-induced pneumonitis [39].

Despite limitations, several of the models developed have demonstrated reasonable predictive power in an independent external validation. There is, however, the need for the predictive power of the models to be improved. Apart from the improvements of model development and reporting as suggested above, there are other avenues being explored. Genetic features have been shown to be associated to post-treatment dysfunctions and the inclusion of these features in the models have been shown to improve predictive power [18, 40, 41], although, this is controversial as the impact of genetic features found to be predictive in some studies failed to be replicated in an independent validation [42]. The extraction of more sophisticated radiomics features including from spatial dose descriptors (e.g. dose maps) and dose features external to the conventional organ of interest for urinary symptoms (i.e. the bladder) should also be considered [29, 30, 43, 44]. Another obvious avenue of model improvement is through the pooling of data from different datasets as highlighted in a QUANTEC report [45] or through open source approaches (e.g. [46], cancerdata.org).

In conclusion, in this study we have provided an independent external validation of predictive models for urinary symptoms following external beam radiotherapy of the prostate. The models vary in performance and transferability illustrating the need for improvements in model development and reporting. Baseline urinary symptoms should always be considered in predictive models. We have provided evidence of the reasonable potential of these models but efforts should be increased to improve model performance.

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), 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. NY would like to acknowledge the support of the Australian government (Scholarship for International Research Fees) and UWA Convocation Travel Award. We are grateful to Michelle Krawiec and Sharon Richardson for delineating bladder structures on trial datasets and Feredica Palorini and Tiziana Rancati for providing further details of their models.

References

[1] van der Schaaf A, Langendijk JA, Fiorino C, Rancati T. Embracing phenomenological approaches to normal tissue complication probability modeling: a question of method. Int J Radiat Oncol Biol Phys. 2015;91:468-71.

[2] Lambin P, van Stiphout RG, Starmans MH, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol. 2013;10:27-40.

[3] Viswanathan AN, Yorke ED, Marks LB, Eifel PJ, Shipley WU. Radiation dose-volume effects of the urinary bladder. Int J Radiat Oncol Biol Phys. 2010;76:S116-22.

[4] Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. JClin Epidemiol. 2001;54:774-81.

[5] Dehing-Oberije C, De Ruysscher D, Petit S, et al. Development, external validation and clinical usefulness of a practical prediction model for radiation-induced dysphagia in lung cancer patients. Radiother Oncol. 2010;97:455-61.

[6] Nalbantov G, Kietselaer B, Vandecasteele K, et al. Cardiac comorbidity is an independent risk factor for radiation-induced lung toxicity in lung cancer patients. Radiother Oncol. 2013;109:100-6.

[7] Guerra JL, Gomez D, Wei Q, et al. Association between single nucleotide polymorphisms of the transforming growth factor beta1 gene and the risk of severe radiation esophagitis in patients with lung cancer. Radiother Oncol. 2012;105:299-304.

[8] Denham JW, Joseph D, Lamb DS, et al. 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. 2014;15:1076-89.

[9] Denham JW, Steigler A, Joseph D, et al. Radiation dose escalation or longer androgen suppression for locally advanced prostate cancer? Data from the TROG 03.04 RADAR trial. Radiother Oncol. 2015;115:301-7.

[10] Yahya N, Ebert MA, Bulsara M, et al. Impact of treatment planning and delivery factors on gastrointestinal toxicity: an analysis of data from the RADAR prostate radiotherapy trial. Radiat Oncol. 2014;9:282.

[11] Ebert MA, Foo K, Haworth A, et al. Gastrointestinal dose-histogram effects in the context of dose-volume-constrained prostate radiation therapy: analysis of data from the RADAR prostate radiation therapy trial. Int J Radiat Oncol Biol Phys. 2015;91:595-603.

[12] Yahya N, Ebert MA, Bulsara M, et al. Dosimetry, clinical factors and medication intake influencing urinary symptoms after prostate radiotherapy: An analysis of data from the RADAR prostate radiotherapy trial. Radiother Oncol. 2015;116:112-8.

