View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by IntechOpen

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter

Prediction of Cancer Patient Outcomes Based on Artificial Intelligence

Suk Lee, Eunbin Ju, Suk Woo Choi, Hyungju Lee, Jang Bo Shim, Kyung Hwan Chang, Kwang Hyeon Kim and Chul Yong Kim

Abstract

Knowledge-based outcome predictions are common before radiotherapy. Because there are various treatment techniques, numerous factors must be considered in predicting cancer patient outcomes. As expectations surrounding personalized radiotherapy using complex data have increased, studies on outcome predictions using artificial intelligence have also increased. Representative artificial intelligence techniques used to predict the outcomes of cancer patients in the field of radiation oncology include collecting and processing big data, text mining of clinical literature, and machine learning for implementing prediction models. Here, methods of data preparation and model construction to predict rates of survival and toxicity using artificial intelligence are described.

Keywords: big data, artificial intelligence, prediction, cancer patient outcomes, radiation oncology

1. Introduction

1.1 Definitions of big data

There are numerous definitions of big data covering attributes from technological needs to key thresholds to social impacts [1]. One popular definition of big data, proposed by Gartner, encompasses the "3Vs: volume, velocity, and variety" [2]. This definition refers to the increasing size of standard datasets, the increasing rate at which they are produced, and the increasing range of formats and representations employed. But there are few numerical quantifications in place to analyze big data. A fourth V, veracity, was added by IBM in 2012 [3]. Veracity describes questions of trust and uncertainty regarding data and results stemming from data. De Mauro et al. proposed an alternative definition of big data, introducing a fifth V (value): "Big data is the information asset characterized by such a high volume, velocity, and variety as to require specific technology and analytical methods for its transformation into value" (**Figure 1**) [1].

IntechOpen

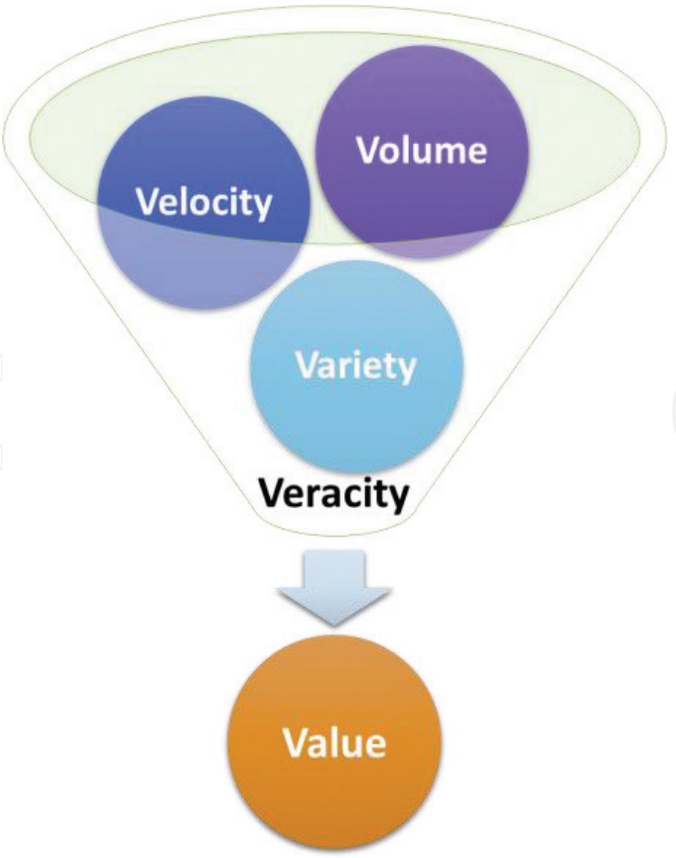

Figure 1.
The 5Vs of big data [4].

