Radiotherapy and Oncology xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original article

# Why validation of prognostic models matters?

Alex Zwanenburg [a,b,c], Steffen Löck [a,c,d,*]

[a] OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden – Rossendorf; [b] National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany, and; Helmholtz Association/Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany; [c] German Cancer Consortium (DKTK), Partner Site Dresden, and German Cancer Research Center (DKFZ), Heidelberg; and [d] Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Germany

ARTICLE INFO

*Article history:*
Received 23 February 2018
Received in revised form 7 March 2018
Accepted 8 March 2018
Available online xxxx

*Keywords:*
Prognostic models
Validation
TRIPOD
Biomarkers

SUMMARY

Prognostic models are powerful tools for treatment personalisation. However, not all proposed models work well when validated using new data, despite impressive results being reported initially. Here, we will use a hands-on approach to highlight important aspects of prognostic modelling, as well as to demonstrate methods to generate generalisable models.



One central objective for the improvement of radiotherapy is treatment personalisation, which for example tailors the prescribed type of chemotherapy, targeted drug or radiation dose to the specific phenotype of the tumour [1]. This approach requires evaluating the radiosensitivity of every tumour before treatment, which is increasingly based on prognostic models that translate clinical, molecular or imaging characteristics of the tumour into a prognosis of patient outcome. Such prognostic models provide scores, e.g., for the risk of local tumour recurrence, likelihood of treatment success, or the probability of developing treatment side effects. Thus, such models may assist a clinician in tailoring the treatment to the patient.

Prognostic models are built using biomarkers. Biomarkers measure a biological process or characteristic, and represent it by a single value. Biomarkers can be extracted from many sources, e.g. clinical records (age, smoking, gender, performance scores), tumour biopsies (gene mutations, copy number alterations, HPV status), blood, and medical imaging (tumour volume, maximum SUV, texture). Building a prognostic model is relatively straightforward. Building a prognostic model that is generalisable and suitable for prospective use is, however, far more challenging. In 2016, the TRIPOD statement was released to provide guidelines for reporting on prognostic models [2,3]. At the same time the statement highlights some methodological pitfalls that need to be avoided, and offers recommendations which ought to be followed. In this work we will assess some of the methodological implications of the TRIPOD statement and give recommendations for building generalisable prognostic models. We will focus specifically on the validation of prognostic models, as validation is used to assess generalisability. To illustrate the importance of validation, we present example cases based on publically available data from the Cancer Genome Atlas Head and Neck Squamous Cell Carcinoma (TCGA-HNSCC) provisional data set [4–6].

## Validation in prognostic modelling

Learning algorithms (learners) use development (training) data to build mathematical models. After model development, the model may be used to predict an outcome for new cases. The generalisability of these models is assessed using validation data. Generalisable models possess similar prognostic power for both development and validation data. The TRIPOD statement defines 4 different types of prognostic modelling strategies, based on how development and validation data are defined and used. In type 1 analyses, a single data set is used for model development. In type 2 analyses still only one data set is available, but the data are split into separate sets for development and validation. Type 2 analyses are often referred to as "internal validation" due to the use of a single data set. Type 3 analyses use a completely separate data set for validation. The validation data are separate in the sense that it has a different origin than the development data, i.e. the data come

* Corresponding author at: OncoRay – National Center for Radiation Research in Oncology, Medical Faculty Carl Gustav Carus, Technische Universität Dresden, Fetscherstr. 74, 01307 Dresden, Germany.
*E-mail address:* steffen.loeck@oncoray.de (S. Löck).

https://doi.org/10.1016/j.radonc.2018.03.004
0167-8140/

from a different treatment centre or a different study. In type 4 analyses no development data are available at all. Instead, an existing prognostic model is provided and applied to a new data set. Analyses of types 3 and 4 are referred to as "external validation".

Here we present different analyses of types 1, 2, and 3 by building prognostic models for overall survival (OS) in a cohort of HNSCC patients and subsequently validate these models.

## Patients

We used data from the Cancer Genome Atlas Head and Neck Squamous Cell Carcinoma (TCGA-HNSCC) provisional data set. Overall survival after treatment was chosen as the endpoint of interest. Clinical descriptors (43) and gene mutation data (285) were used as input for prognostic modelling, leading to a combined set of 328 biomarkers (features), which is described in the supplement. The data were divided into development and validation cohorts based on the contributing TCGA treatment centre, see Fig. 1. 132 patients with available gene mutation data were selected from the *CV* sub-cohort as a development cohort. An additional 126 patients with complete data were selected from the *CN* and *CR* sub-cohorts to serve as a validation cohort. Clinical characteristics of both cohorts are shown in Supplementary Table 1.

## Example cases

We present five example cases demonstrating analyses of type 1 (examples 1 and 2), type 2 (examples 3 and 4) and type 3 (example 5). For these examples we build Cox proportional hazards models. Their prognostic performance is assessed using the concordance index (c-index) [7,8]. The c-index ranges between 0.0 and 1.0. 0.5 corresponds to a random ordering, and 1.0 indicates the ability to perfectly order patients by the predicted log hazard ratios.

