Prospective validation of pathologic complete response models in rectal cancer

Citation for published version (APA):

Document status and date:
Published: 01/01/2017

DOI:
10.1002/mp.12423

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:
repository@maastrichtuniversity.nl
providing details and we will investigate your claim.

Download date: 15 Sep. 2023
Prospective validation of pathologic complete response models in rectal cancer: Transferability and reproducibility

Johan van Soest
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University, Medical Centre+, Maastricht 6062 NA, the Netherlands

Elisa Meldolesi
Department of Radiotherapy, Sacred Heart University Hospital, Rome 00168, Italy

Ruud van Stiphout
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University, Medical Centre+, Maastricht 6062 NA, the Netherlands

Roberto Gatta, Andrea Damiani, and Vincenzo Valentini
Department of Radiotherapy, Sacred Heart University Hospital, Rome 00168, Italy

Philippe Lambin and Andre Dekker
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University, Medical Centre+, Maastricht 6062 NA, the Netherlands

(Received 26 October 2016; revised 10 February 2017; accepted for publication 3 April 2017; published 8 August 2017)

Purpose: Multiple models have been developed to predict pathologic complete response (pCR) in locally advanced rectal cancer patients. Unfortunately, validation of these models normally omit the implications of cohort differences on prediction model performance. In this work, we will perform a prospective validation of three pCR models, including information whether this validation will target transferability or reproducibility (cohort differences) of the given models.

Methods: We applied a novel methodology, the cohort differences model, to predict whether a patient belongs to the training or to the validation cohort. If the cohort differences model performs well, it would suggest a large difference in cohort characteristics meaning we would validate the transferability of the model rather than reproducibility. We tested our method in a prospective validation of three existing models for pCR prediction in 154 patients.

Results: Our results showed a large difference between training and validation cohort for one of the three tested models [Area under the Receiver Operating Curve (AUC) cohort differences model: 0.85], signaling the validation leans towards transferability. Two out of three models had a lower AUC for validation (0.66 and 0.58), one model showed a higher AUC in the validation cohort (0.70).

Discussion: We have successfully applied a new methodology in the validation of three prediction models, which allows us to indicate if a validation targeted transferability (large differences between training/validation cohort) or reproducibility (small cohort differences).

Key words: case mix, pathologic complete response, prediction model, rectal cancer, validation

1. INTRODUCTION

As the field of radiation oncology is moving towards individualized medicine, the need to identify (sub-)groups of patients on the basis of patient and/or tumor features is emerging.1 Machine learning techniques using (routine) clinical patient information are needed to identify these features. Furthermore, machine learning can be used to develop a prognostic model for disease development, or to develop a predictive model where the outcome may vary, based on the applied intervention(s). These prognostic and predictive models are the building blocks for clinical decision support systems (CDSS).2 The promise of these CDSSs is to handle and adapt to insights found in research, relieving the clinical staff from the burden of keeping up with the high volume of publications and the rapidly increasing amount of knowledge.3,4

Before implementing clinical prediction models into a CDSS, these models need validation on different levels.5 These levels can be classified using the TRIPOD statement.6 Although in many studies internal/external validations are included, they normally do not describe validation results to their full extent. According to Justice et al.7 validation of prediction models should describe two aspects: Accuracy validation (performance of the model) and generalizability (how similar/dissimilar are training and validation cohorts and why and how do these differences influence the performance of the model).

Accuracy, or model performance validation, describes the statistical validity of a prognostic or predictive model.8 In general, model performance (or fitness) is determined by the discriminative ability and calibration of a prognostic/predictive model.9 The discriminative ability describes how well a
model correctly classifies a subject into the correct group. Calibration describes the agreement of the frequency between observed and predicted events.