1.2 Differences between statistical analyses and machine learning

Statistical analyses are traditionally conducted using a mathematical formula based on a hypothesis, whereas machine learning is algorithm-based using data without rule-based programming. Statistics aims to infer the relationship between input and output and can explain the outcome of a probability distribution when the hypothesis is satisfied. A predictive model using statistical analyses has high explanatory power but low predictive power. Traditional statistical methods thus depend on a hypothesis. In most cases, machine learning predicts by directly modeling and learning from data, without hypothesis-based or rule-based programming. Machine learning focuses on important features; it ignores noise and outliers by extracting only important features from the data for the predictive model (**Figure 2**).

1.3 Big data in healthcare

Medical big data comprises complex results from a diversity of diseases, treatment methods, outcomes, data resources, analytical methods, and approaches for collecting, processing, and interpreting data [5]. There are various sources of medical big data, such as hospital information systems (HIS), electronic medical records (EMR), order communication records (OCR), picture archiving and communication systems (PACS), patient reports, biomarker data, genomic data, prospective cohort studies, and large clinical trials [6, 7]. There are several distinctive features of medical data that are different from data in other fields. Medical

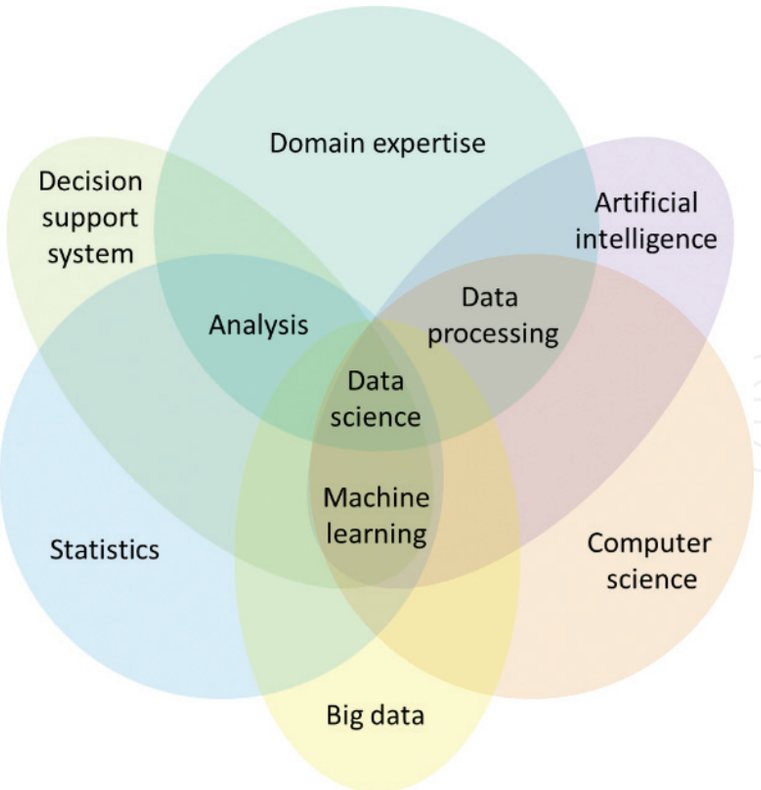

Figure 2.
The field of data science including statistics, big data, and artificial intelligence [8].

data are often difficult to access. Many investigators in the medical field are hesitant to practice open data science for various reasons, including the risk of data misuse by other parties. Medical data are often collected based on established protocols. These protocols commonly include preprocessing to simplify raw data. Both the acquisition and sharing of medical data require institutional approvals (e.g., approvals from an institutional review board), privacy protection for patients, shared agreement over the meaning of certain data elements, and an overall technology infrastructure enabling data sharing (such as a cloud-based system).

1.4 Big data in radiation oncology

In the radiation oncology field, diagnostic and therapeutic data are acquired throughout the course of treatment and during follow-up. Specific to radiation oncology, heterogeneous and voluminous amounts of data must be evaluated. These data exist in different formats across various information systems. Examples include hospital, laboratory, and oncology information systems (HIS, LIS, OIS), picture archiving and communication systems (PACS), and systems to record and verify (R&V) [9]. As expectations for personalized radiotherapy using complex data have increased, studies on outcome predictions using artificial intelligence have also increased. Specifically, studies of decision support systems based on big data have increased [10–12]. Several decision support systems have been developed in radiation oncology. Decision support systems for treatment planning have integrated imaging, dosimetry, biological, and other data in a quantitative manner to provide specific clinical predictions [13]. For example, a treatment planning decision support system that predicts radiation toxicity based on big data now exists [14]. Importantly, validation and standardization are crucial when developing medical decision support systems [15, 16].

