Modelling for Radiation Treatment Outcome

Dutz, Almut; Zwanenburg, Alex; Langendijk, Johannes; Löck, Steffen

Published in:
Image-Guided High-Precision Radiotherapy

DOI:
10.1007/978-3-031-08601-4_13

Publication date:
2022

Citation for published version (APA):
Dutz, A., Zwanenburg, A., Langendijk, J., & Löck, S. (2022). Modelling for Radiation Treatment Outcome. In *Image-Guided High-Precision Radiotherapy* (pp. 285-314). Springer International Publishing AG. https://doi.org/10.1007/978-3-031-08601-4_13

Modelling for Radiation Treatment Outcome

13

Almut Dutz, Alex Zwanenburg, Johannes A. Langendijk, and Steffen Löck

Almut Dutz and Alex Zwanenburg shared first authorship.

A. Dutz
OncoRay—National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden—Rossendorf, Dresden, Germany

Helmholtz-Zentrum Dresden—Rossendorf, Institute of Radiooncology—OncoRay, Dresden, Germany
e-mail: almut.dutz@oncoray.de

A. Zwanenburg
OncoRay—National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden—Rossendorf, Dresden, Germany

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
e-mail: alexander.zwanenburg@nct-dresden.de

J. A. Langendijk
Department of Radiation Oncology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
e-mail: j.a.langendijk@umcg.nl

S. Löck (✉)
OncoRay—National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden—Rossendorf, Dresden, Germany

Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
e-mail: steffen.loeck@oncoray.de

E. G. C. Troost (ed.), *Image-Guided High-Precision Radiotherapy*,
https://doi.org/10.1007/978-3-031-08601-4_13

13.1 Introduction

As many types of cancer are life-threatening, tumour control has a high priority in cancer treatment. However, radical treatment is often limited by the surrounding healthy organs that may lose their functions. These relationships were already considered by Holthusen [1] who described the probability of achieving tumour control and developing normal tissue damage after radiotherapy as a function of radiation dose (Fig. 13.1). The ability to deliver a sufficient tumour dose with a tolerable level of side effects is characterised by the therapeutic window, which defines the target dose prescription as well as dose limits for organs at risk (OARs) [2]. Efforts have been made to widen the therapeutic window and to increase tumour control or to reduce the risk of side effects, e.g. by modified fractionation schemes, new technologies, or biological modulation [3]. For these efforts, statistical modelling is essential, relating patient and treatment-specific risk factors that are associated with tumour radiosensitivity or normal tissue response to defined endpoints.

Statistical models for radiation treatment outcome are becoming increasingly specific and complex. This is caused by two factors. One is the growing amount of patient-specific data that are being collected and made accessible using electronic hospital information systems. With decreasing costs, an increasing number of patients receive in-depth analyses of their tumour tissue, generating multi-omics data that may comprise thousands to millions of parameters from genomic, methylomic, proteomic, radiomic, histomic, and other analyses. In addition, longitudinal data are more commonly acquired, including repeated imaging or liquid biopsies. The second factor comprises advances in computer technology and

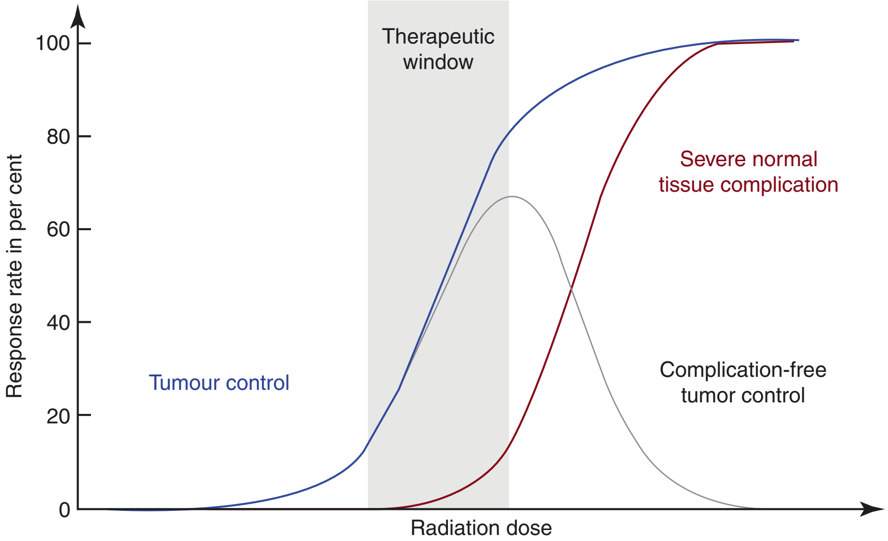

Fig. 13.1 Schematic dose–response curves for tumour control and severe normal tissue complication

machine-learning algorithms, which allow for rapid analyses of these big data and their integration in statistical models.

These developments enable new possibilities: tumour control probability (TCP) models that classify patients into groups with a different risk of treatment failure are essential for biomarker-guided interventional trials. Information from normal tissue complication probability (NTCP) models can be considered in addition to physical parameters during treatment plan optimisation in order to reduce estimated risks of complication (biological treatment plan optimisation). Moreover, predictions of NTCP models may support clinicians in identifying the optimal treatment plan among different planning options or even among different treatment modalities. Another aim that is facilitated by statistical models is adaptive radiotherapy, where radiation treatment is altered during fractionated treatment depending on tumour and normal tissue responses [4, 5].

Since the clinical application of statistical models may substantially affect the treatment of patients, the question arises, what is a good model? A good model addresses a relevant question in a reliable and reproducible manner. It is interpretable. It should either be better than the clinical standard or equivalent to it, but more efficient. Hence, not every model reported in scientific literature will find clinical application. Models may lack generalisability outside the cohort in which they were originally developed. Other models may lack reproducibility due to incomplete reporting. Again, some models may not actually address clinically relevant problems. And finally, models may require data that are too expensive or time-consuming to obtain during clinical routine. To assess these intricacies and to increase the rate of successful translation of models into clinical practice, a general understanding of statistical modelling principles and their application is essential.

In this chapter, we therefore first outline general modelling principles, comprising data types and endpoints, data pre-processing, modelling strategies, and validation procedures (Sect. 13.2). We then provide basic details on modelling tumour response and complication probabilities of normal tissue (Sect. 13.3). Finally, we present two relevant applications of outcome modelling in radiotherapy: the model-based approach for assigning patients to photon or proton-beam therapy based on NTCP models (Sect. 13.4) and radiomics analyses using medical imaging data to predict tumour control (Sect. 13.5).

13.2 Basic Modelling Principles

A model essentially describes the relationship between input variables (features) and an endpoint (outcome). In this section, we describe a modelling workflow and related approaches, see Fig. 13.2 for an overview.

13.2.1 Data

In recent years, the amount and complexity of available data in radiation oncology have increased substantially. Besides demographic, tumour or treatment-related

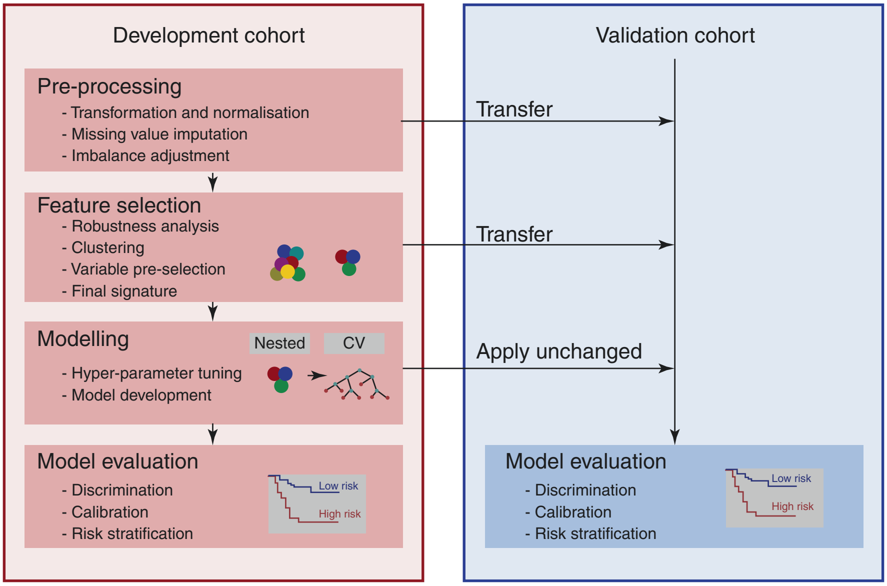

Fig. 13.2 Schematic representation of the described modelling workflow. Data are divided into two separate cohorts. The development cohort is used to create the model, which is subsequently validated on unseen data in the validation cohort. The model development process consists of several steps, including data pre-processing and feature selection, which produces additional parameters. Such parameters (e.g. scale and shift parameters for normalisation of features) are transferred to the validation cohort so that data in both cohorts are pre-processed in the same manner. *CV* cross-validation

factors, dosimetric parameters and pathological findings, increasingly complex features from the analyses of tumour tissue or liquid biopsies and from medical imaging are available. These data are used to predict specific endpoints that may be categorical (e.g. severity grades of side effects), numerical (e.g. hypoxic fraction of the tumour), or survival data (containing an event time and an event indicator, e.g. progression-free survival). Depending on the outcome type, different modelling strategies have to be applied, see Sect. 13.2.5.