[13] Yahya N, Ebert MA, Bulsara M, et al. Urinary symptoms following external beam radiotherapy of the prostate: Dose-symptom correlates with multiple-event and event-count models. Radiother Oncol. 2015;117:277-82.

[14] Lent soma scales for all anatomic sites. Int J Radiat Oncol Biol Phys. 1995;31:1049-91.

[15] Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128-38.
[16] R Development Core Team. R: a language and environment for statistical computing. 2013. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0; 2013.
[17] Mathieu R, Arango JD, Beckendorf V, et al. Nomograms to predict late urinary toxicity after prostate cancer radiotherapy. World J Urol. 2014;32:743-51.
[18] De Langhe S, De Meerleer G, De Ruyck K, et al. Integrated models for the prediction of late genitourinary complaints after high-dose intensity modulated radiotherapy for prostate cancer: making informed decisions. Radiother Oncol. 2014;112:95-9.
[19] Cozzarini C, Rancati T, Carillo V, et al. Multi-variable models predicting specific patient-reported acute urinary symptoms after radiotherapy for prostate cancer: Results of a cohort study. Radiother Oncol. 2015;116:185-91.
[20] Palorini F, Rancati T, Cozzarini C, et al. Multi-variable models of large International Prostate Symptom Score worsening at the end of therapy in prostate cancer radiotherapy. Radiother Oncol. 2016;118:92-8.
[21] Denham JW, Wilcox C, Lamb DS, et al. Rectal and urinary dysfunction in the TROG 03.04 RADAR trial for locally advanced prostate cancer. Radiother Oncol. 2012;105:184-92.
[22] Peeters ST, Hoogeman MS, Heemsbergen WD, et al. Volume and hormonal effects for acute side effects of rectum and bladder during conformal radiotherapy for prostate cancer. Int J Radiat Oncol Biol Phys. 2005;63:1142-52.
[23] Feigenberg SJ, Hanlon AL, Horwitz EM, Uzzo RG, Eisenberg D, Pollack A. Long-term androgen deprivation increases Grade 2 and higher late morbidity in prostate cancer patients treated with three-dimensional conformal radiation therapy. Int J Radiat Oncol Biol Phys. 2005;62:397-405.
[24] Galdas PM, Cheater F, Marshall P. Men and health help-seeking behaviour: literature review. J Adv Nurs. 2005;49:616-23.
[25] Andreassen CN, Alsner J. Genetic variants and normal tissue toxicity after radiotherapy: a systematic review. Radiother Oncol. 2009;92:299-309.
[26] Thor M, Olsson C, Oh JH, et al. Urinary bladder dose-response relationships for patient-reported genitourinary morbidity domains following prostate cancer radiotherapy. Radiother Oncol. 2016.
[27] Oberije C, Nalbantov G, Dekker A, et al. A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: a step toward individualized care and shared decision making. Radiother Oncol. 2014;112:37-43.
[28] Karlsdottir A, Muren LP, Wentzel-Larsen T, Dahl O. Late gastrointestinal morbidity after three-dimensional conformal radiation therapy for prostate cancer fades with time in contrast to genitourinary morbidity. Int J Radiat Oncol Biol Phys. 2008;70:1478-86.
[29] Heemsbergen WD, Al-Mamgani A, Witte MG, van Herk M, Pos FJ, Lebesque JV. Urinary obstruction in prostate cancer patients from the Dutch trial (68 Gy vs. 78 Gy): relationships with local dose, acute effects, and baseline characteristics. Int J Radiat Oncol Biol Phys. 2010;78:19-25.
[30] Ghadjar P, Zelefsky MJ, Spratt DE, et al. Impact of dose to the bladder trigone on long-term urinary function after high-dose intensity modulated radiation therapy for localized prostate cancer. Int J Radiat Oncol Biol Phys. 2014;88:339-44.
[31] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. JClin Epidemiol. 2015;68:134-43.