2. Data preparation

2.1 Multi-institutional data collection

For prediction models using supervised learning, patient's data can be obtained by retrospectively analyzing the outcomes and prognoses of individual cancer patients. Since there can be data collection biases within a single institution, multi-institutional analyses are useful. Furthermore, data from one institution can be used to verify data from another institution. Oncospace (http://oncospace.radonc.jhmi.edu/) is a representative example of a multi-institutional big data platform in the field of radiation oncology. It comprises a database and web-based analysis tools for planning, data import, and outcome predictions [17]. Radiation oncology data sharing has been positively affected by the Oncospace consortium model.

2.2 Literature-based data collection

Data from previously published sources can be applied to prediction models. Representative databases for searching medical literature include PubMed (www.ncbi.nlm.nih.gov./entrez/query.fcgi), ScienceDirect (www.sciencedirect.com), Scirus (www.scirus.com/srsapp), ISI Web of Knowledge (http://www.isiwebofknowledge.com), and Google Scholar (http://scholar.google.com). It is important to obtain as many relevant studies as possible, as loss of studies can lead to bias.

The PRISMA statement recommends that a full electronic search of at least one major database be included [18]. Database searches can be augmented with manual searches of relevant papers, books, abstracts, and conference proceedings. Cross-checking references, capturing citations in review papers, and including communications from scientists working in a relevant field are important methods used to ensure that a comprehensive search is conducted [19].

3. Definitions of cancer patient outcomes

In 1993, the Outcomes Working Group (OWG) of the American Society of Clinical Oncology (ASCO) defined the outcomes of cancer treatment to be used for technical assessment and the development of cancer treatment guidelines [20]. According to the OWG, patient outcomes (e.g., survival rate or quality of life) should be prioritized over cancer outcomes (e.g., toxicity, response, or cost-effectiveness). Since a single outcome is not indicative of the overall patient outcome following cancer treatment, multiple outcomes should be considered [20]. In this chapter, we discuss three important outcomes to consider when choosing a treatment plan: toxicity, response, and survival rate.

3.1 Toxicity

Toxicity (either acute or chronic) is vitally important, with chronic toxicity being particularly critical in children [20]. The Radiation Therapy Oncology Group (RTOG) distinguishes acute and late toxicity from the side effects that occur during radiation therapy and provides guidelines for the clinical management of toxicity graded for each critical organ. Toxicity can be scored using the Common Terminology Criteria for Adverse Events (CTCAE). The CTCAE scoring system is a product of the US National Cancer Institute (NCI) [21]. Toxicity is graded as mild (grade 1), moderate (grade 2), severe (grade 3), or life-threatening (grade 4), with specific

parameters for the organ system involved. Death (grade 5) is used to denote a fatality occurring during treatment [22].

3.2 Response

A solid tumor response assessment usually consists of a bidimensional (World Health Organization criteria, WHO) or unidimensional (response evaluation criteria in solid tumors guidelines, RECIST) measurement of tumors before and after chemotherapy [23, 24].

A treatment response can be grouped into four categories which are as follows: a complete response (CR), with the disappearance of all target lesions; a partial response (PR), with a decrease of greater than 30% of the target lesions; disease progression (DP), with an increase of greater than 20% of the target lesions, the appearance of new lesions, and/or the unequivocal progression of nontarget lesions; and stable disease (SD), with changes in tumor size not otherwise qualifying as PR or PD [23, 25].

3.3 Survival rate

The 5-year survival rate represents the percentage of patients living at least 5 years after a cancer is found. For example, the international 5-year survival rate for patients with lung cancer varies from 5–16% [26].