The quality of a model is highly dependent on the quality of the data. In general, high-quality data must meet the following criteria:

1. **Cover Patient Heterogeneity:** The cohort used to develop the model should represent the population to which it will be applied. For example, a TCP model created using a cohort of patients with locally advanced head and neck squamous cell carcinoma (HNSCC) may not be reliable for predicting TCP of patients with early-stage HNSCC or of patients with pancreatic cancer.
2. **Completeness:** Complete data have no or very few missing feature values. Features that contain many missing values will typically fail to relate to the outcome.

3. **Uniform Labelling:** The outcome is measured in the same manner for all samples. For example, progression-free survival should be measured from the same starting date, e.g. from the start of radiotherapy or diagnosis, but not both. Likewise, follow-up should be conducted similarly for all patients and radiation-induced side effects should be reported using the same grading system and evaluation criteria.
4. **Reproducible Acquisition:** For example, tumour tissue or OARs should be segmented according to standardised clinical guidelines so that extracted parameters can be compared between patients. Imperfect reproducibility can be somewhat mitigated by ensuring that sufficient data are available to identify robust parameters.

The above requirements are generally not easy to fulfil. Covering patient heterogeneity requires a sufficient sample size in order to still detect relevant effects. Moreover, these cohorts should preferably be obtained from different institutions to allow for identifying and correcting institutional biases, e.g. due to different equipment, treatment workflows, or follow-up procedures. Uniform labelling and reproducible acquisition require standardised protocols and guidelines for prospective application and data curation for retrospective studies.

In particular, dosimetric parameters and image features depend on the delineation of OARs and target structures. This should be performed according to standardised contouring guidelines to assure uniform structures. Automated contouring may also be considered. To ensure consistent evaluation of outcomes and reduce inter-observer variability, data on side effects should be collected prospectively using standardised tests or grading systems (e.g. Common Terminology Criteria for Adverse Events [CTCAE]) by continuously trained clinical staff. Predefined long-term follow-up should be preferred, taking care to ensure the completeness of the outcome data. In addition, prospective scoring of various potential predictor variables such as patient-, disease-, and treatment-specific data is required.

Meeting these requirements can be greatly facilitated by the use of digital information systems, such as electronic health records. In addition, structured databases may link the different available clinical systems, e.g. PACS, DICOM servers, biobanks, study databases, and others. This enables standardised and structured data acquisition as well as curation and annotation of data, which in the end facilitates sharing and linking of data with other institutions and thereby the collection of larger datasets.

13.2.2 Data Analysis Strategy

The most important concern of modelling, after identifying the question and the required data, is the strategy used to analyse the data. The analysis strategy defines which data are used to develop the model, and which data are used to subsequently validate it. Validation is important because it has to be demonstrated that the model works as expected. Assessment of the model should be unbiased [6]. Before we describe different analysis strategies, it is important to consider what interrelated sources of bias may occur in an analysis:

1. **Overfitting**: Given sufficient features, models can learn to predict the outcome for development samples without error. This comes at a trade-off, as such models will typically fail to accurately predict the outcome for new samples, i.e. the model overfits the development data (Fig. 13.3). Overfitting is typically associated with increasing model complexity, i.e. the use of a large number of features relative to the sample size, or model algorithms that can capture high-dimensional data, or both. For linear regression models, ten events per feature are often recommended to prevent overfitting as a simple rule [7].
2. **Underfitting**: Underfitting is the opposite of overfitting. A model underfits when an increase in model complexity would have noticeably improved accuracy of model predictions for new samples (Fig. 13.3). Underfitting is relatively uncommon.
3. **Structural Information Leakage**: Information leakage occurs when information concerning the validation data is used during model development. As a result, the error in predictions on the validation dataset will be smaller than without leakage. Leakage occurs in many forms, e.g. the presence of identical samples in development and validation datasets or feature selection based on the combined dataset. Structural information leakage can be entirely prevented through careful data curation and appropriate methodology.
4. **Developer-Driven Information Leakage**: Developer-driven information leakage occurs when the person responsible for developing a model uses results obtained from the validation dataset, e.g. to select a particular model, tweak modelling parameters, or select important features. Developer-driven information leakage is more pernicious than structural leakage because it is difficult to prove or disprove. The best way to avoid this issue is to limit access to the validation dataset entirely until a model has been completely developed. To a lesser extent, this issue may also be addressed by registering the protocol for the modelling experiment, registering the data prior to the experiment, and automating parameter selection and other modelling steps.

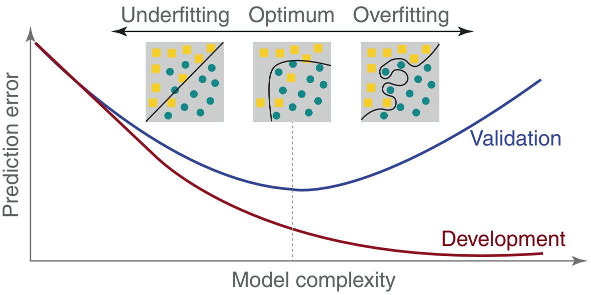

Fig. 13.3 Over- and underfitting during model development. The prediction error of the development data set (red) and the validation data set (blue) is shown as a function of model complexity. For an optimal model, the validation error has a global minimum

Four different types of analysis are described in the TRIPOD guidelines (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) and assessed in terms of their level of evidence [8]:

Type 1: The development data used to create the model are also used to validate the model. An important limitation of this approach is the tendency to produce optimistic biases due to overfitting and information leakage.

Type 2: The available dataset is split into development and validation subsets. While the validation performance will be more realistic than for type 1 analyses, these approaches are still limited because general characteristics may be shared across development and validation sets. Hence, the model is not necessarily generalisable.

Type 3: A separate dataset is used to externally validate the model. This dataset is recruited separately, e.g. from a different study in the same institution, or from a different institution. The latter is preferable because this demonstrates model behaviour and performance in the presence of potential institutional biases.

Type 4: A model is first developed (and published) and then applied to a new dataset. Type 4 analyses provide the most reliable assessment of model performance, as it avoids information leakage.

Type 3 and 4 analyses represent external validation. Models should preferably be assessed using these analyses as the results tend to be more representative of actual model performance. For model development, we moreover recommend splitting the development dataset into internal development and validation subsets, e.g. using repeated (stratified) cross-validation. The internal validation subsets can be used to evaluate whether a model would over- or underfit by comparing model errors between the subsets. Therefore, they can be used to guide the choice for different modelling parameters, e.g. to choose a particular modelling algorithm or a signature of features included in the model.

13.2.3 Data Pre-Processing

Before models can be created, data should be pre-processed. This typically includes steps such as transformation, normalisation, and missing value imputation [9]. Though there is no fixed approach to pre-processing, we propose the following.

First, features and samples that have a large fraction (e.g. >10%) of missing values as well as constant features can be removed. Several modelling algorithms assume that numerical features follow a normal distribution. Hence, the remaining features can be power-transformed to make them follow a normal distribution more closely. A typical transformation is logarithmic transformation, but Box-Cox [10] or Yeo-Johnson power transformations [11] offer a more flexible approach.

Normalisation is used to ensure that each numerical feature has a similar value range, as modelling algorithms can be sensitive to features with greatly varying value ranges. Common methods are standardisation, which centres values at 0 by subtracting the mean value and scales their range by dividing by the standard deviation, and rescaling, which limits feature values to a [0,1] or [−1,1] interval by dividing a feature by the range of its values.

Normalisation is also a common preliminary step to batch normalisation. These methods are used to reduce technical sources of variation between the samples (batch effects), e.g. due to different imaging devices or protocols [12, 13]. All normalisation methods that can be used over the entire data set, can also be employed for batch normalisation, e.g. standardisation [14] or the ComBat algorithm [15]. However, batch normalisation may obfuscate or enhance batch effects due to actual differences in patient outcome between cohorts.

Remaining missing values may cause statistical issues and some modelling algorithms will fail to work if they are present. Therefore, they need to be addressed. One method is simply omitting all samples with missing values. However, this may bias results and leads to the loss of other, perhaps more relevant, information [16]. It is generally better to impute missing data, for which various methods exist [17, 18].

Another issue that may be addressed during pre-processing is imbalance in outcome classes. For example, low-grade radiotoxicity is generally more prevalent than high-grade toxicity. As a consequence, a model that predicts the probability of side effects may overemphasise the more frequent low-grade toxicity (majority class) and be insensitive to the rare high-grade toxicity (minority class). Balancing the outcome classes mitigates this issue, which requires either undersampling the majority class or oversampling the minority class [19, 20]. Both undersampling and oversampling have disadvantages. Undersampling is at the expense of removing samples, whereas oversampling requires the generation of synthetic data. The SMOTE [21] and ADASYN [22] algorithms are commonly used for oversampling. Class imbalances, however, may also be considered outside of pre-processing, e.g. through modelling algorithms that can handle class imbalance [23, 24].

13.2.4 Feature Selection

In modern clinical datasets, the number of features can well exceed the number of samples. However, only some of these features will be important for the outcome. We can make the modelling process more efficient by first excluding non-reproducible features, further reducing the dimensionality of the problem, and only then determining the importance of the remaining features.

In particular, in datasets where the number of features exceeds the number of samples considerably, some features may be highly dependent on the specific experimental conditions and are thus not reproducible in repeated experiments or by other centres. Such features should be excluded. For example, through repeated measurements, it has been established that radiomics features computed from medical imaging have varying degrees of reproducibility [25]. Feature reproducibility may be identified from the literature, by performing repeated measurements, through the use of phantom data, or perturbation of image data in case of radiomics [26]. Sometimes it is not possible to assess robustness and care should be taken in the interpretation of the obtained results.