The used modelling framework is described in the supplement, which includes a detailed description of the models for examples 1 and 2 as well. A summary overview of the results is presented in Table 1.

### *Example 1: type 1 analysis without feature selection*

The first example is a type 1 analysis which draws upon the 132-patient development cohort only. Including all 328 features into the model is not possible. The model does not converge, as there are more features than data points. A commonly used strategy is therefore to perform a univariable analysis and select all features with a significance level below a set threshold, e.g. 0.05 or 0.01, which may be corrected for multiple testing. The selected features are used to build a Cox model. Forward or backward selection strategies are also commonly applied, with backward selection being suited to low-dimensional cases only. For the purpose of our example, we select 20 features with $p < 0.05$. This model achieves a c-index of 0.71.

### *Example 2: type 1 analysis with feature selection*

The second example is similar to example 1, but includes feature selection using LASSO regression [9,10], as well as sequential model-based boosting (SMBO) [11] to select the signature size. Hence the signature consists of only 7 features and the reported model performance is slightly worse than that in example 1, with a c-index of 0.68.

### *Example 3: type 2 analysis using cross-validation*

The third example again only draws upon the development cohort. However, the cohort is now repeatedly split into training and validation sets using stratified cross-validation, and is thus a

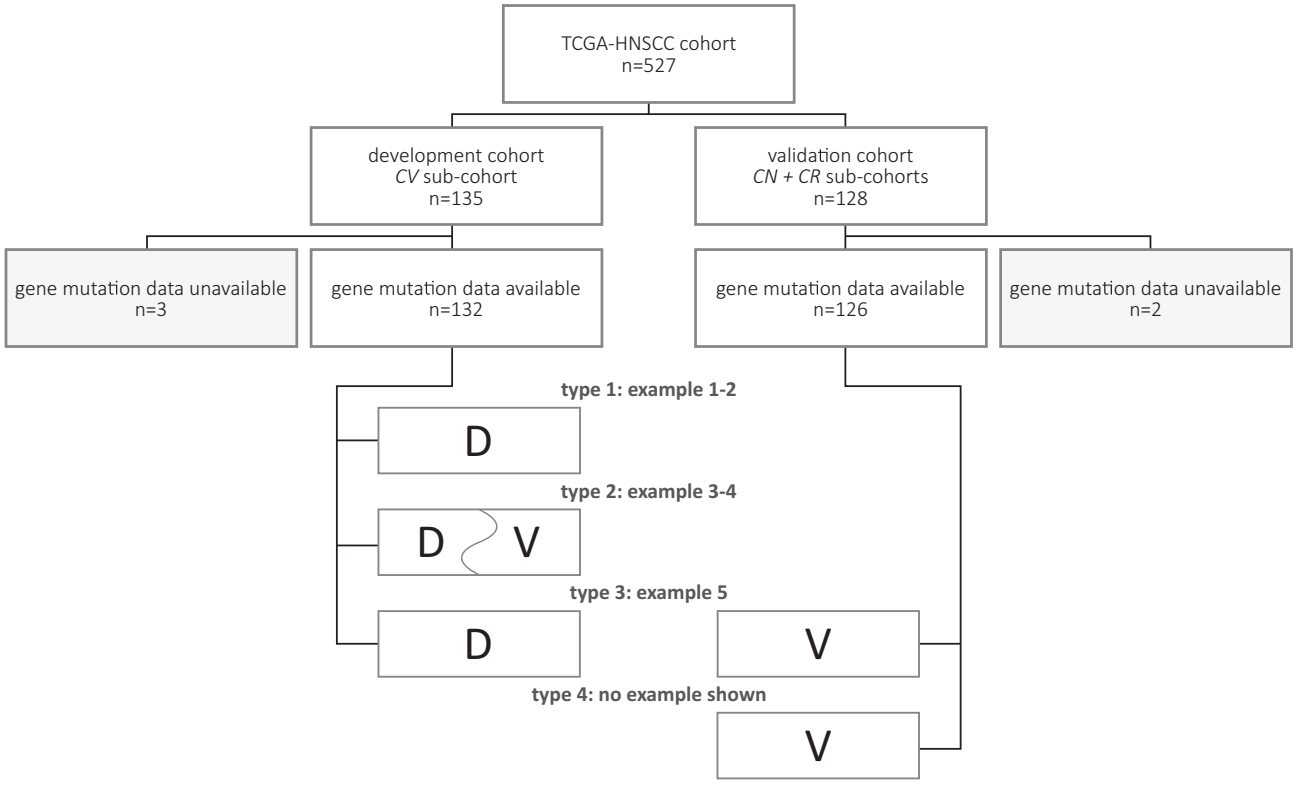


**Fig. 1.** Experimental setup and examples. Sub-cohorts from the TCGA-HNSCC cohort are used as development cohort (D) and validation cohort (V). For examples 1–4 only the development cohort is used for development and validation. In example 5, the validation cohort is used to externally validate prognostic models from examples 1–4.

Please cite this article in press as: Zwanenburg A, Löck S. Why validation of prognostic models matters?. Radiother Oncol (2018), https://doi.org/10.1016/j.radonc.2018.03.004