4. Prediction models

The accurate prediction of a patient's outcome before radiotherapy is an interesting and challenging task (**Figure 3**) [15, 28–30]. Machine learning (ML) methods have become popular with medical researchers. ML techniques can discover and identify patterns and relationships between treatment methods and outcomes. Using complex datasets, ML algorithms are increasingly able to predict outcomes for a specific cancer type [16, 29, 31–34].

The artificial neural net (ANN) and support vector machine (SVM) classifiers are among the most widely used ML algorithms related to cancer patient outcomes. The ANN algorithm has been used for almost 30 years. The SVM tool constitutes a more recent approach to predict cancer outcomes and is popular for its accurate

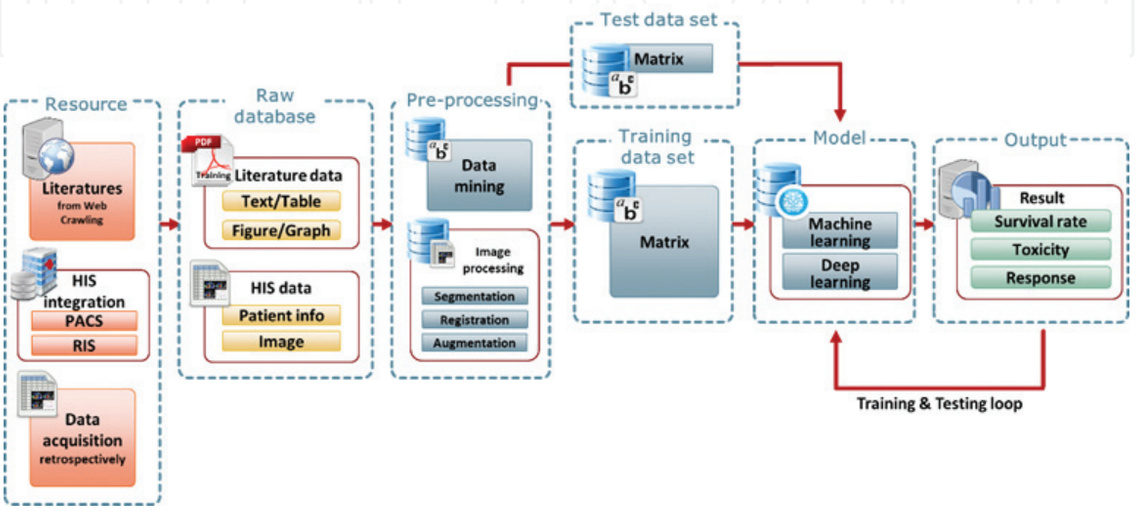

Figure 3.
Workflow of a prediction model, from raw data to the prediction result [27].

predictive performance. The most suitable algorithm choice for prediction depends on various parameters, including the type of data collected, the size of the data samples, the time frame for collection and analysis, and the type of results needed [29].

When using literature to collect data for prediction model implementation, text mining is often needed to transform literature to structured data. A major part of the text mining process involves the crucial stage of preprocessing the literature (i.e., dealing with unstructured data). Preprocessing techniques such as text categorization and term extraction are necessary. The text mining process itself requires the storage of intermediate representations, techniques to analyze intermediate representations, clustering, trend analysis, association rules, and visualization of results [35].

4.1 Toxicity prediction using clinical data

When treating cancer patients, the dual administration of chemotherapy and radiotherapy can cause severe toxicity [36]. Several studies using ANN to predict the toxicity of radiation therapy at various tumor sites have been conducted. Among tumor sites, there is a high probability of radiation toxicity in the head and neck. According to one study in 2002, they tested on clinical data and proved to be able to predict which patients will tolerate a combined chemoradiotherapy and to supply a potential predictive indicator for radiation toxicity. Clinical data were derived from 63 consecutive cases. All patients admitted into the study received induction chemotherapy for three cycles followed by concomitant chemoradiotherapy to treat head and neck cancer. They used an interval arithmetic perceptron (IAP) algorithm that consists of a neural network with a single layer of weights. The prediction performance using 11 input variables is 76.19% of correctly classified cases, whereas the whole network using 38 input variables allows only 53.97% of successes, confirming that reducing the input variables to the salient ones do improve statistical performances [37].