Dimensionality reduction may be approached by projecting the actual feature space to a lower-dimensional feature space, e.g. by principal component analysis or linear discriminant analysis [27]. As an alternative, unsupervised clustering algorithms may be applied to remove highly correlated features. Such features carry essentially the same information and are thus redundant. Moreover, the presence of redundant features may lead to correlation bias [28]. Hence, such features can be replaced by a single cluster feature [29, 30]. Clusters are formed by computing the similarity between pairs of features using certain metrics, such as Spearman's rank correlation coefficient (for numerical features) or McFadden's pseudo R^2 (for any feature type), as input to cluster algorithms [31]. Each cluster can then be represented by a single feature, e.g. the central feature, a meta-feature such as the mean value across all features in the same cluster or the feature that is most strongly related to the outcome.

The remaining features are used in feature selection, aiming to identify the most important features that show the strongest association with the endpoint and should be incorporated into a model [32–34]. However, feature selection results may be sensitive to the underlying dataset [35, 36]. To improve the stability of results, feature selection can be repeated using resampled subsets of the data [37–39]. Feature importance in each subset is then aggregated over the ensemble of subsets to obtain an ensemble feature importance [40]. The final number of included features (signature size) can be determined during hyperparameter optimisation, which is described in the next section. Some modelling algorithms, such as LASSO regression [41] and model-based boosting [42, 43], perform feature selection internally. Still, such algorithms may benefit from filtering irrelevant features and removing redundant ones.

13.2.5 Model Training

Modelling algorithms try to learn the relationship between features and the outcome. Hundreds of algorithms have been devised [44] and their applicability may depend on the type of the considered outcome, see Table 13.1. We generally recommend starting with the use of simple algorithms such as generalised linear models [52] or algorithms based on the least absolute shrinkage and selection operator [41]. The models created by such algorithms are easily understood and reported, and they can be used as baseline models. Given sufficient samples, more complex algorithms such as random forests [53] and extreme gradient boosting [54] may produce models that give better results.

Complex algorithms are characterised by the presence of many model hyperparameters, such as the number of decision trees in a random forest or the learning rate in extreme gradient boosting. However, even simple models have one or more hyperparameters, such as the signature size. Hyperparameters need to be provided manually or determined from the data through an optimisation process. An advantage of automatic optimisation is that it avoids manual bias. Grid search is a common method that samples the hyperparameter space at specified positions, and trains and evaluates a model at each position. This works well for simple models

Table 13.1 Common models and metrics for model discrimination based on categorical, numerical, and survival endpoints

	Categorical endpoint	Numerical endpoint	Survival endpoint
Example models	Logistic regression Support vector machines Neural networks Random forest	Linear regression Random forest Neural networks	Cox regression Boosted-tree regression Survival random forest
Discrimination metrics	Area under the receiver operating characteristic curve (AUC) [45]	Mean-squared error	Concordance Index [46]
	Balanced accuracy [47]	Root-mean-squared error	Censoring-Corrected Concordance Index [48]
	Brier score [49]	Explained variance	Integrated Brier Score [51]
	Matthews correlation coefficient [50]	Median absolute error	
	Sensitivity		
	Specificity		

with few hyperparameters. For high-dimensional hyperparameter spaces, a grid search is no longer efficient. Random search [55] or sequential model-based optimisation [56, 57] is more efficient alternatives. The optimal model hyperparameters are then used to create a final model from the available development samples.

13.2.6 Model Evaluation and Interpretation

Model evaluation shows whether a model has acceptable performance characteristics and whether it generalises well. Models are evaluated on validation samples, e.g. from an external validation dataset. A comparison with development samples may moreover indicate the presence of overfitting and insufficient data heterogeneity in the development data.

There are at least three areas that should be evaluated for a model: model discrimination, model calibration, and model benefit. In addition, model stratification should be assessed for survival endpoints.

An assessment of model discrimination shows how well the model can predict the outcome of samples, and whether it discriminates better than at random. This is done by comparing the predicted outcome with the observed outcome using one or more appropriate metrics (Table 13.1). Note that many metrics for categorical endpoints, that are commonly used in clinical settings, are sensitive to class imbalances in the underlying samples, e.g. sensitivity, specificity, and accuracy [58]. These metrics should be interpreted with caution.

Even though a model may discriminate well, this does not mean that it is well-calibrated [59]. Well-calibrated models have the ability to accurately predict class probability (categorical endpoint), survival probability at a time T (survival endpoint), or value (numerical endpoint) for each sample. For example, a well-calibrated

NTCP model can be used to accurately estimate the probability of the considered radiotoxicity. A well discriminating but not well-calibrated model is capable of distinguishing between samples with and without toxicity, but the predicted probabilities do not correspond to those observed.

Another important part of model evaluation is a comparison with existing models or the clinical standard, or if these do not exist, with null or random models. If a new model is to be translated to the clinic, it should improve upon existing alternatives in terms of predictive power or cost. Additionally, clinical usefulness can be assessed using decision–curve analysis [60–62]. This analysis can be used to determine whether a model would improve decision-making.

Also, the ability of the model to stratify patients into risk groups is clinically relevant and should be assessed [63]. For this purpose, one or more thresholds are determined from the development data and used to form different risk strata. The difference between these strata can then be evaluated by an appropriate significance test [64].

The assessments discussed above only describe model characteristics. Another important aspect of modelling, one that is often overlooked, is model provenance. Many complex modelling algorithms are black boxes in practice. Understanding why an algorithm came to a certain prediction is relevant for any clinical model because it may point out particular biases or incompleteness of the model [65]. The following aspects of a model can be investigated, though this list is not final:

1. **What Is the Importance of Each Feature for the Model?** This can be answered in different ways. For example, in regression models, individual coefficients can be evaluated, e.g. odds ratios for logistic models. A model-agnostic approach expresses feature importance by comparing the discriminatory performance of the developed model between the given dataset and a dataset in which the considered feature is randomly permuted [66].
2. **How Does Each Feature Affect the Outcome?** Explaining how the outcome depends on a feature value may help to elucidate non-linear behaviour or to illustrate potential biases in the model, i.e. feature values that lead to unexpected outcome values. The relationship between a feature and the outcome may, for example, be illustrated by partial dependence plots [67] or individual conditional expectation plots [68].
3. **Which Features Are Similar?** Similar features, such as highly correlated ones, contain mostly the same information. Newly identified important features for a particular outcome should be compared for similarity with established features.

13.2.7 Model Application

After successful evaluation and potential further prospective validation, models may be applied to identify patient subgroups, for example in interventional clinical trials that test the efficacy of treatment modification. Stratified block randomisation, taking the most important confounders into account, should be preferred but may not always be applicable, as discussed in Sect. 13.4. A suitable primary endpoint and the final statistical test have to be chosen, e.g. accounting for competing risks, censored data, and patient drop-out. For sample size planning, a realistic estimate of the expected effect and variability in the primary endpoint is decisive. Monitoring should be performed following good clinical practice including site initiation, interim monitoring, and closeout. Standard operating procedures and procedures for homogenised data acquisition and storage need to be defined in case of several participating centres in order to avoid site-specific bias and missing data. Advanced biomarker-specific trial designs may enhance the success probability of the trial and combine the steps described above [69, 70].

13.3 Introduction to TCP and NTCP Models

In this section, we introduce classic TCP and NTCP models and outline their application in biological treatment plan optimisation and evaluation.

13.3.1 Poisson Model of Tumour Control Probability

Tumour control probability models are used in radiotherapy to estimate the probability of an effective tumour treatment with the planned dose. Common TCP models assume that tumour control is achieved when no single clonogenic cell of the tumour survives after irradiation. They are often based on the linear-quadratic model, which describes the surviving fraction SF of an original cell population irradiated with dose D by

$$\mathrm{SF} = e^{-\left(\alpha D + \beta D^2\right)} \tag{13.1}$$

Here, α and β are tissue-specific parameters describing the mechanisms of cell damage [71]. Combining the surviving fraction SF with the number N_0 of clonogens per tumour before irradiation, the average number of surviving clonogens per tumour $N_0 \cdot \mathrm{SF}$ is obtained. Since the elimination of cells by radiation is a random process and the probability of single cells to survive is low, TCP can be approximated by a Poisson distribution for the case of zero surviving clonogens. The standard model of tumour control is [72]

$$\mathrm{TCP} = e^{-N_0 \cdot \mathrm{SF}} \tag{13.2}$$

This function describes a sigmoidal curve increasing from 0 to 100% with increasing dose. It can be characterised by the dose TCD_{50}, at which 50% of the tumours are controlled, and by the normalised dose–response gradient (or slope) γ_{50}, defining the steepness of the TCP curve at the 50% response level. Under a single-hit assumption ($\beta = 0$), the Poisson TCP model can be quantified by [73]

$$\mathrm{TCP} = 2^{-\exp\left[\frac{2}{\ln 2}\gamma_{50}\left(1 - \frac{D}{\mathrm{TCD}_{50}}\right)\right]} \tag{13.3}$$

To determine TCD_{50} and γ_{50}, clinical studies with varying prescribed dose but fixed number of fractions or dose per fraction have been conducted for several tumour entities. These parameters were tabularised, e.g. by Okunieff et al. [74]. Extensions of this model including tumour repopulation, incomplete repair, hypoxia, and non-uniform dose distributions were considered [75–77].

13.3.2 Modelling of Normal Tissue Complication Probability

NTCP models aim to predict the probability of complications based on the dose distribution in associated irradiated organs. For this purpose, the three-dimensional dose distribution is often reduced to a few simple metrics that can be derived from a dose-volume histogram (DVH). Some of the different methods for modelling clinical outcome data of retrospective patient cohorts and their dose distributions are described as follows [78].