The second aspect, generalizability, can be divided into two components: reproducibility and transferability. Reproducibility describes the accuracy of a prediction model on similar cohorts, where transferability tests the accuracy of a prediction model on cohorts with different characteristics. Similarities or differences between two cohorts are affected by temporal, methodological or geographic aspects. An example of a temporal difference is the emerging influence of HPV on head & neck cancer patients. Methodological differences could originate from different treatments being applied in the same patients or different levels of quality (e.g., clinical routine versus clinical trials). Geographical different origins of the training and validation cohort could make these different in, for example, race and socioeconomic factors. Often these are interrelated with geographical differently located cancer centers treating different patients differently at different times. Often, (external) validation of prediction models only describe the accuracy. The method described by Debray et al. can be used to estimate the difference between the training and validation cohort, measuring the level of generalizability (same characteristics) versus transferability (different characteristics) between training and validation cohorts. By adding this measurement, next to the model performance on the validation set, it gives more insight (without hard boundaries) in which situations a prediction model does (not) work (Fig. 1). Therefore, it is imperative to add this measure in the general model validation process, as it better describes for which cohorts a prediction model was tested.

In this work, we aim to investigate this reproducibility and transferability metric in a prospective validation of three prediction models for pathologic complete response (pCR) in rectal cancer patients. These models have been developed and retrospectively validated by van Stiphout et al. based on prior work identifying prognostic factors for pathologic response. We hypothesize that this prospective validation tests for reproducibility, with comparable (or slightly reduced) model performances.

2. Method and Materials

The three models we validated were learned on three different training cohorts as published previously. These models predict pathological complete response (pCR) based on different groups of available data: (a) only clinically available parameters (clinical model), (b) Clinically available parameters + pretreatment PET parameters (pretreatment PET model), (c) Clinically available parameter + pretreatment PET + post-treatment PET parameters (post-treatment PET model). For the PET parameters, tumors were semi-automatically contoured on PET-CT scans using commercial software (TrueD, Siemens Medical, Erlangen, Germany). Standardized Uptake Value (SUV) thresholding was performed using the gluteus muscle to set the threshold for the automatic contouring, within pre-defined boundaries. The response index (RI) describes the ratio between the pretreatment and post-treatment SUV value of the primary tumor. Pathological complete response was determined as having a T0N0M0 based on the surgical specimen, extracted from the pathology report. Based on the three different datasets, an exhaustive feature selection was performed to train a proximal Support Vector Machine (SVM). Internal validation was performed using a leave-one-out cross-validation. Original cohort datasets for training and validation were at our disposal. The cohort used for our prospective validation was the THUNDER trial cohort (NCT00969657). This cohort consists of 154 patients, from two participating centers (MAASTRO Clinic, Maastricht University Medical Centre+, Netherlands and Sacred Heart University Hospital, Rome, Italy). All patients included in this THUNDER trial gave written informed consent before data was collected.

Univariate cohort differences were tested for statistical significance using Wilcoxon rank sum test (for continuous variables) or Fisher’s exact test (for categorical variables). To correct for multiple (univariate) testing, we calculated an adjusted P-value using the Bonferroni correction which multiplies the P-values by the number of comparisons. In our case, multiplying the P-values by the number of model input and output parameters. Cohort characteristics for the prediction model variables are shown in Tables I, II and III.

Next, we calculated the multivariate cohort differences (MCD) using the method proposed by Debray et al. This method assesses the ability to predict whether a specific patient in our cohort belongs to cohort A (training) or B (validation). When we are able to predict to which cohort patients belong, it would mean that (several of) the underlying prediction model variables have very different distributions (pointing to a validation which would test transferability). In contrast, when we cannot predict to which cohort patients belong, it would mean that the model variables are more homogeneous among the training and validation cohort.
Variable Training Validation (pros) P-value adjusted

# Patients 677 112
Tumor length [cm] (SD) 4.97 (1.73) 5.03 (1.81) 9.63-10^-1 9.63-10^-1

cT
1 4 (0%) 0 (0%)
2 18 (3%) 16 (14%)
3 583 (86%) 70 (63%)
4 72 (11%) 26 (23%)

cN 154 (23%) 9 (8%)
1 307 (45%) 35 (31%)
2 216 (32%) 68 (61%)
pCR 134 (20%) 29 (26%)