4.2 Response prediction using medical images

To better predict tumor responses to chemotherapy, a modeling study using CT and MR images was performed. In breast cancer patients, MR images generated useful clinical markers. MR images of 68 cancer patients were obtained before neoadjuvant chemotherapy, after which 25 patients were CR and 43 were NR. There is no statistically significant difference of each of these image features between the CR and NR case groups ($p > 0.05$). After applying ROC analysis on each of the 39 features, 10 features yielded AUC > 0.6 in classifying between the CR and NR case groups. The artificial neural network yielded an AUC = 0.96 ± 0.03, which is significantly higher than AUC = 0.85 ± 0.05 yielded using a simple feature fusion method ($p < 0.01$). The overall accuracy of response prediction was 94% with a sensitivity of 88% at a specificity of 98% [38].

4.3 Survival rate prediction using immunohistochemical data

In 2003, an ANN analysis proved to be more accurate than a statistical analysis in predicting the survival rate of patients with non-small cell lung cancer (NSCLC). In the study, a predictive model was implemented using data from 125 lung cancer patients. The study used 17 input variables (including five immunohistochemical parameters: p27 percentage, p27 intensity, p53, cyclin D1, and retinoblastoma) and 12 clinicopathological variables (including age, sex, smoking index, tumor size, p factor, pT, pN, stage, and histology). The prediction accuracy of the NSCLC 5-year

survival rate using ANN was 87%, whereas the prediction accuracy using a logistic regression analysis was 78% [39].

4.4 Text mining-based toxicity prediction model

Prediction of radiation toxicity at the treatment planning stage of radiotherapy can improve tumor control and quality of life. However, due to the lack of patient data analyzed retrospectively in actual clinical practice, there is a limit to establish accurate prediction models. Thus, we used semantic data mining method to structure the meta-analysis literature related to radiation pneumonitis and constructed a dataset for machine learning. The 160 peer-reviewed papers related to radiation pneumonitis were structured through semantic data mining (Konan Analytics 4, Konan Technology Inc., Republic of Korea). In a structured learning dataset, the target variable was set to grade 1–5 pneumonitis graded according to the National Cancer Institute Common Toxicity Criteria version 3.0. The predictor variable was set to 10 factors (interstitial lung disease, chronic obstructive pulmonary disease, pulmonary function, age, concurrent chemotherapy, tumor location, mean lung dose, V15, V20, V30). Based on the target variable characteristics, support vector regression algorithm was implemented using the scikit-learn open source toolkit. The accuracy of the regression model was expressed in the form of root-mean-square error (RMSE) comparing the difference between the predicted value and the actual value. In order to evaluate the results of radiation pneumonitis prediction using unstructured data, we compared structured data that retrospectively analyzed 110 cases of lung cancer patients. Therefore, the semantic database of 39,404 cases related to radiation pneumonitis was constructed through semantic data mining. The results of the radiation pneumonitis prediction showed RMSE of 1.307 using a structured semantic database and RMSE of 1.056 using the retrospectively analyzed lung cancer patient data. It was confirmed that there is no difference between prediction model using unstructured data and structured data (RMSE cost difference, 0.251).

5. Limitations

The main obstacle to widely applying AI in the radiation oncology field is the lack of valid data. Only 2–3% of available data adequately capture a patient's current state of health and medical history. Suitable data are, nonetheless, included in certain ongoing clinical trials.

Since no dataset is likely to include all the features needed for an AI analysis, handling of missing data is needed to build a sufficient dataset for machine learning. A researcher can compensate for missing data by interpolating from the surrounding values, filling gaps with average values, or applying new artificial intelligence methods. The "curse of dimension" seen in machine learning with numerous features may make it necessary to select input factors using techniques like principal component analysis (PCA) or feature selection.

