Models by their nature are mathematical approximations of reality, as conveyed by the statement that “all models are wrong but some are useful (1).” Their usefulness in radiation therapy (RT) is highlighted by the roles that outcome models play in improving the quality and efficacy of radiation treatment of tumors by predicting response, individualizing prescriptions, and optimizing and ranking planning options. These models are generally categorized into those for tumor response prediction by tumor control probability (TCP) and those for predicting radiation-induced toxicities by normal-tissue complication probability (NTCP). Traditionally, these models were simplistic and included a few dose and volume metrics summarizing the delivered treatment and baseline patient-specific prognostic risk factors, rendering them less powerful but also less prone to overfitting pitfalls (2).

Recent advances in multimodality imaging and biotechnology have improved the availability of imaging and molecular profile data, resulting in tremendous growth in patient-specific information that can be informative for guiding radiation treatments. This additional information can be incorporated into classical or new TCP/NTCP outcome models to improve their predictive power and bring them closer to providing a more realistic and possibly comprehensive representation of radiation response. Large-scale quantitative analysis of anatomic and functional imaging data has yielded the radiomics domain (3), and that of related high-throughput molecular profiles (eg, genomics, transcriptomics, proteomics, metabolomics) has delivered the radiogenomics field into radiation oncology (4). However, in this context of complex data-driven (big data) modeling, it is important to remember the parsimony...
principle: “Everything should be made as simple as possible, but not simpler (5).” Radiomics/radiogenomics models are largely phenomenological (data-driven) and more complex than traditional TCP/NTCP; therefore, their clinical value should be carefully evaluated using rigorous statistics and scientifically sound methods. This includes addressing the following 2 pertinent questions: (1) Are these more complicated models of TCP/NTCP statistically and scientifically valid? and (2) Is the gain in the predictive power/radiobiological understanding over classical models of TCP/NTCP worth the added complexity?

This editorial aims to highlight current affairs in radiomics/radiogenomics modeling, with a special focus on identifying frequent sources of uncertainty and presenting some best practices for model evaluation and validation in this RT big data era, before such models can be adopted in clinical trials or deployed in radiation oncology practice (6).

Main Sources of Uncertainty

Source data uncertainties

RT data (eg, dose, clinical endpoints, imaging, molecular markers) are usually collected in heterogeneous conditions that may lead to various uncertainties such as noise interference, calibration error, sparse data sets, inconsistent measurement, redundant variables/dimensions, wrong labeling, and missing outcomes, to name a few. An important lesson from the Quantitative Analysis of Normal Tissue Effects in the Clinic (QUANTEC) project is that the scarcity of quality outcomes data, not the weakness of the existing models, was a major impediment in developing and validating models of outcomes that may have clinical utility (7).

This issue is an even more significant problem when developing more complex models based on heterogeneous data from radiomics and radiogenomics. The complexity of a useful prediction model (eg, number of parameters or variables) is determined by the size and quality of the outcomes data used for model development. Specifically, this complexity is determined in practice by the number of cases with clinical events of interest that a model can properly fit, not by the number of features extracted from radiomics and radiogenomics data that may be lost in the process as a result of limited sample size. For instance, a fundamental challenge related to radiogenomics (eg, genome-wide associations) is the very large number of sequence alterations (~11 million single nucleotide polymorphisms) and small sample sizes available. Therefore, adjustments for multiple testing and appropriate dimensionality reduction techniques are imperative when building such radiogenomics models (8, 9), and the same is true in the case of radiomics models (10).

Model selection uncertainties

The modeling approach used for any specific TCP/NTCP endpoint may be subject to selection bias. For instance, the model complexity may be inappropriate for the RT data that were collected, resulting in an overtrained or underfitted model. Moreover, the assumed model may not faithfully reflect the original data behavior, or temporal information may not be carefully captured. This is particularly challenging in the case of radiomics/radiogenomics, where data suffer from multicollinearity (ie, the variables are intercorrelated), and model selection needs to account for minimum redundancy while maximizing relevance as part of its construction (11). To handle such data, computer algorithms based on artificial intelligence techniques such as machine learning (ML) are being frequently used instead of traditional regression models for their ability to learn from data and generalize to samples that have not been seen before (external data) by the algorithm. In such a case of ML modeling of radiomics/radiogenomics, the model is developed by first representing the regularities of observation data based on a presumed formulation of the model. Then, given a reasonable collection of training data, one would optimize a metric of interest and estimate these ML model-fitting parameters. The trained model is expected to generalize well to unseen test data; however, this cannot be guaranteed unless data uncertainties are accounted for during the training phase and the model has undergone a rigorous validation process that includes data suffering from high-uncertainty with existing noise in the data (12). Therefore, the modern RT outcome models should tackle the issue of model regularization and accommodate or compensate for a variety of uncertainties and weaknesses in the model construction by, for example, providing adequate examples to the ML algorithm of such noisy cases during the training phase.