After performing tests to describe univariate and multivariate cohort differences, we compared the distributions of predicted probabilities in the training and validation cohorts by calculating mean probabilities and corresponding standard deviations. Furthermore, we evaluated the prediction model performance on both cohorts using the Area under the Receiver Operating Curve (AUC), Hosmer-Lemeshow C-statistic and Brier score to determine the discriminative ability, calibration and accuracy, respectively. These performance measures have different characteristics: The AUC specifies the ability to make a threshold, separating the probabilities for a given outcome into a binary yes/no result (discriminative performance). Unfortunately, this AUC doesn’t take the distance between a probability and the actually measured outcome into account, hence only determines the best operating (threshold) point. In contrast, calibration measures how well the predicted probability is comparable to the actual incidence of the outcome. For example, the Hosmer–Lemeshow C-statistic splits patients into n groups, based on ordered prediction probabilities, and uses the Chi-square test to assess statistical significant differences between observed and predicted outcomes. Finally, aspects from both discriminative ability and calibration are available in accuracy measurements. One of these measurements is the Brier score, which is the mean squared error between probabilities and observed outcomes. This score is not suitable as a single measure; however is useful when comparing different models with equal outcomes and/or cohorts. For more information regarding these model performance metrics, we would like to refer to Steyerberg et al.

For robustness purposes, we used bootstrapping as a resampling technique (R = 1000), and applied this method to the discrimination (AUC) and accuracy (Brier score) measurements. All calculations and statistical analysis were performed using R (version 3.3.2). A generalized workflow of the applied methods is shown in Fig. 2.

3. RESULTS

The multivariate cohort differences are shown in Table IV. For every prediction model, we made a separate multivariate
TABLE IV. Multivariate differences model for clinical, pretreatment PET and post-treatment PET prediction model, based on original and current prospective validation cohort.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>P-value</th>
<th>CD-AUC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical prediction model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>2.72</td>
<td>5.05×10⁻⁴</td>
<td>0.69 (0.61–0.71)</td>
</tr>
<tr>
<td>Tumor length</td>
<td>0.04</td>
<td>5.09×10⁻¹</td>
<td></td>
</tr>
<tr>
<td>cT</td>
<td>0.11</td>
<td>6.69×10⁻¹</td>
<td></td>
</tr>
<tr>
<td>cN</td>
<td>-1.01</td>
<td>5.65×10⁻⁹</td>
<td></td>
</tr>
<tr>
<td>pCR</td>
<td>0.57</td>
<td>2.36×10⁻²</td>
<td></td>
</tr>
<tr>
<td><strong>Pretreatment PET prediction model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1.32</td>
<td>1.23×10⁻¹</td>
<td>0.85 (0.78–0.89)</td>
</tr>
<tr>
<td>Max tumor diameter</td>
<td>0.51</td>
<td>5.18×10⁻⁶</td>
<td></td>
</tr>
<tr>
<td>cN</td>
<td>-1.36</td>
<td>1.06×10⁻⁶</td>
<td></td>
</tr>
<tr>
<td>Tumor location</td>
<td>-0.66</td>
<td>1.59×10⁻³</td>
<td></td>
</tr>
<tr>
<td>SUV max</td>
<td>-0.07</td>
<td>6.21×10⁻³</td>
<td></td>
</tr>
<tr>
<td>pCR</td>
<td>-0.81</td>
<td>8.31×10⁻²</td>
<td></td>
</tr>
<tr>
<td><strong>Post-treatment PET prediction model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.35</td>
<td>7.71×10⁻¹</td>
<td>0.62 (0.51–0.67)</td>
</tr>
<tr>
<td>Response index</td>
<td>-0.01</td>
<td>4.91×10⁻¹</td>
<td></td>
</tr>
<tr>
<td>SUV max pre/post</td>
<td>0.09</td>
<td>3.35×10⁻¹</td>
<td></td>
</tr>
<tr>
<td>Tumor length</td>
<td>0.04</td>
<td>6.30×10⁻¹</td>
<td></td>
</tr>
<tr>
<td>pCR</td>
<td>0.31</td>
<td>4.69×10⁻¹</td>
<td></td>
</tr>
</tbody>
</table>

The CD-AUC values in Table IV, we can probably state that the clinical and pre–post PET prediction models are being validated for reproducibility, where the pretreatment PET prediction model is being validated for transferability. The CD-AUC of the pretreatment PET prediction model indicated a high predictive ability whether a patient belongs to the training/validation cohort. This is also expressed in the (multivariate) cohort differences model coefficients (Table IV) deviating from 0.

