Title
Machine learning and modeling: Data, validation, communication challenges.

Permalink
https://escholarship.org/uc/item/3mm0d3c7

Journal
Medical physics, 45(10)

ISSN
0094-2405

Authors
El Naqa, Issam
Ruan, Dan
Valdes, Gilmer
et al.

Publication Date
2018-10-01

DOI
10.1002/mp.12811

Peer reviewed

Machine learning and modeling: Data, validation, communication challenges

Issam El Naqa[a)]
Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA

Dan Ruan
Department of Radiation Oncology, University of California Los Angeles, Los Angeles, CA, USA

Gilmer Valdes
Department of Radiation Oncology, University of California Los San Francisco, San Francisco, CA, USA

Andre Dekker
GROW-School for Oncology and Developmental Biology, Department of Radiation Oncology (MAASTRO), Maastricht University Medical Center, Maastricht, The Netherlands

Todd McNutt
Department of Radiation Oncology, John Hopkins University, Baltimore, MD, USA

Yaorong Ge
Department of Software and Information Systems, University of North Carolina, Charlotte, NC, USA

Q. Jackie Wu
Department of Radiation Oncology, Duke University Medical Center, Durham, NC, USA

Jung Hun Oh and Maria Thor
Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Wade Smith
Department of Radiation Oncology, University of Washington, Seattle, WA, USA

Arvind Rao
Department of Radiation Oncology, MD Anderson, Houston, TX, USA
Department of Bioinformatics and Computational Biology, MD Anderson, Houston, TX, USA

Clifton Fuller
Department of Radiation Oncology, MD Anderson, Houston, TX, USA

Ying Xiao
Department of Radiation Oncology, University of Pennsylvania, Philadelphia, PA, USA

Frank Manion, Matthew Schipper, Charles Mayo, Jean M. Moran, and Randall Ten Haken
Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA

(Received 13 September 2017; revised 28 December 2017; accepted for publication 22 January 2018; published 24 August 2018)

With the era of big data, the utilization of machine learning algorithms in radiation oncology is rapidly growing with applications including: treatment response modeling, treatment planning, contouring, organ segmentation, image-guidance, motion tracking, quality assurance, and more. Despite this interest, practical clinical implementation of machine learning as part of the day-to-day clinical operations is still lagging. The aim of this white paper is to further promote progress in this new field of machine learning in radiation oncology by highlighting its untapped advantages and potentials for clinical advancement, while also presenting current challenges and open questions for future research. The targeted audience of this paper includes newcomers as well as practitioners in the field of medical physics/radiation oncology. The paper also provides general recommendations to avoid common pitfalls when applying these powerful data analytic tools to medical physics and radiation oncology problems and suggests some guidelines for transparent and informative reporting of machine learning results. © *2018 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.12811]

Key words: big data, machine learning, radiation oncology

1. INTRODUCTION

Machine learning (ML) embraces an evolving branch of computational algorithms that were originally designed to emulate living beings' intelligence by learning from the surrounding environment. The term was coined by Arthur Samuel in his seminal work in the 1950s where he described machine learning as "a field of study that gives computers the ability to learn without being explicitly programmed."[1] Machine learning as a branch of the artificial intelligence field draws upon ideas from diverse disciplines such as probability and statistics, information theory, psychology, control theory, and philosophy.[2–4] It has been successfully applied to many different fields including pattern recognition,[4] computer vision,[5] spacecraft engineering,[6] finance,[7] computational biology,[8,9] and medical applications.[10,11] Developed ML algorithms are currently considered one of the main workhorses in the new era of *Big Data* to potentially overcome challenges related to the excessive burden of manual curation, data veracity, and the analysis of complex patterns. In this sense, ML algorithms can both add to and complement traditional statistical modeling methods.