[32] Bentzen SM, Dorr W, Gahbauer R, et al. Bioeffect modeling and equieffective dose concepts in radiation oncology--terminology, quantities and units. Radiother Oncol. 2012;105:266-8.

[33] Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLoS Med. 2007;4:e297.

[34] Meldolesi E, van Soest J, Dinapoli N, et al. An umbrella protocol for standardized data collection (SDC) in rectal cancer: a prospective uniform naming and procedure convention to support personalized medicine. Radiother Oncol. 2014;112:59-62.

[35] Barnett GC, De Meerleer G, Gulliford SL, Sydes MR, Elliott RM, Dearnaley DP. The impact of clinical factors on the development of late radiation toxicity: results from the Medical Research Council RT01 trial (ISRCTN47772397). Clin Oncol (R Coll Radiol). 2011;23:613-24.

[36] Steinberger E, Kollmeier M, McBride S, Novak C, Pei X, Zelefsky MJ. Cigarette smoking during external beam radiation therapy for prostate cancer is associated with an increased risk of prostate cancer-specific mortality and treatment-related toxicity. BJU Int. 2015;116:596-603.

[37] Sutani S, Ohashi T, Sakayori M, et al. Comparison of genitourinary and gastrointestinal toxicity among four radiotherapy modalities for prostate cancer: Conventional radiotherapy, intensity-modulated radiotherapy, and permanent iodine-125 implantation with or without external beam radiotherapy. Radiother Oncol. 2015;117:270-6.

[38] Jain S, Loblaw DA, Morton GC, et al. The effect of radiation technique and bladder filling on the acute toxicity of pelvic radiotherapy for localized high risk prostate cancer. Radiother Oncol. 2012;105:193-7.

[39] Vogelius IR, Bentzen SM. A literature-based meta-analysis of clinical risk factors for development of radiation induced pneumonitis. Acta Oncol. 2012;51:975-83.

[40] Kerns SL, Stone NN, Stock RG, Rath L, Ostrer H, Rosenstein BS. A 2-stage genome-wide association study to identify single nucleotide polymorphisms associated with development of urinary symptoms after radiotherapy for prostate cancer. JUrol. 2013;190:102-8.

[41] Fachal L, Gomez-Caamano A, Barnett GC, et al. A three-stage genome-wide association study identifies a susceptibility locus for late radiotherapy toxicity at 2q24.1. Nat Genet. 2014;46:891-4.

[42] Barnett GC, Coles CE, Elliott RM, et al. Independent validation of genes and polymorphisms reported to be associated with radiation toxicity: a prospective analysis study. Lancet Oncol. 2012;13:65-77.

[43] Palorini F, Cozzarini C, Gianolini S, et al. First application of a pixel-wise analysis on bladder dose-surface maps in prostate cancer radiotherapy. Radiother Oncol. 2016.

[44] Yahya N, Ebert MA, Bulsara M, et al. 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. Med Phys. 2016;43:2040.

[45] Deasy JO, Bentzen SM, Jackson A, et al. Improving normal tissue complication probability models: the need to adopt a "data-pooling" culture. Int J Radiat Oncol Biol Phys. 2010;76:S151-4.

[46] Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26:1045-57.

[47] Denham JW, Wilcox C, Joseph D, et al. Quality of life in men with locally advanced prostate cancer treated with leuprorelin and radiotherapy with or without zoledronic acid

(TROG 03.04 RADAR): secondary endpoints from a randomised phase 3 factorial trial. Lancet Oncology. 2012;13:1260-70.

Captions

Table 1: List of considered predictive models for urinary symptoms following external beam radiotherapy of the prostate, the event rate in the development cohort and the corresponding event rate in the validation cohort.

Table 2: Comparison of distribution of features between the model development cohort and model validation cohort. Categorical features are specified as percentage; continuous features are specified to match the summary statistics used in the model development report.