1. **DVH-Reduction Models:** Based on the data published by Emami et al. [79], the empirical Lyman-Kutcher-Burman (LKB) model was developed. The LKB model describes the dose-response as a function of irradiated volume by reducing the DVH to a single metric to estimate model parameters for specific OARs [80–83]. The model includes TD_{50}, m and n as parameters. The parameter $\mathrm{TD}_{50}(V)$ is the tolerance dose for uniform irradiation of a partial volume V of an OAR at which 50% of patients are likely to experience a specific toxicity. The parameter m represents the slope at the steepest part of the dose-response curve. The parameter n describes the volume effect of the investigated OAR [84]. Serially structured organs such as the spinal cord show $n \approx 0$, while parallel organs are characterised by $n \approx 1$. Taking fractional irradiation into account, the LKB-NTCP model for a uniform dose D to a volume V of an OAR is given by

$$\mathrm{NTCP}_{\mathrm{LKB}} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(\frac{u^2}{2}\right) \mathrm{d}u \tag{13.4}$$

$$\text{with } t = \frac{D - \mathrm{TD}_{50}(V)}{m\mathrm{TD}_{50}}, \tag{13.5}$$

and

$$\mathrm{TD}_{50}(V) = \frac{\mathrm{TD}_{50}(V_{\mathrm{OAR}})}{V^n}, \tag{13.6}$$

where V_{OAR} represents the entire volume of the considered OAR.

However, dose distributions to OARs are non-uniform. The inhomogeneous dose distribution can be reduced to a single metric that produces the same probability of a given side effect as a corresponding uniform dose distribution. Such a metric is the widely used generalised equivalent uniform dose gEUD given by

$$g\text{EUD} = \left(\sum_i v_i D_i^a\right)^{1/a} \tag{13.7}$$

where D_i is the dose defined for each dose bin i in a differential DVH. v_i is the volume in a dose bin i and a is a volume parameter that is equivalent to $1/n$. This 'homogeneous' dose can then be applied as $D = g\text{EUD}$ in the LKB model in Eq. (13.5).

2. **Tissue-Architecture Models**: These more mechanistic models are based on the functional architecture of the tissue by introducing functional subunits of an OAR. These can be anatomical substructures, such as nephrons of the kidney, or the largest cell group that still functions as long as it comprises a surviving clonogen [78]. These functional subunits can be arranged in serial or parallel order, or in a combination of both. In parallel organs, functional subunits are performing rather independently so that side effects occur after the irradiated volume exceeds a critical value. Side effects that arise from irradiation of parallel organs depend on the mean dose deposited in these organs (e.g. liver, lung, or kidney). Källman et al. [85] suggested the relative seriality model, in which an organ consists of several serial and parallel structures whose reaction is described by Poisson statistics. The volume effect is characterised by a parameter s indicating the relative seriality of the organ, i.e. the proportion of serial subunits of an organ. A serial organ is characterised by large values ($s \approx 1$) and parallel organs by small values ($s \ll 1$). Other models based on the assumption that NTCP can be determined by functional subunits are for example the critical volume model [86] or the critical element model [87].
3. **Multiple-Metric Models:** The above-mentioned models predict the complication probability for one specific side effect based on the dose to a corresponding OAR. However, some complications are caused by the irradiation of different OARs, e.g. swallowing dysfunction following the irradiation of superior pharyngeal constrictor muscle and the supraglottic larynx [88] or heart valvular dysfunction by the irradiation of heart and lung [89]. To correct for this in LKB models, an interaction gEUD variable for both OARs can be introduced [89]. Moreover, side effects may also be related to dose-independent clinical parameters, such as age, radiation technique, gender, or chemotherapy [88, 90]. Multivariable logistic regression models are appropriate to include both clinical and dosimetric parameters. They are defined by

$$\text{NTCP}_{\text{Logistic}} = \left(1 + \mathrm{e}^{-g(x)}\right)^{-1}, \tag{13.8}$$

$$\text{with } g(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{13.9}$$

Here, β_i denote model coefficients and x_i are the p individual explanatory variables.

Example: Logistic NTCP models for acute side effects after cranial proton-beam therapy were developed and validated in independent patient cohorts treated at three different proton therapy centres, based on methodology described in Sect. 13.2 [91]. Alopecia grade ≥ 2 showed a strong association to the dose–volume parameter *D*5% of the skin in repeated cross-validation performed on the development cohort (AUC = 0.82, Fig. 13.4a). The corresponding NTCP model (Table 13.2) was applied to the two remaining validation cohorts, which showed similar AUC values (0.77 and 0.85, Fig. 13.4b). While the calibration slopes were close to one in validation, the intercept deviated from zero, possibly due to centre-specific differences in toxicity assessment (Fig. 13.4c).

13.3.3 Application: Biological Treatment Plan Optimisation and Evaluation

During the last decades, fluence modulated beam delivery techniques, such as intensity-modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT), successively replaced conventional 3D-conformal radiotherapy (3D-CRT). The greatest benefit of these inverse planning techniques is the multiplicity of dynamically adjustable machine parameters, allowing the creation of highly conformal treatment plans. In contrast to 3D-CRT, dose distributions in OARs can be adjusted to a much larger degree. Additionally, hardware and computing technologies evolved rapidly. Hence, more complex dose calculation algorithms could be translated into clinical routine.

To account for these developments, more and more advanced approaches for creating and evaluating treatment plans have to be designed. One of these approaches currently discussed among clinicians and medical physicists is biological treatment planning. This approach replaces the commonly used physical dose–volume parameters, which are only surrogates for biological effects, with biological measures during treatment plan optimisation and evaluation.

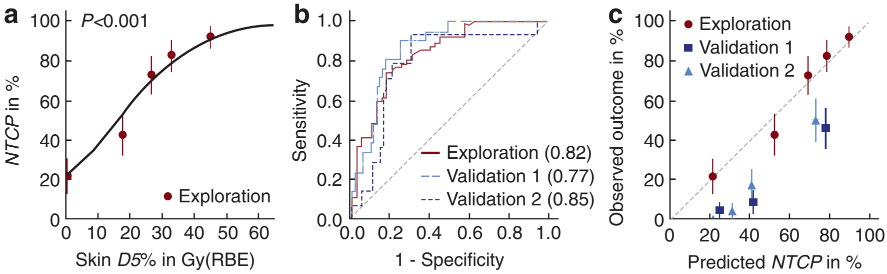

Fig. 13.4 NTCP models for acute alopecia grade ≥2 after cranial proton-beam therapy. (**a**) regression curve, (**b**) receiver operating characteristic curves, and (**c**) calibration plot are displayed. AUC values for each cohort are given in brackets. Data points and error bars represent mean and standard deviation of patient sets. Adapted from Dutz et al. [91]

Table 13.2 NTCP model for acute alopecia grade ≥2 after cranial proton-beam therapy, from Dutz et al. [91]

Model parameters	Model coefficients	(95% confidence interval)	*p*-value
Skin *D*5% in Gy(RBE)	0.081	(0.05 – 0.11)	<0.001
Constant	−0.94	(−2.91 to −0.27)	

RBE: relative biological effectiveness

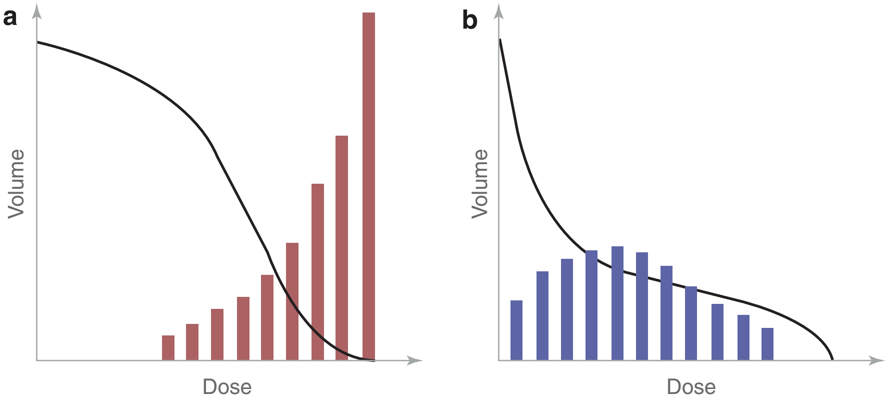

Fig. 13.5 Concept of biological optimisation for (**a**) a serial organised OAR and (**b**) a parallel organised OAR. The DVH curve is influenced by a single *g*EUD objective that achieves the same volume effect as multiple physical dose–volume objectives represented by the bars. Their different weights are expressed by the bar lengths. Adapted from [94]

For biological treatment plan optimisation, one approach is to use multivariable NTCP models directly in the objective function [92]. More common in modern treatment planning systems are optimiser functions that implement a biological objective, e.g. based on *g*EUD (Eq. 13.7), as the main optimisation parameter that adjusts the DVH curve as a whole instead of several physical dose-volume objectives [93]. For serial OARs, a high volume parameter ($a > 10$) is used to prevent dose maxima, while for parallel OARs, a low parameter value ($a = 1$) is used to reduce the mean dose [94]. In contrast to single physical dose-volume objectives (e.g. $D_{\max}$), biological objectives influence the entire DVH curve, see Fig. 13.5.