Machine learning could be further subdivided per the nature of the data labeling into: supervised, unsupervised, and semi-supervised.[3,6,12] Supervised learning is used to estimate an unknown (input, output) mapping from known (input, output) samples, where the output is "labeled" (e.g., classification or regression). This is the most commonly used approach in radiotherapy applications such as planning evaluation or outcomes prediction using known labels provided by experts or clinical endpoints. In unsupervised learning, only input samples are given to the learning system and inferences are drawn without labeled responses (e.g., clustering and estimation of probability density function [PDF]) such as visualization of higher dimensional data, some respiratory motion management studies, and contouring, which has typically been based on clustering methods and is currently trending toward supervised deep learning.[13] Semi-supervised learning is a combination of both supervised and unsupervised learning methods. The part of the data, which is labeled, could be used to infer the unlabeled portion (e.g., text/image retrieval systems) through transductive learning, or to induce the general mapping from input to output by inductive learning. Additionally, in semi-supervised learning unlabeled data could be used to infer high-order representations of data to aid the supervised component of the learning task[14] with application examples such as interactive prostate segmentation[15] and xerostomia (dry mouth) prediction in head and neck cancer.[16]

Although there are several ongoing efforts to provide guidelines for developing and reporting ML results,[17,18] with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement receiving wide range endorsements for predictive modeling,[17] there are yet no universal consensus recommendations for ML in general or in the setting of medical physics and radiation oncology specifically. This white paper aims to (a) further promote progress in the new ML field in radiation oncology by highlighting its untapped advantages and potential for clinical advancement to newcomers; (b) present current challenges and open questions for further research by newcomers and practitioners; and (c) provide general recommendations to active researchers to avoid common pitfalls and suggest guidelines for transparent and informative reporting of ML results for medical physics and radiation oncology applications.

1.A. Use case examples in radiation oncology

In recent years, ML has witnessed an increased use in radiation oncology with focused sessions at the annual meetings of the American Association of Physicists in Medicine (AAPM). However, initial applications of ML in radiotherapy have started in the mid 1990s by training artificial neural networks (ANNs) for automating treatment planning evaluation,[19] beam orientation customization,[20] and standardization (knowledge-based planning),[21] for instance. Later applications in the mid-2000s focused on predicting normal tissue toxicity in different sites.[22–24] Currently, these methods are applied to many aspects of radiation oncology including: tumor response modeling,[25–31] radiation physics quality assurance (QA),[32] auto-segmentation for normal tissue and target delineation,[33–36] treatment planning,[37–39] image-guided radiotherapy (RT),[40,41] and respiratory motion management.[42,43] Details about these and other applications are reviewed in the literature.[44] Further and future applications of ML may also expand into:

- Identifying potential hardware and software safety- and quality-related risks prior to treating a patient.
- Using ML-aided decision support systems to improve the efficiency and the consistency of current diagnosis and treatment tools, and subsequently raising average physician performance during residency training or clinical practice.
- Identifying and analyzing underlying pan-omics (images, genetics, dosimetric indices, and clinical information) data for patient-specific treatment regime stratification (drug-RT and combined therapy) and predicting radiotherapy outcomes.
- Relating images/genetics to outcomes, and identifying latent pheno-/geno-/image prognostic features.
- Complementing existing response models with better learning of data-derived information, for outlier analysis, hypothesis modification, and model refinement.
- Conducting clinical trials using ML algorithms as a guidance for optimal treatment strategies.[45]

1.B. Recommendations for general ML application in radiation oncology

For successful application of ML approaches in general and in medical physics and radiation oncology in particular, there are five main issues that need to be cautiously considered.

1.B.1. First, characterize the problem properly

One needs to properly represent the problem at hand in terms of the input/output data, the desired results, assumptions made, and the interpretation of their associated outputs in relation to specific clinical goals. Care needs to be taken to minimize the risk of false positive findings or overfitting the data, via multiple testing adjustments, false discovery analysis, or other methods to avoid data dredging or p-hacking problems.[46] Similar considerations should inform which metrics are used to evaluate performance, as they will be used to judge suitability of the model. For example, standard maximization of the area under a receiver operating characteristic curve (AUC) is a convenient and easy-to-comprehend metric. However, it assumes that specificity and sensitivity are of equal importance to the decision-maker, which may not always be the case for physicians or physicists working in the clinic.[47–50] Furthermore, ascribing unwarranted clinical significance sometimes to results based on marginal AUC values (e.g., <0.7), does occur often and should be cautiously addressed in scientific publications and presentations.[51,52] ML provides a model approximation of reality (correlations), to make clinically relevant predictions with associated error characteristics. This should be reported and is appropriate for the decision space and ensures that results are more likely to be used robustly.