Figure 1: Calibration curves for models A to M. Calibration of a model describes the extent to which the predicted probability matches the observed probability. Models E and N were not included (see text for details). The *x*-axis indicates the prediction obtained from the predictive models, and the *y*-axis indicates the observed proportions. Patients were divided into groups of 20 (represented by the triangles) based on the increasing predicted risk. Triangles can fall on the same spot resulting in smaller number of triangles shown. The line of unity, at 45 degrees, represents ideal agreement between observed and predicted probabilities. The vertical lines at the bottom indicate the distribution of the predicted probability. The solid line is the fitted logistic function and the dotted line is the non-parametric function.

Supplementary material A: Distributions of clinical features in the validation cohort. Continuous distributions are specified as mean ± standard deviation (range), categorical variables are specified as number of patients (%).

Supplementary Material B: Expressions for models D to N. Models A to C were based on the nomograms from the original publication [17].

Supplementary Material C: Model performance based on internal validation and independent external validation.

Table 1: List of considered predictive models for urinary symptoms following external beam radiotherapy of the prostate, the event rate in the development cohort and the corresponding event rate in the validation cohort.

Publication	Model	Event rate	
		Development	Validation
		Event/total (%)	
Mathieu et al. 2014 **[17]**			
	A. Global late urinary toxicity grade ≥2	183/965 (19)	272/746 (36)
	B. Late urinary frequency grade ≥2	92/965 (10)	216/746 (29)
	C. Late dysuria grade ≥2	36/965 (4)	53/746 (7)
De Langhe et al. 2014 **[18]**			
	D. Late haematuria 2+*	36/262 (14)	17/746 (2)†
	E. Late nocturia grade 2+*	29/264 (11)	166/748 (22)‡
Cozzarini et al. 2015 **[19]** §			
	F. Acute feeling of incomplete bladder emptying	18/231 (8)	104/678 (15)
	G. Acute frequency	35/220 (16)	204/653 (31)
	H. Acute intermittency	22/260 (8)	114/660 (17)
	I. Acute urgency	32/219 (15)	184/652 (28)
	J. Acute weak stream	44/221 (20)	157/614 (26)
	K. Acute straining	19/248 (8)	79/695 (11)
	L. Acute nocturia	42/229 (18)	260/643 (40)
Palorini et al. 2015 **[20]** ¶			
	M. Acute increase of IPSS score ≥10	77/380 (20)	255/718 (36)
	N. Acute increase of IPSS score ≥15	28/380 (7)	131/718 (18)

Note: IPSS – International Prostate Symptoms Score; * - ≥2 increase of grade; † - based on LENT-SOMA; ‡ - derived from IPSS questionnaire Question 7; § - patients with baseline ≥4 were removed from analysis; ¶ - patients with baseline ≥20 were removed from analysis

Table 2: Comparison of distribution of features between the model development and validation cohort. Categorical features are specified as percentage; continuous features are specified to match the summary statistics used in the model development report.