6. Conclusions

Due to the increasing size of datasets, the increasing rate at which they are produced and the increasing range of formats employed, predictive analysis studies using big data and artificial intelligence have also increased. In the radiation oncology field, there are ongoing trials to implement AI for predictive analyses.

Outcomes such as survival rate, tumor response, and radiation toxicity are important to cancer patients and physicians alike. In some cases, ANN is superior to conventional statistical analyses in predicting a cancer patient's prognosis. Recently, an ensemble model has emerged, combining the advantages of various ML algorithms to make predictions. Although it is sometimes difficult to interpret the processes and results obtained from artificial intelligence techniques, the current research into explainable artificial intelligence (XAI) can help to provide insight [40]. Given the lack of retrospectively analyzed data, there are limits to collecting learning data of high quality. This limitation might be overcome by data mining the clinical literature. In summary, the increased use of big data and complex variables in medicine suggests that AI will become increasingly crucial in predicting cancer patient outcomes.

Acknowledgements

This project was supported by the Korean Small and Medium Business Administration (Grant No. C0558199 and No. C0558032), the Ministry of Science, ICT and Future Planning in Korea (2017R1A2B2004012), and Korea University (K1722451).

Conflict of interest

The researcher claims no conflicts of interest.

Author details

Suk Lee[1*], Eunbin Ju[1], Suk Woo Choi[2], Hyungju Lee[3], Jang Bo Shim[1], Kyung Hwan Chang[4], Kwang Hyeon Kim[5] and Chul Yong Kim[1]

1 Department of Radiation Oncology, College of Medicine, Korea University, Seoul, Republic of Korea

2 Medical Standard Co., Ltd., Gyeonggi, Republic of Korea

3 Konan Technology Inc., Seoul, Republic of Korea

4 Department of Radiation Oncology, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul, Republic of Korea

5 Proton Therapy Center, National Cancer Center, Gyeonggi, Republic of Korea

*Address all correspondence to: sukmp@korea.ac.kr

IntechOpen

© 2018 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. (cc) BY

References

[1] Andrea DM, Marco G, Michele G. A formal definition of big data based on its essential features. Library Review. 2016; **65**(3):122-135

[2] Douglas L. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Gartner. 2001

[3] Beyer MA, Laney D. The Importance of 'Big Data': A Definition. Stamford, CT: Gartner; 2012. pp. 2014-2018

[4] Jeff. Big Data, Digital Marketing, Social Listening [Internet]. 2018. Available from: http://chinetekstrategy.com/blog/2017/12/28/social-listening-big-data [Accessed: 2018-11-05]

[5] Dinov ID. Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. Gigascience. 2016;**5**(1):12

[6] Lee CH, Yoon HJ. Medical big data: Promise and challenges. KidneyResearch and Clinical Practice. 2017;**36**(1):3-11

[7] Slobogean GP et al. Bigger data, bigger problems. Journal of Orthopaedic Trauma. 2015;**29**:S43-S46

[8] Palmer S. Data Science for the C-Suite. New York: Digital Living Press; 2015

[9] Kessel KA, Combs SE. Data management, documentation and analysis systems in radiation oncology: A multi-institutional survey. Radiation Oncology. 2015;**10**(1):230

[10] Ree A, Redalen K. Personalized radiotherapy: Concepts, biomarkers and trial design. The British Journal of Radiology. 2015;**88**(1051):20150009

[11] Lambin P et al. Decision support systems for personalized and participative radiation oncology. Advanced Drug Delivery Reviews. 2017; **109**:131-153

[12] Huilgol N. Big data in radiation oncology. Journal of Cancer Research and Therapeutics. 2016;**12**(3):1107-1108

[13] Lee S, Cao YJ, Kim CY. Physical and radiobiological evaluation of radiotherapy treatment plan. In: Evolution of Ionizing Radiation Research. Rijeka, Croatia: InTech; 2015

[14] Lee S et al. Predictive Solution for Radiation Toxicity Based on Big Data; 2017