1.B.2. Second, include sufficient data volume and quality in training

It cannot be overemphasized that ML algorithms are data-driven approaches and their performances are intrinsically dependent on the data provenance, volume and quality assurance of training data, and outlier identification.[53] Assembly of large patient datasets containing both treatment parameters and outcomes to investigate linkages using ML can be a significant challenge. ML applications generally perform better with more training data, particularly as more input/output variables are added and the model complexity is increased. The goal of an ML study is to learn and understand the training data interdependencies and potentially generalize based on them. However, inference on causation is not directly attainable with the most popular algorithm. For instance, in the classical application of ML for modeling radiation-induced toxicities, it is understood that radiation is the main causative agent. However, the exact role of other variables (co-variates), both clinical and biological, beyond dose-effect modification or complementarity may require further experimentation or gathering new variables that were not included in the original analysis. This is, in a sense, quite similar to inference in traditional statistics. However, ML methods are well suited in such predictive modeling scenarios because of the following: (a) the flexibility and inherent ability of many ML algorithms to navigate complex high-dimensional data space and identify nonlinear/nonmonotonic patterns (e.g., via kernel mapping or nonlinear activation functions); and (b) many ML algorithms can also measure oversensitivity to data or identify possible "gaps" in the modeling process, that is, areas where the model actually failed to fit the data. An example is shown in modeling outcomes of radiotherapy with support vector machines (SVM), where many of the training data points (dose-volume metrics) were located in the "margin" region between the classifier boundaries indicating missing discriminant information from the used data based on this particular model.[54]

1.B.3. Third, model parsimony and generalizability

To be useful, the model needs to generalize beyond the training observation into out-of-sample data. To achieve this goal, the model generally needs to be kept as simple as possible but not simpler. This property, known as *parsimony*, follows from Occam's razor that states "among competing hypotheses, the hypothesis with the fewest assumptions should be selected."[55] However, deep learning algorithms with their large number of layers for learning data representation and performing model prediction in the same architecture, may present a future challenge to this classical notion,[56] but the overall objective remains the same, that is, to achieve generalizability to out-of-sample data. This could be evaluated using resampling methods (cross-validation or bootstrapping), bias-variance trade-offs (Cramer-Rao) or analytically by using complexity measures such as Vapnik–Chervonenkis (VC) dimension, for instance.[57] External validation of models in cohorts, which were acquired independently from the discovery cohort (e.g., from another Radiation Oncology department) is still considered the gold standard for true estimates of performance and generalizability of prediction models. For example, models for optimal organ sparing in treatment planning can be evaluated using cross-institutional data, which can ensure that the training data represents the general practice and also provides generalizability of the model.[58] Finally, data guarantees that ensure equivalence of the training and testing datasets are essential for robust model evaluation and application. Given new published models being considered for clinical use in critical decisions, the medical physics community should take a leading role to treat these models as medical devices including formal acceptance and commissioning to ensure that the right algorithm or model are applied to the right application and that the model results make sense in a given clinical situation.

1.B.4. Fourth, quality assurance of ML algorithm selection

The set of machine learning algorithms and associated public-domain implementations are expanding at a rapid pace with several open-source platforms. Application of different algorithms to the same dataset may yield variable results for predictors found to be significantly associated with the outcome of interest.[18,20] However, this may also suggest a potential limitation of self-critical assessment of published ML models or realistic confidence levels with implications for their practical clinical value. Typically, the best model is the simplest model with the fewest assumptions, following the parsimony principle mentioned above,[55] however, the selected model should also include estimates of its uncertainties (confidence levels) that can be performed analytically or using statistical resampling methods (e.g., bootstrapping).[59–61] Also, there are issues related to interpreting or combining results of different ML algorithms or for defining criteria for objectively selecting the approach best suited to particular clinical investigation in radiation oncology. More generally, post model selection inference is an important topic with relevance to ML methods, which should be considered.[62] Standard data analysis often ignores the model selection step and as a result overstates the significance of the findings by ignoring the uncertainty associated with such model selection, which should be reported.[63]

