HHS Public Access

Author manuscript

Int J Radiat Oncol Biol Phys. Author manuscript; available in PMC 2020 August 05.

Published in final edited form as:

Int J Radiat Oncol Biol Phys. 2018 November 15; 102(4): 1070–1073. doi:10.1016/j.ijrobp.2018.08.022.

Radiotherapy outcomes models in the era of radiomics and radiogenomics: Uncertainties and validation

Issam El Naqa, PhD[1,*], **Gaurav Pandey, PhD**[2], **Hugo Aerts, PhD**[3], **Jen-Tzung Chien, PhD**[4], **Christian Nicolaj Andreassen, MD PhD**[5], **Andrzej Niemierko, PhD**[6], **Randall K. Ten Haken, PhD**[1]

[1]Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan, USA.

[2]Department of Genetics and Genomic Sciences and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[3]Departments of Radiation Oncology & Radiology Dana-Farber Cancer Institute Brigham and Women's Hospital, Harvard Medical School, Boston, USA.

[4]Department of Electrical and Computer Engineering and the Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan.

[5]Departments of Clinical Medicine and Experimental Clinical Oncology, Aarhus, Denmark.

[6]Department of Radiation Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, USA.

Abstract

Recent advances in imaging and biotechnology have tremendously improved the availability of quantitative imaging (radiomics) and molecular data (radiogenomics) for radiotherapy patients. This big data development with its comprehensive nature promises to transform outcome modeling in radiotherapy from few dose-volume metrics into utilizing more data-driven analytics. However, it also presents new profound challenges and creates new tasks for alleviating uncertainties arising from dealing with heterogeneous data and complex big data analytics. Therefore, more rigorous validation procedures need to be devised for these radiomics/radiogenomics models compared to traditional outcome modeling approaches previously utilized in radiation oncology, before they can be safely deployed for clinical trials or incorporated into daily practice. This editorial highlights current affairs, identifies some of the frequent sources of uncertainties, and presents some of the recommended practices for radiomics/radiogenomics models' evaluation and validation.

Keywords

outcome models; radiomics; radiogenomics; big data; uncertainties; model validation

[*]**Corresponding author**: Issam El Naqa, University of Michigan, Ann Arbor, 48103, MI, USA.

Conflict of Interest: None.

Introduction

Models by their own nature are mathematical approximations of reality as conveyed in the statement that "*All models are wrong but some are useful*." Their usefulness in radiotherapy (RT) is highlighted by the roles outcome models perform to improve the quality and efficacy of radiation treatment of tumors by predicting response, individualizing prescriptions, 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 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 [1].

Recent advances in multi-modality 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 anatomical and functional imaging data has yielded the *radiomics* domain [2], while that of related high-throughput molecular profiles (genomics, transcriptomics, proteomics, metabolomics, etc.) has delivered the *radiogenomics* field into radiation oncology [3]. However, in this context of complex data-driven (*big data*) modeling, it's important to remember the parsimony principle "*Everything should be made as simple as possible, but not simpler*." 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 two pertained questions: (i) whether these more complicated models of TCP/NTCP are statistically and scientifically valid? And (ii) 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 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 into radiation oncology practice[4].

Main sources of uncertainty

Source data uncertainties

RT data (e.g., dose, clinical endpoints, imaging, molecular markers) are usually collected in heterogeneous conditions that may lead to various uncertainties like noise interference, calibration error, sparse datasets, inconsistent measurement, redundant variables/dimensions, wrong labeling and missing outcomes, to name a few. An important lesson from the QUANTEC project is that the sparsity 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 [5]. This issue is even more significant problem when developing more complex models based on heterogeneous data from radiomics and radiogenomics. The complexity of a useful prediction model (e.g., 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, and 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 (e.g., genome-wide associations) is the very large number of sequence alterations (~11 million SNPs) and small sample sizes available. Therefore, adjustments for multiple testing and appropriate dimensionality reduction techniques are imperative when building such radiogenomics models [6,7] and the same is true in the case of radiomics models [8].

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 over-trained or under-fitted 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 the data suffers from multi-collinearity (i.e., the variables are inter-correlated) problem, and model selection needs to account for minimum redundancy while maximizing relevance as part of its construction [9]. To handle such data, computer algorithms based on artificial intelligence techniques such as machine learning (ML) are being frequently utilized 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 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 to ensure that the ML model can handle existing noise in the data[10]. 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 providing adequate examples to the ML algorithm of such noisy cases during the training phase, for instance.

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 datasets 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 [11]. 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 (e.g., race, sex, age, staging, technique, etc.). Differences in such characteristics may define the scope that a radiomics or radiogenomics may or may not apply. For instance, it is less likely that a radiomics model developed for predicting local control on an early stage lung cancer patients' population treated with hypofractionation will work on another locally advanced population treated with conventional fractionation, due to the different feature patterns that may be captured in the model. This does not necessarily preclude the fact that there can be 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 [12].