Model	Feature	Development	Validation
A	Anti-coagulant treatment	21	16

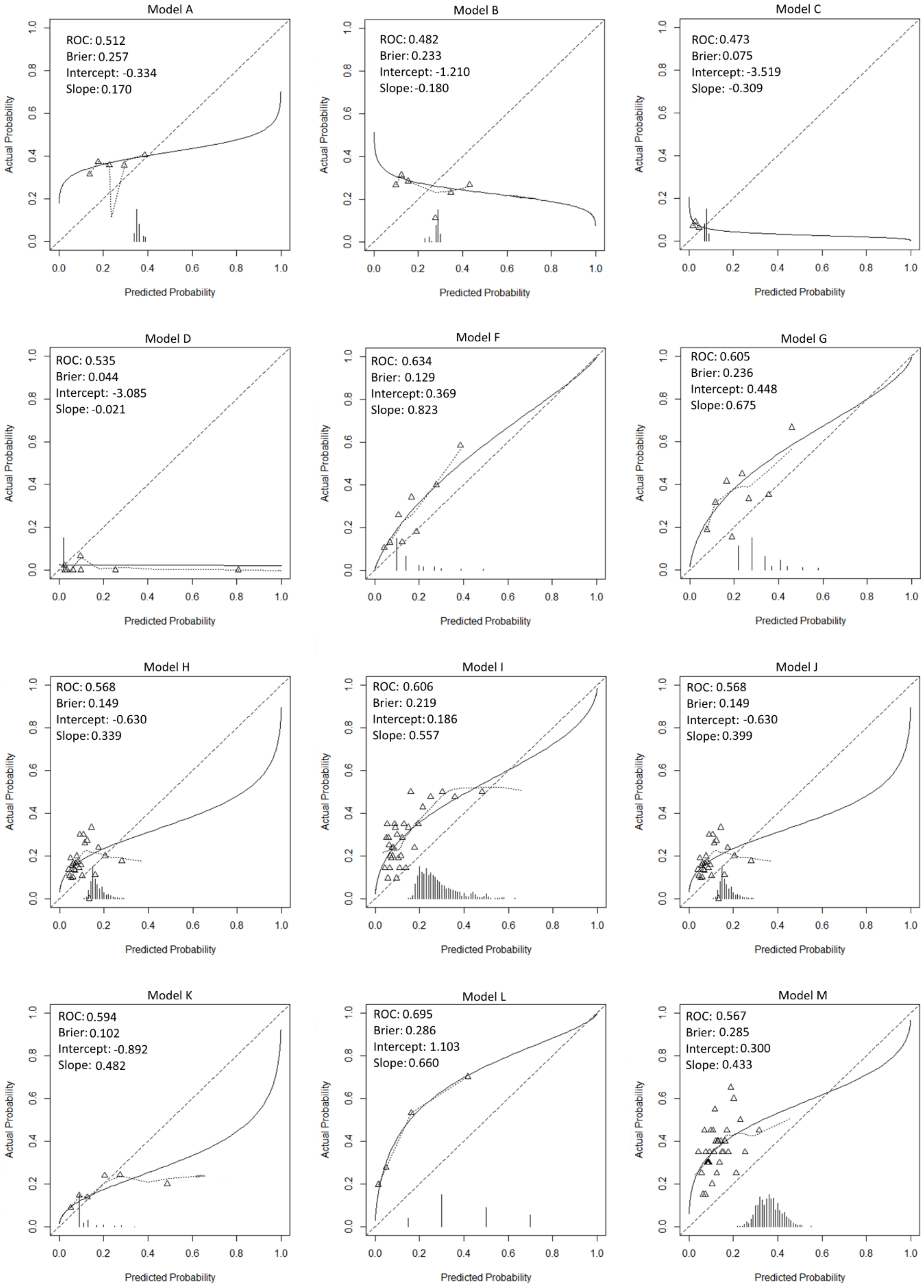
A, B, C	Total dose (Gy)	65 Gy - 15;
		70 Gy - 44;
		80 Gy - 41
			66 Gy - 13;
			70 Gy - 56;
			74 Gy - 30