1.B.5. Fifth and finally, make models and/or results intuitive

A major limitation in the acceptance of ML by the larger medical community has been hailed as the "black box" stigma, where the ML algorithm maps a given input data to output predictions without providing any additional insight into the system mapping. That is, providing an intuitive interpretation of the learned process could be missing, which impedes clinical practitioners from better understanding their data and entrusting the ML model predictions.[44] Interpretability is also important in generating new knowledge, hypotheses, and in identifying biomarkers that could guide treatment prescription or technology design by a ML response model, for example. Another example is in the case of organs-at-risk (OAR) dose-volume histogram (DVH) prediction models for treatment planning, where further interpretation of the ML results indicated that the main factors affecting the mean value and slope of the DVH curve were related to the mean distance and the slope of the distance to the target. Such analysis corroborated prior intuitions and studies that attempted to link patient geometry to planning results, and helped with the understanding of the ML results.[64] Although there are inherently interpretable ML algorithms, for instance decision trees, Bayesian networks, or generalized linear models (e.g., logistic regression), they are usually outperformed in terms of accuracy by ensemble methods or deep neural nets (for large datasets).[12] The aversion to black box models in medicine goes beyond the instinctive fear of being the first adopter of new technologies. For instance, a "black box" neural net that was developed to infer whether patients with pneumonia could be discharged from a hospital was found to inadvertently label asthmatic patients as low risk.[65–67] Due to the nature of the training data used, this mistake could have not been fixed without using an interpretable model or deeper understanding of the modeling results.[65–67] The development of accurate and interpretable models is an active area of research and recent progress has been made using different ML archetictures.[67–71] This area of research requires special attention from the ML community working in biomedicine generally and radiation oncology specifically for the sake of machine learning algorithms to gain the broader acceptance they deserve.[72] In addition, while ML results (predictors) for disease and toxicity outcomes have the potential to improve physician decision-making, information overload is an emerging issue as practitioners have increasing amounts of information available.[73,74] Incorporating results into a decision support tool, which intelligently can synthesize many types and many sources of information is likely to facilitate increased adoption of new ML results.[75,76]

1.C. Open issues and suggestions for ongoing ML research in radiation oncology

There are many ongoing issues related to applying ML as part of the clinical workflow or prospective clinical trial designs that need further consideration by the research community. These include but are not limited to:

- Access to and standardization of the radiation oncology pan-omics data (clinical, dosimetric, imaging, etc.) and allowing interactive learning/labeling strategies to further enrich such datasets. This is currently being aided by task group efforts such as TG-263.[77]
- Maintaining high data quality requirements and the ability to train the ML under realistic clinical scenarios with noisy conditions, especially when dealing with Big data, for instance. This is aided by efforts in the community to publish reusable datasets such as the Medical Physics Dataset Article (MPDA) efforts.
- Development of robust methods to quantitate the impact of incomplete, adverse, and uncertain labeling on model predictions and associated performance guarantees.
- Address inconsistency issues related to hierarchal fitting of heterogeneous vs homogenous datasets.
- Development of accurate and interpretable algorithms. However, as noted by Breiman, "Framing the question as the choice between accuracy and interpretability is an incorrect interpretation of what the goal of a statistical analysis is. The goal is not interpretability, but accurate information."[78] A balance between these two issues may be, nevertheless, needed for broader clinical acceptance or to correct spurious correlations when important cofounders are missing from the training data.
- Standardizing the validation process (Internally only, Internally and externally) by adopting recommendations from the TRIPOD guidelines,[17] ML practitioners,[18] or developing own medical physics/radiation oncology-specific guidelines.
- Evaluating and applying ML not only as related to research topics but also within clinical practice such as daily quality assurance checks.
- Extending the application of rapid sharing and distributed learning paradigms[79] as more data become available from everyday clinical practice.
- Development of robust methods to incorporate results into existing clinical decision-making practices.[75]
- Developing methods for sample size estimation for using ML (e.g., learning curves),[80] which is currently underutilized for most applications and would be useful

when incorporating ML into clinical trial designs,[45] where scarce resource of patients willing to enroll and ethical issues limit the studied population size.