In order to adequately apply these biological functions, the tissue-specific parameter a has to be known for all OARs. Using the relationship $a = 1/n$, it can be determined from published LKB models, e.g. in Luxton et al. [95]. Before application in clinical routine, these parameters should be calibrated on clinic-specific data. In case a cannot be calibrated, different generic initial values depending on the type of OAR (parallel or serially organised) have been recommended to be used for plan optimisation, e.g. outlined in the AAPM task group report 166 [93]. However, the use of these non-calibrated initial parameters requires an additional uncertainty analysis. For treatment plan selection and evaluation, EUD can be used to rank tentative treatment plans. Also, TCP/NTCP models can be used to make patient-specific predictions of outcome and then select a specific treatment plan. Although several dose–response models have been developed and are continuously updated, they continue to be very simplistic [96]. For some clinical situations and tumour entities, several competing models may be available. For example, the predictions of models for the same side effect but developed based on data from different tumour entities may differ (e.g. lung or heart toxicity for lung and breast cancer patients). Hence, clinicians should determine the appropriate biological models for each tumour entity, clinical setting, and radiation modality.

Some treatment planning systems may already include a library of models or model parameters with default values. However, these published models have been developed at other institutions including different patient populations, treatment planning systems, dose calculation algorithms, fractionation schemes, etc. The patient characteristics may differ substantially such that further variables may affect the considered endpoint. Thus, TCP/NTCP models used in biological plan evaluation have to be calibrated based on the institutional situation before use. This requires a comprehensive collection of outcome data and large patient cohorts. If multivariable models are to be implemented that contain additional clinical variables (e.g. comorbidities, age, or concomitant therapies), this information must also be available. Since this complex calibration for different tumour entities and OARs is not feasible for most clinics, partial biological optimisation is currently used, combining biological and physical objectives.

The main limitation of biological treatment planning lies in the uncertainties of the biological models. Due to the increasing amount of patient-specific data and the development of advanced modelling strategies, a reduction of these uncertainties seems feasible. This would allow for implementing such biological techniques widely into clinical practice in the future.

13.4 Case 1: Patient Selection for Proton-Beam Therapy: The Model-Based Approach

One example for the application of NTCP models is the patient assignment to proton-beam therapy (PBT). Although the number of operating PBT facilities is increasing worldwide, the high technical and time expenditure leads to high costs of this treatment modality. Therefore, it is important to offer PBT to those patients who may benefit most from it compared to conventional photon therapy (XRT).

Randomised controlled trials (RCT) are considered as the highest evidence for practice change in oncology. However, there are challenges in performing RCTs to compare different radiotherapy techniques or modalities. The heterogeneity between centres in terms of treatment planning systems, quality assurance, training skills, image guidance techniques, treatment adaptation, immobilisation strategies, etc., may be so pronounced that it may be difficult to generalise results from RCTs into clinical routine [97, 98]. In addition, for trials comparing PBT and XRT in terms of reduced late side effects, the considered endpoints may manifest many years after radiotherapy. Thus, results from large long-term RCTs may be obsolete as radiotherapy (and PBT in particular) is still a rapidly evolving technology [99]. To reduce the number of patients and thus the duration of the study, a proper preselection of eligible patients is necessary.

A feasible approach to meet these challenges and to identify patients suitable for PBT is based on comparative NTCP modelling, the so-called model-based approach [99]. In the Netherlands and Denmark, it has already been implemented in clinical practice for patients with various tumour sites, including HNSCC, non-small cell lung cancer, breast cancer, and mediastinal lymphoma. This section discusses the principles of the two-phase model-based approach as proposed by Langendijk et al. [99] and applications.

13.4.1 Principles of the Model-Based Approach

The model-based approach consists of two phases: model-based selection and validation. The first phase, in turn, comprises three steps: development and validation of NTCP models, patient-specific plan comparison, and estimation of the clinical benefit of PBT. The individual steps are explained in more detail below.

13.4.1.1 Phase α: Model-Based Selection

Patients are selected according to their reduction of side effect probabilities under PBT compared to XRT. If this reduction exceeds a given threshold, those patients will be suitable for PBT treatment. The side effect probabilities for each patient are estimated using NTCP models.

1. **Development and Validation of NTCP Models:** NTCP models have to be developed and externally validated for different entities and relevant side effects that may occur following XRT or PBT. General aspects on development and validation of NTCP models are described in Sect. 13.3.2. Most NTCP models have been derived from data of patients treated with XRT. NTCP models can already differ between various XRT techniques [88, 100, 101]. Since dose distributions of XRT and PBT may show even stronger differences, XRT-based models need to be validated on prospectively collected PBT patient data. In case of negative validation, the development of technique-specific NTCP models may become necessary. Continuous NTCP validation and updating may be implemented, for example, in the framework of a rapid learning health care system [102, 103].
2. **Individual in silico Planning Comparative Studies:** For each patient, two treatment plans are created, one with protons and the other with a state-of-the-art XRT technique. The values of the dosimetric parameters that are supposed to be important in the selected NTCP models should be reduced, if possible, during treatment planning.
3. **Estimation of the Clinical Benefit:** The in silico treatment plans and NTCP models are used to estimate the difference (ΔNTCP) in side effect probabilities between XRT and PBT:

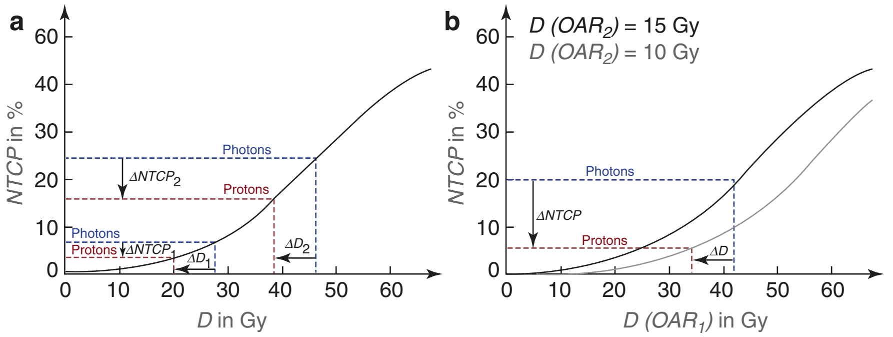

Fig. 13.6 Model-based approach according to [99]. (**a**) An equal dose difference between proton and photon treatment plan may translate into different NTCP reductions in a univariable model. (**b**) In a multivariable model including dosimetric predictors of two different OARs, the NTCP difference may be even higher if PBT is able to reduce dose to both OARs simultaneously

$$\text{”}\,\mathrm{NTCP} = \mathrm{NTCP}_{\mathrm{XRT}} - \mathrm{NTCP}_{\mathrm{PBT}} \tag{13.10}$$

Figure 13.6a shows that a similar dose difference between a photon and a proton treatment plan may lead to different NTCP reductions, depending on the slope of the NTCP curve at the considered dose values. In a multivariable model including dosimetric predictors of two different OARs, the difference in NTCP between the proton and photon plan may be even higher if PBT is able to spare both OARs simultaneously [88], see Fig. 13.6b. A patient is finally selected for PBT if the extent of NTCP reduction in the PBT plan compared to XRT exceeds a given threshold. This threshold depends on the severity of the side effects, with lower thresholds for more severe toxicities. For toxicities of CTCAE grade 2, 3, and 4–5, the Dutch Society of Radiation Oncology suggests thresholds of 10%, 5%, and 2% points, respectively [97]. In some cases, multiple side effects are considered in the selection procedure (NTCP profiles). Here, both the NTCP difference of every single endpoint as well as the summarised NTCP difference for all considered endpoints must exceed different thresholds [97]. If the NTCP difference remains below the recommended threshold, the patient is treated with state-of-the-art XRT.

13.4.1.2 Phase *β*: Model-Based Clinical Evaluation

The initial hypothesis of reduced side effects after PBT compared to XRT is evaluated during model-based clinical validation. Patients who were selected for PBT during phase α are enrolled in prospective clinical evaluation studies and are treated with the proton treatment plan created during step 2. The finally observed toxicity rates of patients treated with PBT are then compared to the initially predicted proton NTCP values to detect possible shortcomings of the applied NTCP models [97]. Furthermore, it can be tested whether the observed toxicity rate following PBT is indeed lower than the estimated NTCP values for XRT (calibration in the large [104]).

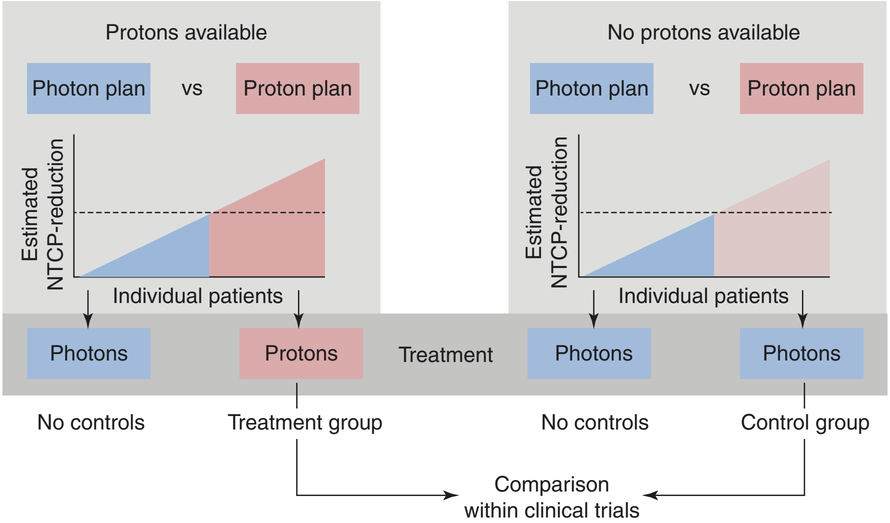

Fig. 13.7 Schematic overview of model-based clinical evaluation according to [99]

Moreover, the real outcome of PBT and XRT can be compared directly using prospectively collected patient data from cohorts treated with one of the treatment modalities (Fig. 13.7). Both patient groups of such clinical trials need to be selected according to the same selection procedure. The control group includes patients who would have been candidates for PBT but were still treated with XRT, e.g. historical cohorts [99] or patients treated in radiotherapy centres without access to PBT.