The comparison of distributions of predicted probabilities in the training and validation cohorts for all three prediction models are shown in Fig. 3. For the clinical + pretreatment PET model, the mean and standard deviations for the predicted probabilities are almost equal in both training and validation cohorts. The post-treatment PET model shows a higher average probability in the validation dataset, with a smaller standard deviation. The latter could be due to the small number of patients available for this prediction model.

4. DISCUSSION

The In this work, we have successfully performed a prospective validation (TRIPOD statement type 4) of three previously developed prediction models, and applied an additional method to assess the differences between training and validation cohorts. In addition to the traditional accuracy validation, our analysis gives additional information to clinicians whether the validation was performed on a similar or different cohort (in terms of population characteristics), and therefore whether the validation assessed the reproducibility (possible same clinical setting), or transferability (possible different clinical setting) of a prediction model. As these measures are relatively easy to interpret, they could be used when commissioning prognostic models for use in clinical practice, by assessing whether the population in a certain clinic is different from the population where the model was trained on.

We would advise to validate prediction models on trial and routine clinical cohorts as also suggested by Booth and Tannock27 and proposed in the V ATE project.28,29 The quality of cohorts from clinical trials are needed to identify which variables need to be reported in clinical practice. Afterwards, training/validating models (using the methods explained here) on routine clinical data would increase the cohorts available to learn/validate upon as was done by Shen et al.30 Furthermore, validation in a clinical setting could also reduce the turnaround time between developing/validating and using predictive models in clinical practice; enabling rapid learning healthcare and subsequently decision support.2,3
When evaluating the results, the significance of the univariate differences ($P$-values between training and validation cohort; Tables I, II and III) generally overlapped with the multivariate cohort differences, described by the covariate weight $P$-values (Table IV). But several variables which were significant in the univariate variable assessment lost their significance in the multivariate assessment (e.g., clinical T-stage); or became significant (e.g., pCR). In our opinion, this
could be affected by differences in sample sizes between training and validation, or the effect of testing one variable versus testing the complete characteristic of a patient. Although this correlation reduces the added value of the cohort differences metric, we still think this metric is an added value as a single measure to assess cohort differences: to determine whether external validation results measure reproducibility or transferability. Secondly, statistical tests only measure significant differences; the cohort differences model can reveal subtle differences which only become apparent in a multivariate analysis.

For the post-treatment PET model, our main hypothesis is that the prediction model was overfitted on the training cohort (pCR positive outcomes = 26). When calculating a sample size for model training, we would use 10 events per variable as used in this rule of thumb. As an example, the training cohort would need 30, 40 and 30 events (pCR) for the clinical, pretreatment PET and post-treatment PET model, respectively. When considering a pCR percentage of 20%, this would result in a population size of 150, 200 and 150 patients, respectively. As a result, only the clinical prediction model training cohort would be considered large enough. Regarding the validation cohort, the only studies investigating model validation cohort sizes up to our knowledge are by Collins et al. and Vergouwe et al. They do state that 100 events would be a minimum, meaning that the required sample size would be 500 patients, considering a pCR event rate of 20%. Therefore we have to state that our validation might be underpowered, however, could only be accomplished by large multicenter trials. This also means that the cohort difference model and AUC values cannot reliably detect a difference in cohorts in underpowered datasets.

Future work would include the validation of the clinical pCR prediction model in a routine clinical cohort, and investigate applicability of prediction models in clinical practice.

5. CONCLUSION

In general, we would advise to apply the explained methods when validating (existing) prediction models, as it puts prediction model performance in perspective of the heterogeneity between training and validation cohorts. Our workflow (Fig. 1) could therefore be used as a guideline.

Based on these results, we can also state that the clinical prediction model performed well when reproducing results in the current prospective validation. The pre- and post-treatment PET prediction models were unfortunately underpowered in both training and validation cohorts.

CONFLICT OF INTEREST

This work was partially funded by Varian Medical Systems (VATE & SAGE project).

Author to whom correspondence should be addressed: Electronic mail: johan.vansoest@maastro.nl; Telephone: +31 (0) 88 44 55 578.

REFERENCES