- Determine what evidence is available to substantiate inferential claims when p-values are not available for testing significance of the variables inside the model (e.g., for random forests or penalized regression methods such as Elastic Net). This is an active area of research and progress has been made in some instances but more work remains (e.g., Elastic Net Cox models).[62,81]

1.D. Recommendations for publications related to ML

This is an important time period in the early emerging history of application of ML and AI into health care data. Authors of papers have the opportunity in each instance, not only to present the results of their particular model but also to shape expectations in the community for how results should be evaluated, communicated, and applied. A few recommendations are:

- Identify why and what criteria were used to choose the ML algorithm. Over time, certain methodologies may be preferred for specific applications.
- Define and apply proper criteria for evaluating ML results as presented, for instance, by Kang et al.[72] or discussed by Japkowicz et al.[82] Generally, ML performance is evaluated empirically using learning curves, information theoretic techniques, and statistical resampling methods.[12]
- Authors of ML studies should discuss conformance to criteria such as those above and adopt the TRIPOD checklist for more informative and transparent reporting.[17] The limitations of the authors' methodology and implications of those limitations should be openly and collegially discussed.
- Construct publically available benchmark datasets with known interactions and include checks of algorithms' sensitivity and specificity in identifying these interactions. Such datasets can currently undergo a peer-review authentication and formally published receiving a unique Digital Object Identifier (DOI) as offered by MPDA and others that would provide a necessary description of the dataset and its potential usage. With the availability of these open access benchmark datasets, publications on applications of ML approaches to clinical data could also include application of the algorithms to the benchmark dataset to define a context for assessment of uncertainties. This may enable demonstration that the algorithm can find the "known" answer, before asserting its ability to find the unknown answers.
- Publications on ML models should generally meet known statistical standards in the literature for the number of patients, fraction of events used, presentation of scientific evidence, and balancing statistical and clinical significance.[45,83] Any substantial deviations should be rigorously addressed as an issue to avoid p-hacking or false discovery pitfalls in the proposed approach.
- When using a resampling technique (cross-validation or bootstrapping) to estimate the predictive performance of a model, it is critical that all aspects of the analysis (including selection of tuning parameters, variable selection, and model specification) have undergone proper bias-variance correction processes to mitigate bias from training data (overfitting) while still achieving similar performance on unseen data (low variance) between train and test distributions).[84,85] However, it is noted that this process may vary from one ML algorithm to another. For example, a radiomics analysis with large number of variables and small sample size applied information gain for feature selection and estimated performance based on the .632 + estimator, which has lower bias-variance compromise between training and testing errors compared to generic cross-validation or bootstrapping.[86]
- When reporting on new biomarkers (e.g., gene expression or radiomics feature), it is desirable to contrast the predictive performance of a model based only on standard clinical factors as a benchmark to allow the reader to understand how much the new biomarkers would improve prediction performance over current models.

2. CONCLUSIONS

Application of ML algorithms in radiotherapy is witnessing tremendous resurgence with the rapid increase in patient-specific information and the data generated from all aspects of the radiotherapy processes. This paper highlighted the many potential opportunities for ML in the medical physics and radiation oncology future and some of the current challenges. It also provided general recommendations to get the most out of these powerful tools and avoid common pitfalls as well as some guidelines were suggested for useful publication of ML results.

ACKNOWLEDGMENTS

The authors especially thank the anonymous three reviewers who have done a tremendous work to assist improving the presentation and the quality of this white paper.

[a] Author to whom correspondence should be addressed. Electronic mail: ielnaqa@med.umich.edu

REFERENCES

1. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev.* 1959;3:210–229.
2. Mitchell TM. *Machine Learning*. New York, NY: McGraw-Hill; 1997.
3. Alpaydin E. *Introduction to Machine Learning*. 3rd ed. Cambridge, MA: The MIT Press; 2014.