13.4.2 Application: Proton-Beam Therapy for Head and Neck Cancer

The model-based approach for patient selection for PBT has been introduced into clinical practice in some European countries. Arts et al. [105] investigated the impact of treatment accuracy, in terms of setup and range uncertainties, on the selection procedure based on a cohort of 78 patients with oropharyngeal cancer. They analysed the number of patients selected for PBT based on four NTCP models and using the above-mentioned ΔNTCP thresholds of 10% and 5% points for grade 2 and grade 3 side effects, respectively (Table 13.3). To analyse the impact of the treatment accuracy, three different planning target volume (PTV) margins for IMRT plans as well as five different setup and range robustness settings for intensity modulated proton-beam therapy (IMPT) were applied. In a setting of a 3 mm PTV margin for IMRT and 3 mm setup and 3% range error for IMPT, a total of 77% of patients were selected for PBT if the corresponding threshold was exceeded in at least one of the four NTCP models. For all models, the more robust the IMPT plans were for the same PTV margin, the fewer patients were selected for

Table 13.3 Investigated NTCP models predicting side effects following the irradiation head and neck cancer patients in the study by Arts et al. [105]

Side effect and NTCP model	Severity grade	Time after RT	Model predictors	Model type
Tube feeding dependence Wopken et al. [106]	3	6 months	Mean dose of the superior and inferior PCM, contralateral parotid, and cricopharyngeal muscle Advanced T stage Weight loss (moderate/severe) Accelerated radiotherapy Chemoradiation Radiotherapy plus cetuximab	Logistic regression model
Reduced parotid flow Dijkema et al. [107]	2	1 year	Mean dose in parotid glands	LKB model
Patient-rated problems swallowing solid food[a] Christianen et al. [88]	2	6 months	Mean dose superior PCM and supraglottic larynx Age	Logistic regression model
Patient-rated xerostomia[a] Beetz et al. [108]	2	6 months	Mean dose contralateral parotid gland Baseline xerostomia score	Logistic regression model

LKB Lyman–Kutcher–Burman; *PCM* pharyngeal constrictor muscle; *RT* radiotherapy

[a] Assessed with the head-and-neck cancer-specific quality of life questionnaire EORTC QLQ-H&N35

PBT. With the same robustness settings of the IMPT plan, more patients were selected for PBT the greater the PTV margin of the IMRT plan. The study by Arts et al. [105] showed that, in addition to the choice of an appropriate threshold for each severity grade, treatment accuracy also affects the proportion of patients selected for PBT.

13.5 Case 2: Radiomics

Medical imaging is commonly acquired prior to and during radiation treatment. It may contain information on disease diagnosis or treatment outcome and thereby improve corresponding TCP or NTCP models. It can thus enable further treatment personalisation, e.g. by selecting patients for specific treatments [109]. In a radiomics analysis, information is extracted from each image and quantitatively assessed. Radiomics draws upon mathematically well-defined ('hand-crafted') features, automated feature generation based on deep learning algorithms, or both

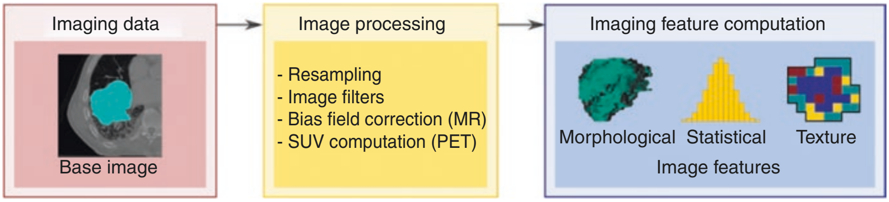

Fig. 13.8 Schematic representation of image processing and feature calculation. *MR* magnetic resonance, *SUV* standardised uptake value, *PET* positron emission tomography. Adapted from [29]

[110, 111]. Radiomics has been applied to numerous tasks in radiation oncology including TCP and NTCP modelling for several tumour entities [112–115].

13.5.1 The Radiomics Workflow

A radiomics analysis using hand-crafted features consists of several steps, as illustrated in Fig. 13.8 and explained in the following.

1. **Image Acquisition and Reconstruction:** A patient is scanned in a medical imaging device according to a specific protocol. Software, usually provided by the vendor, then converts the acquired image data into something interpretable by human readers.
2. **Segmentation:** This usually aims to characterise part of an image, e.g. the primary tumour or different OARs. Clinical experts or (semi-)automatic algorithms segment or delineate the image to identify the regions of interest (ROIs).
3. **Image Processing:** Image processing primarily harmonises images across patients. For example, voxels (3D pixels) in images of different patients are resampled to the same dimensions to decrease variability of radiomics features related to different voxel sizes in the reconstructed images [116]. Another component of image processing is the use of image filters, e.g. to remove noise or emphasise edges, blobs, or directional structures [117]. A general image processing scheme for radiomics is described by the Image Biomarker Standardisation Initiative (IBSI) [118].
4. **Radiomics Feature Computation:** After image processing, radiomics features are computed from the ROI. This generates a feature value for each image. Many common features were standardised by the IBSI, and described in their documentation [118].
5. **Modelling:** The previous steps yield radiomics feature values that are easily converted to a tabular format. The modelling component of a radiomics analysis is therefore not specific to radiomics but follows the principles described earlier in this chapter.

Radiomics based on deep learning is quite similar, but the radiomics feature computation and modelling steps are typically replaced entirely by a deep learning algorithm. Manual segmentation may not be required, as a deep learning network is capable of learning what aspects and regions of the image are of interest, given sufficient data. Some image processing is usually required because of constraints on the input of deep learning algorithms.

13.5.2 Application: Radiomics for Adaptive Treatment

One way to personalise radiotherapy is to monitor the treatment progress and adapt treatment correspondingly [4]. Treatment progress may potentially be monitored by medical imaging and its comprehensive radiomics analysis. This particular subfield of radiomics is called delta-radiomics because radiomics features are computed from images acquired at different time points [119, 120].

We have previously performed a delta-radiomics analysis to assess whether computed tomography (CT) imaging during treatment can be used to classify patients with locally advanced HNSCC into a high and a low-risk group for loco-regional recurrence [121]. We will describe this study here as an example of how radiomics could be used for adaptive treatment.

The study involved three patient cohorts, a development cohort ($n = 48$), a validation cohort ($n = 30$), and a cohort that was only used to assess the robustness of radiomics features ($n = 18$). The patients in the development and validation cohort were followed up for several years, and loco-regional recurrence was recorded. Patients in these cohorts were scanned prior to treatment (CT_0) and during the second week of treatment (CT_2). Based on the primary tumour contours, we computed 1583 radiomics features with the IBSI-compliant software MIRP [122] for every imaging dataset.

We identified 269 robust features, which we computed from CT_0 and CT_2. Furthermore, we computed the difference between the two time points, i.e. 269 delta features. The three feature sets were compared for modelling loco-regional control (LRC) in a TRIPOD type 3 survival analysis: using CT_0-features only, CT_2-features only, and the combined set including delta features. Modelling followed the steps outlined in Sect. 13.2. Features in the development cohort were pre-processed by standardisation and subsequent clustering of similar features (Spearman correlation $\rho > 0.90$). We then determined variable importance by performing feature selection using six different methods on 1000 bootstraps of the data and aggregated the feature ranks. Subsequently, we optimised model hyperparameters for six different algorithms through grid search in a cross-validation scheme. Models were then trained on 1000 bootstraps of the development cohort and combined into an ensemble model for each combination of feature selection method and learner. In total, we created 36 ensemble models, based on earlier findings that indicated that a combination of different methods should be assessed to reduce the risk of accidental findings [29]. Furthermore, it could be assessed whether an increase in model complexity justifies a decrease in model explainability by better performance.

We then validated all models in the validation dataset, see Fig. 13.9. Model performance was assessed using a concordance index (C-index; 0.5: random, 1.0

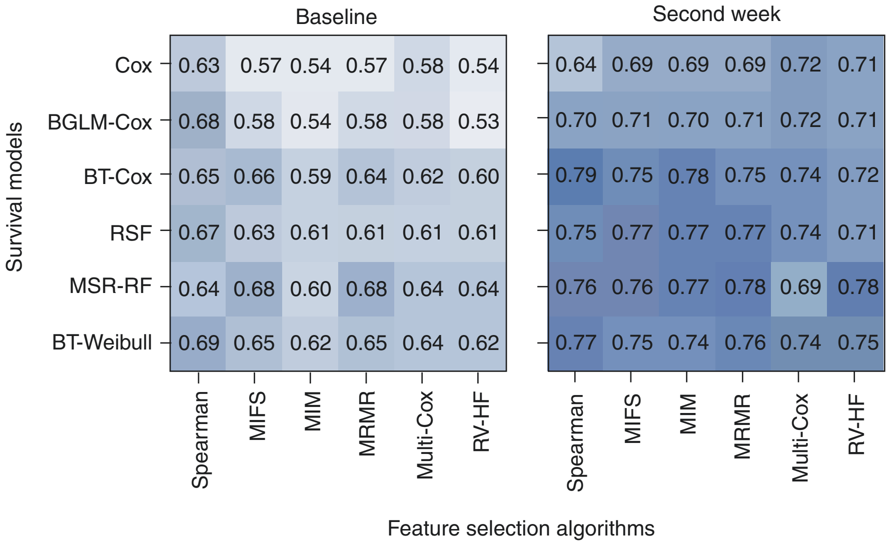

Fig. 13.9 Concordance indices of radiomics models (0.5: random, 1.0: perfect discrimination) based on treatment planning CT images (left panel) and on CT images after the second week of treatment (right panel). The performance of several survival models based on imaging features selected from different feature selection algorithms is shown for the validation cohort. For details, see [121]

perfect discrimination) [46]. Stratification into low- and high-risk groups for locoregional recurrence was evaluated using a log-rank test. We found that models based on the CT_2 (C-index: 0.73 ± 0.04, mean ± standard deviation over all models) and combined feature sets (0.70 ± 0.05, not shown) exceeded the performance of models using CT_0 only (0.62 ± 0.04). The combined feature set ($p = 0.06$) and CT_2 only ($p = 0.005$) enabled better performance compared to CT_0.

