Radiotherapy and Oncology 120 (2016) 339–345

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Morbidity of prostate RT

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

Noorazrul Yahya [a,b,*], Martin A. Ebert [a,c], Max Bulsara [d], Angel Kennedy [c], David J. Joseph [c,e], James W. Denham [f]

[a] School of Physics, University of Western Australia, Australia; [b] School of Health Sciences, National University of Malaysia, Malaysia; [c] Department of Radiation Oncology, Sir Charles Gairdner Hospital; [d] Institute for Health Research, University of Notre Dame, Fremantle; [e] School of Surgery, University of Western Australia; and [f] School of Medicine and Public Health, University of Newcastle, Australia

ARTICLE INFO

Article history:
Received 11 March 2016
Received in revised form 11 May 2016
Accepted 15 May 2016
Available online 28 June 2016

Keywords:
Independent external validation
Predictive model
Prostate radiotherapy
Normal tissue complications
Urinary symptoms

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 area under 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.

 Radiotherapy and Oncology 120 (2016) 339–345

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 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

* Corresponding author at: School of Physics, University of Western Australia, Stirling Hwy, Crawley, Western Australia 6009, Australia.
E-mail addresses: noorazrul.yahya@research.uwa.edu.au, azrulyahya@ukm.edu.my (N. Yahya).

http://dx.doi.org/10.1016/j.radonc.2016.05.010
0167-8140/

(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 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-months follow-up post-randomisation. The median follow-up for RADAR is 72 months. 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 is 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