4. Bishop CM. *Pattern Recognition and Machine Learning*. New York, NY: Springer; 2006.
5. Apolloni B. *Machine Learning and Robot Perception*. Berlin: Springer-Verlag; 2005.
6. Ao S-I, Rieger BB, Amouzegar MA. *Machine Learning and Systems Engineering*. Dordrecht, NY: Springer; 2010.
7. Györfi LS, Ottucsák GG, Walk H. *Machine Learning for Financial Engineering*. Singapore; London: World Scientific; 2012.
8. Mitra S. *Introduction to Machine Learning and Bioinformatics*. Boca Raton: CRC Press; 2008.
9. Yang ZR. *Machine Learning Approaches to Bioinformatics*. Hackensack, NJ: World Scientific; 2010.
10. Cleophas TJ. *Machine Learning in Medicine*. New York, NY: Springer; 2013.
11. Malley JD, Malley KG, Pajevic S. *Statistical Learning for Biomedical Data*. Cambridge: Cambridge University Press; 2011.
12. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*. Vol 1: Berlin: Springer series in statistics Springer; 2001.
13. Lustberg T, vanSoest J, Gooding M, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother Oncol*. 2018;126:312–317.
14. Chapelle O, Schlkopf B, Zien A. *Semi-Supervised Learning*. Cambridge, MA: The MIT Press; 2010.
15. Park SH, Gao Y, Shi Y, Shen D. Interactive prostate segmentation using atlas-guided semi-supervised learning and adaptive feature selection. *Med Phys*. 2014;41:111715.
16. Soares I, Dias J, Rocha H, Khouri L, Do Carmo Lopes M, Ferreira B. Semi-supervised self-training approaches in small and unbalanced datasets: application to xerostomia radiation side-effect. In Kyriacou E, Christofides S, Pattichis CS, eds. *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016: MEDICON 2016, March 31st-April 2nd 2016, Paphos, Cyprus*. Cham: Springer International Publishing; 2016:828–833.
17. Collins GS, Reitsma JB, Altman DG, Moons KM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Ann Intern Med*. 2015;162:55–63.
18. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18:e323.
19. Willoughby TR, Starkschall G, Janjan NA, Rosen II. Evaluation and scoring of radiotherapy treatment plans using an artificial neural network. *Int J Radiat Oncol Biol Phys*. 1996;34:923–930.
20. Rowbottom CG, Webb S, Oldham M. Beam-orientation customization using an artificial neural network. *Phys Med Biol* 1999;44:2251–2262.
21. Wells DM, Niederer J. A medical expert system approach using artificial neural networks for standardized treatment planning. *Int J Radiat Oncol Biol Phys* 1998;41:173–182.
22. Gulliford SL, Webb S, Rowbottom CG, Corne DW, Dearnaley DP. Use of artificial neural networks to predict biological outcomes for patients receiving radical radiotherapy of the prostate. *Radiother Oncol*. 2004;71:3–12.
23. Munley MT, Lo JY, Sibley GS, Bentel GC, Anscher MS, Marks LB. A neural network to predict symptomatic lung injury. *Phys Med Biol*. 1999;44:2241–2249.
24. Su M, Miften M, Whiddon C, Sun X, Light K, Marks L. An artificial neural network for predicting the incidence of radiation pneumonitis. *Med Phys*. 2005;32:318–325.
25. NAqA IE, Deasy JO, Mu Y, et al. Datamining approaches for modeling tumor control probability. *Acta Oncol*. 2010;49:1363–1373.
26. Valdes G, Solberg TD, Heskel M, Ungar L, Simone CB II. Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy. *Phys Med Biol*. 2016;61:6105.
27. El Naqa I, Bradley J, Blanco AI, et al. Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors. *Int J Radiat Oncol Biol Phys*. 2006;64:1275–1286.
28. Oberije C, Nalbantov G, Dekker A, et al. A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: a step toward individualized care and shared decision making. *Radiother Oncol*. 2014;112:37–43.