Our results indicate that imaging obtained during treatment can be more suited to identify patients at lower or higher risk of tumour recurrence than pre-treatment imaging. Though this effect should be validated in a larger dataset, the results do show the potential for image-guided treatment adaptation. For instance, if the tumour has a very low risk of recurring, treatment may be stopped early, while in case of a high recurrence risk, the patient and clinician may choose to pursue an extended radiation treatment. In the future, such options for treatment adaptation may become available for patients with a clear prognosis based on precise and validated models.

13.6 Summary and Outlook

Due to the growing amount of patient-specific data and corresponding advances in computer technology and adapted machine-learning algorithms, models predicting tumour control or normal tissue complications are becoming increasingly complex, which in turn may allow for more accurate predictions. This brings forward new fields of application, e.g. in personalised radiotherapy, for model-based patient selection or biological treatment planning.

It is thus essential to understand the basic principles of model development and validation, which we have presented in this chapter. We outlined important aspects of data quality, data pre-processing, feature selection, model development, model evaluation, and model validation. The application of these concepts was presented for NTCP modelling within the model-based approach selecting patients for photon or proton-beam therapy and for adaptive TCP modelling based on radiomics analyses from pre-treatment and in-treatment CT imaging.

In future, data science and artificial intelligence may play a central role in the development of high-precision radiotherapy. For these developments, homogeneous patient cohorts of sufficient sample size are required. This necessitates the formation of large cooperative networks pooling their data or federated learning strategies with decentralised data storage [123]. Furthermore, data publication according to the FAIR principles [124] will ensure the continued improvement of models on radiation treatment outcome.

References

1. Holthusen H. Erfahrungen über die Verträglichkeitsgrenze für Röntgenstrahlen und deren Nutzanwendung zur Verhütung von Schäden. Strahlentherapie. 1936;57:254–69.
2. Karger CP. Klinische Strahlenbiologie. In: Schlegel W, Karger CP, Jäkel O, editors. Medizinische Physik: Grundlagen—Bildgebung—Therapie—Technik. Berlin, Heidelberg: Springer; 2018. p. 451–72.
3. Baumann M, Krause M, Overgaard J, et al. Radiation oncology in the era of precision medicine. Nat Rev Cancer. 2016;16:234–49. https://doi.org/10.1038/nrc.2016.18.
4. Ajdari A, Niyazi M, Nicolay NH, et al. Towards optimal stopping in radiation therapy. Radiother Oncol. 2019;134:96–100.
5. Beaton L, Bandula S, Gaze MN, Sharma RA. How rapid advances in imaging are defining the future of precision radiation oncology. Br J Cancer. 2019;120:779–90.
6. Zwanenburg A, Löck S. Why validation of prognostic models matters? Radiother Oncol. 2018;127:370–3.
7. van Smeden M, de Groot JA, Moons KG, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16:163.
8. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1–73.
9. García S, Luengo J, Herrera F. Data preprocessing in data mining. Berlin: Springer International Publishing; 2015.
10. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc Series B Stat Methodol. 1964;26:211–52.
11. Yeo I, Johnson RA. A new family of power transformations to improve normality or symmetry. Biometrika. 2000;87:954–9.
12. Orlhac F, Boughdad S, Philippe C, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med. 2018;59:1321–8.
13. Orlhac F, Frouin F, Nioche C, et al. Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology. 2019;291:53–9.

14. Chatterjee A, Vallières M, Dohan A, et al. Creating robust predictive radiomic models for data from independent institutions using normalization. IEEE Trans Radiat Plasma Med Sci. 2019;3:210–5.
15. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
16. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol. 1995;142:1255–64.
17. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087–91.
18. Luengo J, García S, Herrera F. On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowl Inf Syst. 2012;32: 77–108.
19. He H, Garcia EA. Learning from imbalanced data. In: IEEE Transactions on Knowledge and Data Engineering; 2008. pp 1263–1284.
20. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5:221–32.
21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority oversampling technique. J Artif Intell Res. 2002;16:321–57.
22. He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); 2008. pp 1322–1328.
23. Kubat M, Holte R, Matwin S. Learning when negative examples abound. In: Machine learning: ECML-97. Berlin, Heidelberg: Springer; 1997. p. 146–53.
24. O'Brien R, Ishwaran H. A random forests quantile classifier for class imbalanced data. Pattern Recogn. 2019;90:232–49.
25. Zwanenburg A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. Eur J Nucl Med Mol Imaging. 2019;46:2638–55.
26. Zwanenburg A, Leger S, Agolli L, et al. Assessing robustness of radiomic features by image perturbation. Sci Rep. 2019a;9:614.
27. Cunningham JP, Ghahramani Z. Linear dimensionality reduction: survey, insights, and generalizations. J Mach Learn Res. 2015;16:2859–900.
28. Tolosi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics. 2011;27:1986–94.
29. Leger S, Zwanenburg A, Pilz K, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. Sci Rep. 2017;7:13206.
30. Park MY, Hastie T, Tibshirani R. Averaged gene expressions for regression. Biostatistics. 2007;8:212–27.
31. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Hoboken: John Wiley & Sons; 2009.
32. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
33. Li J, Cheng K, Wang S, et al. Feature selection: a data perspective. ACM Comput Surv (CSUR). 2018;50:94.
34. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507–17.
35. Haury A-C, Gestraud P, Vert J-P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS One. 2011;6:e28210.
36. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl Inf Syst. 2007;12:95–116.
37. Abeel T, Helleputte T, Van de Peer Y, et al. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. 2010;26:392–8.
38. Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc Series B Stat Methodol. 2010;72:417–73.

39. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Daelemans W, Goethals B, Morik K, editors. Machine learning and knowledge discovery in databases. Berlin, Heidelberg: Springer; 2008. p. 313–25.
40. Wald R, Khoshgoftaar TM, Dittman D, et al. An extensive comparison of feature ranking aggregation techniques in bioinformatics. In: 2012 IEEE 13th International Conference on Information Reuse Integration (IRI); 2012. pp 377–384.
41. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Stat Methodol. 1996;58:267–88.
42. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. Stat Sci. 2007;22:477–505.
43. Hofner B, Boccuto L, Göker M. Controlling false discoveries in high-dimensional situations: boosting with stability selection. BMC Bioinformat. 2015;16:144.
44. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15:3133–81.
45. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn. 2001;45:171–86.
46. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med. 2004;23:2109–23.
47. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition; 2010. pp 3121–3124.
48. Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30:1105–17.
49. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78:1–3.
50. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975;405:442–51.
51. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc. 2007;102:359–78.
52. Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc Ser A. 1972;135: 370–84.
53. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
54. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. pp 785–794.
55. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13:281–305.
56. Feurer M, Klein A, Eggensperger K, et al. Efficient and robust automated machine learning. In: Cortes C, Lawrence ND, Lee DD, et al., editors. Advances in neural information processing systems 28. New York: Curran Associates, Inc.; 2015. p. 2962–70.
57. Hutter F, Hoos HH, Leyton-Brown K. Sequential model-based optimization for general algorithm configuration. In: Coello CAC, editor. Learning and intelligent optimization. Berlin, Heidelberg: Springer; 2011. p. 507–23.
58. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal. 2002;6:429–49.
59. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
60. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak. 2008;8:53.
61. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26:565–74.

62. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18.
63. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13:33.
64. Mallett S, Royston P, Waters R, et al. Reporting performance of prognostic models in cancer: a review. BMC Med. 2010;8:21.
65. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning; 2017. arXiv [stat.ML].
66. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a Variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res. 2019;20:1–81.
67. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–232.
68. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black Box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat. 2015;24:44–65.
69. Antoniou M, Kolamunnage-Dona R, Jorgensen AL. Biomarker-guided non-adaptive trial designs in phase II and phase III: a methodological review. J Pers Med. 2017;7:1.
70. Lin J-A, He P. Reinventing clinical trials: a review of innovative biomarker trial designs in cancer therapies. Br Med Bull. 2015;114:17–27.
71. Joiner MC. Quantifying cell kill and cell survival. In: Joiner MC, van der Kogel A, editors. Basic clinical radiobiology. 4th ed. Boca Raton: CRC Press; 2009.
72. Bentzen SM. Dose–response relationships in radiotherapy. In: Joiner MC, van der Kogel A, editors. Basic clinical radiobiology. 4th ed. Boca Raton: CRC Press; 2009.
73. Warkentin B, Stavrev P, Stavreva N, et al. A TCP-NTCP estimation module using DVHs and known radiobiological models and parameter sets. J Appl Clin Med Phys. 2004;5:50–63.
74. Okunieff P, Morgan D, Niemierko A, Suit HD. Radiation dose-response of human tumors. Int J Radiat Oncol Biol Phys. 1995;32:1227–37.
75. Roberts SA, Hendry JH. A realistic closed-form radiobiological model of clinical tumor-control data incorporating intertumor heterogeneity. Int J Radiat Oncol Biol Phys. 1998;41:689–99.
76. Sanchez-Nieto B, Nahum AE. The delta-TCP concept: a clinically useful measure of tumor control probability. Int J Radiat Oncol Biol Phys. 1999;44:369–80.
77. Webb S, Nahum AE. A model for calculating tumour control probability in radiotherapy including the effects of inhomogeneous distributions of dose and clonogenic cell density. Phys Med Biol. 1993;38:653–66.
78. Gulliford S. Modelling of Normal tissue complication probabilities (NTCP): review of application of machine learning in predicting NTCP. In: El Naqa I, Li R, Murphy MJ, editors. Machine learning in radiation oncology: theory and applications. Cham: Springer International Publishing; 2015. p. 277–310.
79. Emami B, Lyman J, Brown A, et al. Tolerance of normal tissue to therapeutic irradiation. Int J Radiat Oncol Biol Phys. 1991;21:109–22.
80. Burman C, Kutcher GJ, Emami B, Goitein M. Fitting of normal tissue tolerance data to an analytic function. Int J Radiat Oncol Biol Phys. 1991;21:123–35.
81. Kutcher GJ, Burman C. Calculation of complication probability factors for non-uniform normal tissue irradiation: the effective volume method gerald. Int J Radiat Oncol Biol Phys. 1989;16:1623–30.
82. Kutcher GJ, Burman C, Brewster L, et al. Histogram reduction method for calculating complication probabilities for three-dimensional treatment planning evaluations. Int J Radiat Oncol Biol Phys. 1991;21:137–46.
83. Lyman JT. Complication probability as assessed from dose-volume histograms. Radiat Res Suppl. 1985;8:S13–9.
84. Gulliford SL, Partridge M, Sydes MR, et al. Parameters for the Lyman Kutcher Burman (LKB) model of Normal tissue complication probability (NTCP) for specific rectal complications observed in clinical practise. Radiother Oncol. 2012;102:347–51.

85. Källman P, Agren A, Brahme A. Tumour and normal tissue responses to fractionated non-uniform dose delivery. Int J Radiat Biol. 1992;62:249–62.
86. Niemierko A, Goitein M. Modeling of normal tissue response to radiation: the critical volume model. Int J Radiat Oncol Biol Phys. 1993;25:135–45.
87. Niemierko A, Goitein M. Calculation of normal tissue complication probability and dose-volume histogram reduction schemes for tissues with a critical element architecture. Radiother Oncol. 1991;20:166–76.
88. Christianen MEMC, Schilstra C, Beetz I, et al. Predictive modelling for swallowing dysfunction after primary (chemo)radiation: results of a prospective observational study. Radiother Oncol. 2012;105:107–14.
89. Cella L, Palma G, Deasy JO, et al. Complication probability models for radiation-induced heart valvular dysfunction: do heart-lung interactions play a role? PLoS One. 2014;9:e111753.
90. Wijsman R, Dankers F, Troost EGC, et al. Multivariable normal-tissue complication modeling of acute esophageal toxicity in advanced stage non-small cell lung cancer patients treated with intensity-modulated (chemo-)radiotherapy. Radiother Oncol. 2015;117:49–54.
91. Dutz A, Lühr A, Agolli L, et al. Development and validation of NTCP models for acute side-effects resulting from proton beam therapy of brain tumours. Radiother Oncol. 2019;130:164–71.
92. Kierkels RGJ, Korevaar EW, Steenbakkers RJHM, et al. Direct use of multivariable normal tissue complication probability models in treatment plan optimisation for individualised head and neck cancer radiotherapy produces clinically acceptable treatment plans. Radiother Oncol. 2014;112:430–6.
93. Li XA, Alber M, Deasy JO, et al. The use and QA of biologically related models for treatment planning: short report of the TG-166 of the therapy physics committee of the AAPM. Med Phys. 2012;39:1386–409.
94. Fogliata A, Thompson S, Stravato A, Tomatis S, Scorsetti M, Cozzi L. On the gEUD biological optimization objective for organs at risk in photon optimizer of eclipse treatment planning system. J Appl Clin Med Phys. 2018;19(1):106–14. https://doi.org/10.1002/acm2.12224.
95. Luxton G, Keall PJ, King CR. A new formula for normal tissue complication probability (NTCP) as a function of equivalent uniform dose (EUD). Phys Med Biol. 2008;53:23–36.
96. Niemierko A. Biological optimization. In: Bortfeld T, Schmidt-Ullrich R, De Neve W, Wazer DE, editors. Image-guided IMRT. Berlin, Heidelberg: Springer; 2006. p. 199–216.
97. Langendijk JA, Boersma LJ, Rasch CRN, et al. Clinical trial strategies to compare protons with photons. Semin Radiat Oncol. 2018;28:79–87.
98. Widder J, van der Schaaf A, Lambin P, et al. The quest for evidence for proton therapy: model-based approach and precision medicine. Int J Radiat Oncol Biol Phys. 2016;95:30–6.
99. Langendijk JA, Lambin P, De Ruysscher D, et al. Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach. Radiother Oncol. 2013;107:267–73.
100. Beetz I, Schilstra C, van Luijk P, et al. External validation of three dimensional conformal radiotherapy based NTCP models for patient-rated xerostomia and sticky saliva among patients treated with intensity modulated radiotherapy. Radiother Oncol. 2012b;105:94–100.
101. Troeller A, Yan D, Marina O, et al. Comparison and limitations of DVH-based NTCP models derived from 3D-CRT and IMRT data for prediction of gastrointestinal toxicities in prostate cancer patients by using propensity score matched pair analysis. Int J Radiat Oncol Biol Phys. 2015;91:435–43.
102. Lambin P, Roelofs E, Reymen B, et al. “Rapid learning health care in oncology”—an approach towards decision support systems enabling customised radiotherapy. Radiother Oncol. 2013;109:159–64.
103. Lambin P, Zindler J, Vanneste B, et al. Modern clinical research: how rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. Acta Oncol. 2015;54:1289–300.

104. 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.
105. Arts T, Breedveld S, de Jong MA, et al. The impact of treatment accuracy on proton therapy patient selection for oropharyngeal cancer patients. Radiother Oncol. 2017;125:520–5.
106. Wopken K, Bijl HP, van der Schaaf A, et al. Development of a multivariable normal tissue complication probability (NTCP) model for tube feeding dependence after curative radiotherapy/chemo-radiotherapy in head and neck cancer. Radiother Oncol. 2014;113:95–101.
107. Dijkema T, Raaijmakers CPJ, Ten Haken RK, et al. Parotid gland function after radiotherapy: the combined Michigan and Utrecht experience. Int J Radiat Oncol Biol Phys. 2010;78:449–53.
108. Beetz I, Schilstra C, van der Schaaf A, et al. NTCP models for patient-rated xerostomia and sticky saliva after treatment with intensity modulated radiotherapy for head and neck cancer: the role of dosimetric and clinical factors. Radiother Oncol. 2012a;105:101–6.
109. Morin O, Vallières M, Jochems A, et al. A deep look into the future of quantitative imaging in oncology: a statement of working principles and proposal for change. Int J Radiat Oncol Biol Phys. 2018;102:1074–82.
110. Hatt M, Le Rest CC, Tixier F, et al. Radiomics: data are also images. J Nucl Med. 2019;60:38S–44S.
111. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14:749–62.
112. van Dijk LV, Brouwer CL, van der Schaaf A, et al. CT image biomarkers to improve patient-specific prediction of radiation-induced xerostomia and sticky saliva. Radiother Oncol. 2017;122:185–91.
113. van Dijk LV, Langendijk JA, Zhai T-T, et al. Delta-radiomics features during radiotherapy improve the prediction of late xerostomia. Sci Rep. 2019;9:12483.
114. van Dijk LV, Noordzij W, Brouwer CL, et al. 18F-FDG PET image biomarkers improve prediction of late radiation-induced xerostomia. Radiother Oncol. 2018a;126:89–95.
115. van Dijk LV, Thor M, Steenbakkers RJHM, et al. Parotid gland fat related magnetic resonance image biomarkers improve prediction of late radiation-induced xerostomia. Radiother Oncol. 2018b;128:459–66.
116. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys. 2017;44:1050–62.
117. Depeursinge A, Al-Kadi OS, Ross Mitchell J. Biomedical texture analysis: fundamentals, tools and challenges. Cambridge: Academic Press; 2017.
118. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. 2020;295:328–38.
119. Carvalho S, Leijenaar RTH, Troost EGC, et al. Early variation of FDG-PET radiomics features in NSCLC is related to overall survival-the “delta radiomics” concept. Radiother Oncol. 2016;118:S20–1.
120. Cunliffe A, Armato SG 3rd, Castillo R, et al. Lung texture in serial thoracic computed tomography scans: correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development. Int J Radiat Oncol Biol Phys. 2015;91:1048–56.
121. Leger S, Zwanenburg A, Pilz K, et al. CT imaging during treatment improves radiomic models for patients with locally advanced head and neck cancer. Radiother Oncol. 2019;130:10–7.
122. Zwanenburg A, Leger S, Starke S, Löck S. Medical image radiomics processor. Version 1.0URL; 2019b. https://github.com/oncoray/mirp
123. Jochems A, Deist TM, van Soest J, et al. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital—a real life proof of concept. Radiother Oncol. 2016;121:459–67.
124. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018.