[15] Lambin P et al. Predicting outcomes in radiation oncology–Multifactorial decision support systems. Nature Reviews. Clinical Oncology. 2013;**10**(1): 27-40

[16] Jochems A et al. Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries. International Journal of Radiation Oncology, Biology, Physics. 2017;**99**(2): 344-352

[17] Bowers MR et al. Oncospace consortium: A shared radiation oncology database system designed for personalized medicine and research. International Journal of Radiation Oncology Biology Physics. 2015;**93**(3): E385

[18] Moher D et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews. 2015; **4**(1):1

[19] Haidich A-B. Meta-analysis in medical research. Hippokratia. 2010;**14** (Supp. 1):29

[20] Fetting J et al. Outcomes of cancer treatment for technology assessment and cancer treatment guidelines. Journal of Clinical Oncology. 1996;**14**(2): 671-679

[21] US Department of Health and Human Services. Common terminology criteria for adverse events (CTCAE) version 4.0. National Institutes of Health, National Cancer Institute. 2009;**4**(3)

[22] Savarese DM. Common Terminology Criteria for Adverse Events. UpToDate. Waltham, MA: UpToDate; 2013

[23] Shanbhogue AKP, Karnad AB, Prasad SR. Tumor response evaluation in oncology: Current update. Journal of Computer Assisted Tomography. 2010; **34**(4):479-484

[24] Park JO et al. Measuring response in solid tumors: Comparison of RECIST and WHO response criteria. Japanese Journal of Clinical Oncology. 2003; **33**(10):533-537

[25] Duffaud F, Therasse P. New guidelines to evaluate the response to treatment in solid tumors. Bulletin du Cancer. 2000;**87**(12):881-886

[26] Butler CA et al. Variation in lung cancer survival rates between countries: Do differences in data reporting contribute? Respiratory Medicine. 2006; **100**(9):1642-1646

[27] Eunbin Ju SL, Kim KH, Choi SW, Chang KH, Cao YJ, Shim JB, et al. Quantitative analysis of weight of prognostic factors related to radiation pneumonitis using statistical analysis and artificial neural network. In: American Society for Radiation Oncology(ASTRO) Annual Meeting. 2018

[28] El Naqa I et al. Predicting radiotherapy outcomes using statistical learning techniques. Physics in Medicine & Biology. 2009;**54**(18):S9

[29] Kourou K et al. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal. 2015; **13**:8-17

[30] Feng M et al. Machine learning in radiation oncology: Opportunities, requirements, and needs. Frontiers in Oncology. 2018;**8**:110

[31] Oermann EK et al. Using a machine learning approach to predict outcomes after radiosurgery for cerebral arteriovenous malformations. Scientific Reports. 2016;**6**:21161

[32] 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**(3):459-467

[33] Lustberg T et al. Implementation of a rapid learning platform: Predicting 2-year survival in laryngeal carcinoma patients in a clinical setting. Oncotarget. 2016;**7**(24):37288

[34] Lambin P et al. Rapid learning health care in oncology–An approach towards decision support systems enabling customised radiotherapy. Radiotherapy and Oncology. 2013; **109**(1):159-164

[35] Feldman R et al. Mining the biomedical literature using semantic analysis and natural language processing techniques. Biosilico. 2003;**1**(2):69-80

[36] Su M et al. An artificial neural network for predicting the incidence of radiation pneumonitis. Medical Physics. 2005;**32**(2):318-325

[37] Drago GP et al. Forecasting the performance status of head and neck cancer patient treatment by an interval arithmetic pruned perceptron. IEEE Transactions on Biomedical Engineering. 2002;**49**(8):782-787

[38] Aghaei F et al. Computer-aided breast MR image feature analysis for prediction of tumor response to chemotherapy. Medical Physics. 2015; **42**(11):6520-6528

[39] Hanai T et al. Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression. Cancer Science. 2003;**94**(5):473-477

[40] Gunning D. Explainable Artificial Intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web; 2017

IntechOpen