29. Bradley J, Deasy JO, Bentzen S, El Naqa I. Dosimetric correlates for acute esophagitis in patients treated with radiotherapy for lung carcinoma. *Int J Radiat Oncol Biol Phys*. 2004;58:1106–1113.
30. Hope AJ, Lindsay PE, El Naqa I, et al. Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters. *Int J Radiat Oncol Biol Phys*. 2006;65:112–124.
31. Zhou Z, Folkert M, Cannon N, et al. Predicting distant failure in early stage NSCLC treated with SBRT using clinical parameters. *Radiother Oncol*. 2016;119:501–504.
32. Kalet AM, Gennari JH, Ford EC, Phillips MH. Bayesian network models for error detection in radiotherapy plans. *Phys Med Biol*. 2015;60: 2735–2749.
33. Valdes G, Scheuermann R, Hung C, Olszanski A, Bellerive M, Solberg T. A mathematical framework for virtual IMRT QA using machine learning. *Med Phys*. 2016;43:4323–4334.
34. Valdes G, Morin O, Valenciaga Y, Kirby N, Pouliot J, Chuang C. Use of TrueBeam developer mode for imaging QA. *J Appl Clin Med Phys*. 2015;16:322–333.
35. Carlson JN, Park JM, Park S-Y, Park JI, Choi Y, Ye S-J. A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. *Phys Med Biol*. 2016;61:2514.
36. Li Q, Chan MF. Predictive time-series modeling using artificial neural networks for Linac beam symmetry: an empirical study. *Ann N Y Acad Sci*. 2017;1387:84–94.
37. Moore KL, Brame RS, Low DA, Mutic S. Experience-based quality control of clinical intensity-modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys*. 2011;81:545–551.
38. Wu B, Ricchetti F, Sanguineti G, et al. Data-driven approach to generating achievable dose-volume histogram objectives in intensity-modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys*. 2011;79:1241–1247.
39. Wu B, Ricchetti F, Sanguineti G, et al. Patient geometry-driven information retrieval for IMRT treatment plan quality control. *Med Phys*. 2009;36:5497–5505.
40. Guidi G, Maffei N, Meduri B, et al. A machine learning tool for re-planning and adaptive RT: a multicenter cohort investigation. *Physica Med*. 2016;32:1659–1666.
41. Guidi G, Maffei N, Vecchi C, et al. A support vector machine tool for adaptive tomotherapy treatments: prediction of head and neck patients criticalities. *Physica Med*. 2015;31:442–451.
42. Ruan D, Keall P. Online prediction of respiratory motion: multidimensional processing with low-dimensional feature learning. *Phys Med Biol*. 2010;55:3011–3025.
43. Isaksson M, Jalden J, Murphy MJ. On using an adaptive neural network to predict lung tumor motion during respiration for radiotherapy applications. *Med Phys*. 2005;32:3801–3809.
44. El Naqa I, Li R, Murphy MJ, eds. *Machine Learning in Radiation Oncology: Theory and Application*. Switzerland: Springer International Publishing; 2015. https://doi.org/10.1007/978-3-319-18305-3.
45. Kohannim O, Hua X, Hibar DP, et al. Boosting power for clinical trials using classifiers based on multiple biomarkers. *Neurobiol Aging*. 2010;31:1429–1442.
46. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of P-hacking in science. *PLoS Biol*. 2015;13: e1002106.
47. Smith WP, Phillips MH. Comment on “ROC analysis in patient specific quality assurance” [Med. Phys. 40(4), 042103 (7 pp.) (2013)]. *Med Phys*. 2015;42:4411–4412.
48. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach Learn*. 2009;77:103–123.
49. Phillips MH, Smith WP, Parvathaneni U, Laramore GE. Role of positron emission tomography in the treatment of occult disease in head-and-neck cancer: a modeling approach. *Int J Radiat Oncol Biol Phys*. 2011;79:1089–1095.
50. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *Eur Radiol*. 2015;25:932–939.
51. Rice ME, Harris GT. Comparing effect sizes in follow-up studies: ROC Area, Cohen’s d, and r. *Law Hum Behav*. 2005;29:615–620.

52. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biol Psychiat*. 2006;59:990–996.
53. Sheng Y, Ge Y, Yuan L, Li T, Yin FF, Wu QJ. Outlier identification in radiation therapy knowledge-based planning: a study of pelvic cases. *Med Phys*. 2017;44:5617–5626.
54. El Naqa I, Bradley JD, Lindsay PE, Hope AJ, Deasy JO. Predicting radiotherapy outcomes using statistical learning techniques. *Phys Med Biol*. 2009;54:S9–S30.
55. Britannica TEOE. Occam's razor. In *Encyclopædia Britannica*,. London, UK: Encyclopædia Britannica, inc.; 2015.
56. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2017.
57. Cherkassky VS, Mulier F. *Learning From Data: Concepts, Theory, and Methods*. 2nd ed. Hoboken, NJ: IEEE Press: Wiley-Interscience; 2007.
58. Lian J, Yuan L, Ge Y, et al. Modeling the dosimetry of organ-at-risk in head and neck IMRT planning: an intertechnique and interinstitutional study. *Med Phys*. 2013;40:121704.
59. Gammerman A, Vovk V. Prediction algorithms and confidence measures based on algorithmic randomness theory. *Theoret Comput Sci*. 2002;287:209–217.
60. Jiang B, Zhang X, Cai T. Estimating the confidence interval for prediction errors of support vector machine classifiers. *J Mach Learn Res*. 2008;9:521–540.
61. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn*. 1997;30:1145–1159.
62. Javanmard A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression. *J Mach Learn Res*. 2014;15:2869–2909.
63. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Stat Sci*. 1999;14:382–401.
64. Yuan L, Ge Y, Lee WR, Yin FF, Kirkpatrick JP, Wu QJ. Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans. *Med Phys*. 2012;39:6868–6878.
65. Cooper GF, Aliferis CF, Ambrosino R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med*. 1997;9:107–138.
66. Cooper GF, Abraham V, Aliferis CF, et al. Predicting dire outcomes of patients with community acquired pneumonia. *J Biomed Inform*. 2005;38:347–366.
67. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Paper presented at: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2015.
68. Lou Y, Caruana R, Gehrke J. Intelligible models for classification and regression. Paper presented at: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining 2012.
69. Valdes G, Luna JM, Eaton E, Simone CB. MediBoost: a patient stratification tool for interpretable decision making in the era of precision medicine. *Sci Rep*. 2016;6.
70. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv preprint arXiv:160605386. 2016.
71. Luo Y, El Naqa I, McShan DL, et al. Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis. *Radiother Oncol*. 2017;123:85–92.
72. Kang J, Schwartz R, Flickinger J, Beriwal S. Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. *Int J Radiat Oncol Biol Phys*. 2015;93:1127–1135.
73. Singh H, Spitzmueller C, Petersen NJ, Sawhney MK, Sittig DF. Information overload and missed test results in electronic health record-based settings. *JAMA Intern Med*. 2013;173:702–704.
74. Duncan J. Information overload: when less is more in medical imaging. *De Gruyter*. 2017;4:179–183.
75. Lambin P, van Stiphout RG, Starmans MH, et al. Predicting outcomes in radiation oncology–multifactorial decision support systems. *Nat Rev Clin Oncol*. 2013;10:27–40.
76. Smith WP, Kim M, Holdsworth C, Liao J, Phillips MH. Personalized treatment planning with a model of radiation therapy outcomes for use in multiobjective optimization of IMRT plans for prostate cancer. *Radiat Oncol*. 2016;11:38.
77. Mayo CS, Moran JM, Bosch W, et al. AAPM TG-263: standardizing nomenclatures in radiation oncology. *Int J Radiat Oncol Biol Phys*. 2017;99:E552.
78. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statist Sci*. 2001;16:199–231.
79. Dekker A, Vinod S, Holloway L, et al. Rapid learning in practice: a lung cancer survival decision support system in routine patient care data. *Radiother Oncol*. 2014;113:47–53.
80. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak*. 2012;12:8–8.
81. Wu Y. Elastic net for cox's proportional hazards model with a solution path algorithm. *Stat Sin*. 2012;22:27–294.
82. Japkowicz N, Shah M. Performance evaluation in machine learning. In: El Naqa I, Li R, Murphy MJ, eds. *Machine Learning in Radiation Oncology: Theory and Applications*. Switzerland: Springer-Verlag; 2015:41–56.
83. Page P. Beyond statistical significance: clinical interpretation of rehabilitation research literature. *Int J Sports Phys Ther*. 2014;9:726–736.
84. Harris RF. *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*. New York: Basic Books; 2017.
85. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. New York, NY: Springer; 2009.
86. Vallieres M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol*. 2015;60:5471–5496.