B	Diabetes	7	13
D	Bladder volume receiving ≥ 70 Gy (cc) †		
	No haematuria	5 (0-22)	0 (0-63)
	Haematuria	8 (1-22)	0 (0-0)
D	Prior TURP	14.9	11.5
E	Clinical target volume (cc) †		
		41 (7-129)	49 (15-199) *
		54 (17-127)	46 (20-147) *
E	Min dose to clinical target volume (Gy) †		
	No nocturia	72 (64-79)	67 (4-82) *
	Nocturia	73(67-78)	67 (52-73) *
F, G, H, I, K	Smoke	16.4	13
G	Age†	71 (46-82)	70 (49-85)
H	Neoadjuvant hormonal therapy	51.9	100
J	Anti-hypertensives	47.3	49*
J	Irradiation of seminal vesicle	61.5	38
G, H	Absolute weekly DSH at 12.5 Gy (cm^2) ‡	NA	0
I	Absolute weekly DSH at 5 Gy (cm^2) ‡	NA	97 (77-120)
L	Absolute weekly DSH at 11.5 Gy (cm^2) ‡	NA	0
F	Baseline Q1	NA	0 – 57; 1 – 26; 2 – 10; 3 – 8
G	Baseline Q2	NA	0 – 30; 1 – 40; 2 – 17; 3 – 13
I	Baseline Q4	NA	0 – 58; 1 – 27; 2 – 10; 3 – 5
J	Baseline Q5	NA	0 – 50; 1 – 28; 2 – 13; 3 – 10
K	Baseline Q6	NA	0 – 76; 1 – 17; 2 – 5; 3 – 2
L	Baseline Q7	NA	0 – 12; 1 – 45; 2 – 27; 3 – 16
M, N	Neoadjuvant hormonal therapy	56	100
M	Planning target volume‡	129 (95-169)	183 (152-225)
M	Absolute weekly DSH at 8.5 Gy (cm^2) ‡	56 (41-81)	54 (43-67)

M	Age (years) ‡	71 (67-75)	70 (64-74)
M	Hypertension	54	49
M, N	Use of cardiovascular drug	34	29*
M	Body mass index (kg/m^2) ‡	26 (24-29)	28 (25-30)
N	Use of hypercholesterolemia drugs	16	30*
N	Absolute weekly DSH at 12 Gy	10 (0-30)	0

Note: TURP - transurethral resection of the prostate; DSH - dose-surface histogram; *Definition varied from the development cohort, refer text for details; † - median (range); ‡ - median (interquartile range), NA: not available.

Figure 1: Calibration curves for models A to M. Calibration of a model describes the extent to which the predicted probability matches the observed probability. Models E and N were not included (see text for details). The *x*-axis indicates the prediction obtained from the predictive models, and the *y*-axis indicates the observed proportions. Patients were divided into groups of 20 (represented by the triangles) based on the increasing predicted risk. Triangles can fall on the same spot resulting in smaller number of triangles shown. The line of unity, at 45 degrees, represents ideal agreement between observed and predicted probabilities. The vertical lines at the bottom indicate the distribution of the predicted probability. The solid line is the fitted logistic function and the dotted line is the non-parametric function.


Model A
ROC: 0.512
Brier: 0.257
Intercept: -0.334
Slope: 0.170
Model B
ROC: 0.482
Brier: 0.233
Intercept: -1.210
Slope: -0.180
Model C
ROC: 0.473
Brier: 0.075
Intercept: -3.519
Slope: -0.309
Model D
ROC: 0.535
Brier: 0.044
Intercept: -3.085
Slope: -0.021
Model F
ROC: 0.634
Brier: 0.129
Intercept: 0.369
Slope: 0.823
Model G
ROC: 0.605
Brier: 0.236
Intercept: 0.448
Slope: 0.675
Model H
ROC: 0.568
Brier: 0.149
Intercept: -0.630
Slope: 0.339
Model I
ROC: 0.606
Brier: 0.219
Intercept: 0.186
Slope: 0.557
Model J
ROC: 0.568
Brier: 0.149
Intercept: -0.630
Slope: 0.399
Model K
ROC: 0.594
Brier: 0.102
Intercept: -0.892
Slope: 0.482
Model L
ROC: 0.695
Brier: 0.286
Intercept: 1.103
Slope: 0.660
Model M
ROC: 0.567
Brier: 0.285
Intercept: 0.300
Slope: 0.433
Actual Probability
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0

Supplementary material A: Distributions of clinical features in the validation cohort. Continuous distributions are specified as mean ± standard deviation (range), categorical variables are specified as number of patients (%).