The fields of statistics and machine learning have offered several approaches for dealing with data modeling uncertainty [13]. 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 [14]. Another approach to alleviate the adverse effects of uncertainty and other quality issues in radiomics/radiogenomics data by applying *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 dataset 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 datasets. Due to these advantages, integrative approaches have been effective for addressing several biomedical problems, such as biomarker discovery [15], molecular interactions [16], as well as RT outcomes with radiomics and radiogenomics data [17,18].

Model validation

An essential part of the model inference process, particularly, when dealing with more complex models of radiomics or radiogenomics, is performing 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/or 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 by conducting a likelihood ratio test to measure the goodness-of-fit as in the case of traditional TCP/NTCP models [19]. However, this may not be plausible in the case of more complex radiomics or radiogenomics models with ML and heterogeneous datasets, 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 [20]. 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 [21]. 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 the model development and external validation on an independent test set(s), which is the gold standard for evaluating generalizability, which can be done centrally by collecting all the data into one database or by utilizing distributed learning to facilitate data sharing [22].

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 (MSE) 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 (ROC) Curve (AUC) score [23]. 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 of a specific type of cancer versus healthy individuals in the general population. Since accuracy and AUC score are not reliable in such scenarios, class-specific measures sensitive to class imbalance, especially the *precision-recall-F-measure* trio, are recommended [23,24].

Conclusions

The emergence of radiomics and radiogeneomics provides new opportunities to develop more informative outcome models for radiotherapy. However, by the same token, these opportunities present new profound 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 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 judgement being clinical assessment of benefits versus costs.

References

[1]. El Naqa I A guide to outcome modeling in radiotherapy and oncology: Listening to the data. Boca Raton, FL: CRC Press: Taylor & Francis Group, 2018; 368

[2]. Lambin P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer 2012;48:441–446. [PubMed: 22257792]

[3]. Rosenstein BS, et al. Radiogenomics: Radiobiology enters the era of big data and team science. Int J Radiat Oncol Biol Phys 2014;89:709–713. [PubMed: 24969789]

[4]. Benedict SH, El Naqa I Klein EE. Introduction to big data in radiation oncology: Exploring opportunities for research, quality assessment, and clinical care. International Journal of Radiation Oncology • Biology • Physics;95:871–872.

[5]. Bentzen SM, et al. Quantitative analyses of normal tissue effects in the clinic (quantec): An introduction to the scientific issues. International journal of radiation oncology, biology, physics 2010;76:S3–S9.

[6]. Andreassen CN, et al. Radiogenomics - current status, challenges and future directions. Cancer Lett 2016;382:127–136. [PubMed: 26828014]

[7]. Oh JH, et al. Computational methods using genome-wide association studies to predict radiotherapy complications and to identify correlative molecular processes. Sci Rep 2017;7:43381. [PubMed: 28233873]

[8]. Yip SS Aerts HJ. Applications and limitations of radiomics. Physics in medicine and biology 2016;61:R150–166. [PubMed: 27269645]

[9]. Hanchuan P, Fuhui L Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 2005;27:1226–1238. [PubMed: 16119262]

[10]. Hastie T, Tibshirani R Wainwright M. Statistical learning with sparsity : The lasso and generalizations. Boca Raton: CRC Press, Taylor & Francis Group, 2015.

[11]. Head ML, et al. The extent and consequences of p-hacking in science. PLOS Biology 2015;13:e1002106. [PubMed: 25768323]

[12]. Aerts HJWL, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications 2014;5:4006.

[13]. Aggarwal CC Philip SY. A survey of uncertain data algorithms and applications. IEEE Transactions on Knowledge and Data Engineering 2009;21:609–623.

[14]. Gelman A Bayesian data analysis. Boca Raton: CRC Press, 2014.

[15]. Sorani MD, et al. Clinical and biological data integration for biomarker discovery. Drug Discov Today 2010;15:741–748. [PubMed: 20558318]

[16]. Pandey G, et al. An integrative multi-network and multi-classifier approach to predict genetic interactions. PLoS Comput Biol 2010;6.

[17]. Luo Y, et al. Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via bayesian network analysis. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 2017;123:85–92. [PubMed: 28237401]

[18]. Luo Y, et al. Development of a fully cross-validated bayesian network approach for local control prediction in lung cancer. IEEE Transactions on Radiation and Plasma Medical Sciences 2018:1–1. [PubMed: 29930991]

[19]. Steyerberg EW. Clinical prediction models : A practical approach to development, validation, and updating. New York, NY: Springer, 2009.

[20]. Arlot S Celisse A A survey of cross-validation procedures for model selection. Statistics surveys 2010;4:40–79.

[21]. Collins GS, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement. Annals of Internal Medicine 2015;162:55–63. [PubMed: 25560714]

[22]. Jochems A, 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. Radiotherapy and Oncology 2016;121:459–467. [PubMed: 28029405]

[23]. Lever J, Krzywinski M Altman N. Classification evaluation. Nature Methods 2016;13:603.

[24]. Saito T Rehmsmeier M The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10:e0118432. [PubMed: 25738806]