How to Learn From Uncertainties

Although there has been tremendous growth in patient-specific information, the number of patients receiving RT has not increased proportionally, leading to limited-size high-uncertainty data sets with a lot more variables to evaluate. This trend leaves the traditional statistical analysis more susceptible to the data dredging (P-hacking or inflated significance) phenomenon with spurious associations (13). Moreover, it is necessary to ensure that the samples used for RT model development and validation are representative of the general population characteristics that are under investigation, both in number and diversity (eg, race, sex, age, staging, technique). Differences in such characteristics may define the scope that a radiomics or radiogenomics model may or may not apply. For instance, it is less likely that a radiomics model developed for predicting local control in a population of patients with early stage lung cancer treated with hypofractionation will work in another locally advanced population treated with conventional fractionation because of the different feature patterns that may be captured in the model. This does not necessarily preclude the possibility of prognostic radiomics models, independent
of the treatment type, or pan-cancer models across different sites, but this needs to be specified and evaluated as part of the original model’s design itself (14).

The fields of statistics and ML have offered several approaches for dealing with data modeling uncertainty (15). For instance, in Bayesian methods the uncertainties in data or an assumed model can be represented by prior probability distributions, and the model is calculated as a marginal likelihood by integrating over the randomness or uncertainty of parameters or algorithms. These priors can be chosen empirically based on past information or constructed mathematically based on desired statistical properties (16). Another approach to alleviate the adverse effects of uncertainty and other quality issues in radiomics/radiogenomics data is to apply data integration methodologies, which has been particularly effective in “omics” applications using networks or graph-based models. This is generally successful because the uncertainty and other quality issues in one data set are allayed by the complementary information in the others and can reinforce the consensus signal for modeling the clinical endpoint of interest in the various data sets. Because of these advantages, integrative approaches have been effective for addressing several biomedical problems, such as biomarker discovery (17), molecular interactions (18), and RT outcomes with radiomics and radiogenomics data (19, 20).

Model Validation

An essential part of the model inference process, particularly when dealing with more complex models of radiomics or radiogenomics, is to perform a rigorous validation process to assess the goodness of an outcome model and its ability to generalize to external (out-of-sample) data. Goodness can be defined in many terms, such as prediction performance, model complexity, and model robustness (e.g., Akaike/Bayesian information criterion, false discovery rate, or free energy principle). A conventional method to select or reject a model for a given outcome and set of data is to conduct a likelihood ratio test to measure the goodness-of-fit as in the case of traditional TCP/NTCP models (21). However, this may not be plausible in the case of more complex radiomics or radiogenomics models with ML and heterogeneous data sets, where the goal is redefined to balance bias (model fit) with variance (generalizability to out-of-sample data). Therefore, methods based on statistical resampling (e.g., cross-validation and bootstrapping) are more appropriate (22). More formally, the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) recommendation developed a 22-item checklist for the reporting of studies developing, validating, or updating a predictive model, such as radiomics/radiogenomics RT outcome models (23). The checklist includes items related to study design and data quality, in addition to issues related to validation and performance reporting of multivariate models. For validation, the TRIPOD recommendations emphasize the importance of internal validation (e.g., resampling) as a necessary part of model development and external validation on an independent test set(s), which is the gold standard for evaluating generalizability and can be done centrally by collecting all the data into 1 database or by using distributed learning to facilitate data sharing (24).

Another critical component of the evaluation process is the metric(s) used to measure the (prediction) performance of these models. For continuous/real-valued outcomes, the most commonly used metrics for this regression purpose are mean squared error and (Pearson’s or Spearman’s) correlation coefficients. In the case of discrete-valued outcomes (classes/labels), where the model inference task is termed classification, the most commonly used metrics are accuracy, cross-entropy error, and area under the receiver operating characteristic curve score (25). However, in many scenarios, the numbers of samples included in the various classes may be (severely) imbalanced, such as when counting the number of patients with a specific type of cancer versus healthy individuals in the general population. Because accuracy and area under the receiver operating characteristic curve score are not reliable in such scenarios, class-specific measures sensitive to class imbalance, especially the precision-recall-F-measure trio, are recommended (25, 26).

Conclusions

The emergence of radiomics and radiogenomics provides new opportunities to develop more informative outcome models for RT. However, by the same token, these opportunities present profound new challenges for mitigating uncertainties and validating these models rigorously for clinical trials or daily practice. The best existing “classical” models include dose and volume metrics summarizing the delivered treatment and baseline patient-specific prognostic risk factors. Indices extracted from radiomics and radiogenomics big data are not expected to replace those conventional model covariates, but rather, if incorporated, may add to the model’s predictive power and its understanding. Therefore, specific demonstrations and more rigorous validation are needed to guarantee that these new models are robust with the ultimate judgment being the clinical assessment of benefits versus costs.

References