Factors		**Missing**
Physical & Trial factors		
Age	69 ± 7(49-85) years	3
BMI	27.98 ± 4.12 (17.17-45.77) kg/m^2	22
ECOG Performance Status (=1)	123 (16%)	1
Arm	A (191), B (187), C (192), D (184) (refer to [21, 47]	0
Bladder volume	219.4±89.9 (61.0-561.7) cm^3	13
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

	No. of patients with no missing information: 711	

Abbreviations; OR- Odds ratio; BMI - body mass index; ECOG - ECOG Performance Status; NIDDM – non-insulin dependent diabetes mellitus; IDDM – insulin dependent diabetes mellitus; ACE - angiotensin-converting-enzyme; NSAID – non-steroidal anti-inflammatory drugs; PC- principal component.

Supplementary material B: Expressions for models D to N. Models A to C were based on the nomograms from the original publication [17].

Model	Feature	Symbol
A	Anti-coagulant treatment	x_1
A, B, C	Total dose (Gy)	x_2
B	Diabetes	x_3
D	Bladder volume receiving ≥ 70 Gy (cc) †	x_4
D	Prior TURP	x_5
E	Clinical target volume (cc) †	x_6
E	Min dose to clinical target volume (Gy) †	x_7
F, G, H, I, K	Smoke	x_8
G	Age†	x_9
H	Neoadjuvant hormonal therapy	x_{10}
J	Anti-hypertensives	x_{11}
J	Irradiation of seminal vesicle	x_{12}
G, H	Absolute weekly DSH at 12.5 Gy (cm^2) ‡	x_{13}
I	Absolute weekly DSH at 5 Gy (cm^2) ‡	x_{14}
L	Absolute weekly DSH at 11.5 Gy (cm^2) ‡	x_{15}
F	Baseline Q1	x_{16}
G	Baseline Q2	x_{17}
I	Baseline Q4	x_{18}
J	Baseline Q5	x_{19}
K	Baseline Q6	x_{20}
L	Baseline Q7	x_{21}
M, N	Neoadjuvant hormonal therapy	x_{22}
M	Planning target volume‡	x_{23}
M	Absolute weekly DSH at 8.5 Gy (cm^2) ‡	x_{24}
M	Age (years) ‡	x_{25}
M	Hypertension	x_{26}
M, N	Use of cardiovascular drug	x_{27}

M	Body mass index (kg/m^2) ‡	x_{28}
N	Use of hypercholesterolemia drugs	x_{29}
N	Absolute weekly DSH at 12 Gy	x_{30}

Model	**Model expressions; ln *p*/(1-*p*) =; where *p* is the probability of event**
D	$-3.67+0.40x_{4}+1.43x_{5}$
E	$-2.21+0.31x_{6}+0.36x_{7}$
F	$-3.11+0.50x_{16}+1.15x_{8}$
G	$-2.47+0.43x_{17}+1.02x_{8}+0.02x_{13}$
H	$0.57-0.06x_{9}+0.72x_{8}+0.04x_{13}+1.10x_{10}$
I	$-3.56+0.67x_{18}+0.57x_{8}+0.01x_{14}$
J	$-3.43+0.47x_{19}+0.57x_{11}+0.76x_{12}$
K	$-2.89+0.96x_{20}+0.57x_{8}$
L	$-4.26+1.31x_{21}+0.02x_{15}$
M	$3.168-0.668x_{22}+0.0015x_{23}+0.008x_{24}-0.056x_{25}+0.470x_{26}+0.007x_{27}*x_{24}-0.060x_{28}$
N	$-2.891+0.01x_{29}*x_{30}+0.007x_{27}*x_{30}-0.619x_{22}+0.021x_{30}$

Supplementary material C: Model performance based on internal validation and independent external validation.

	Area under the receivers operating characteristic (ROC) curve	
Model	**Internal validation**	**Independent external validation**
A	-	0.512
B	-	0.482
C	-	0.473
D	0.67	0.535
E	0.60	-
F	0.67	0.634
G	0.63	0.605
H	0.69	0.568
I	0.69	0.606
J	0.63	0.568
K	0.59	0.594
L	0.75	0.695
M	0.67	0.567
N	0.71	-