Using radiomics and deep learning-based imaging biomarkers to predict radiotherapy outcomes and toxicity

Using radiomics and deep learning-based imaging biomarkers to predict radiotherapy outcomes and toxicity

Dissertation

to obtain the degree of Doctor at the Maastricht University,

on the authority of the Rector Magnificus,

Prof. dr. Pamela Habibović,

in accordance with the decision of the Board of Deans,

to be defended in public

on Tuesday 4th July 2023, at 13:00 hours

by

Zhen Zhang

Supervisor	Prof.dr.ir. A.L.A.J. Dekker
Co-supervisors	Dr. L.Y.L. Wee
	Dr. A. Traverso
Assessment Committee	Prof. Liesbeth Boersma (chair)
	Dr. Ralph Brecheisen
	Prof. Coen Hurkmans, TU Eindhoven
	Dr. Tian-Tian Zhai, Cancer Hospital of Shantou University Medical College, Shantou, China

Index

Chapter 1: Introduction and Outline of Thesis

1.1 The role of radiation therapy in the management of cancer patients

Radiation therapy is an important treatment modality in the management of cancer. Approximately 77% of lung cancer patients have an indication to receive radiation therapy [1]. As technology evolves and new drugs are entering cancer care, integrated or comprehensive cancer treatment is becoming mainstream [2]. Specifically, the main treatments for cancer include surgery, chemotherapy, immunotherapy, radiotherapy, the last of which plays an important role in the integrated treatment. [3]. The use of radiotherapy in combination with surgery can improve survival, i.e., extend the survival time of patients treated with neoadjuvant or consolidation radiotherapy [4, 5]. Concurrent chemoradiotherapy has become a standard treatment option for many cancers, such as locally advanced non-small cell lung cancer, for which it is the primary curative treatment. [6]. With the advent of immunotherapy, the pairing with radiotherapy, including optimal dose fractions, is under active investigation.

Several studies have demonstrated that radiotherapy and immunotherapy can have potentially synergistic effects [7, 8]. The PACIFIC study, a milestone in immunotherapy, established the status of consolidation immunotherapy following concurrent chemoradiotherapy [9]. Radiotherapy can attenuate immune resistance, induce the release of TGF-β, and the upregulation of PD-L1 expression as well as the reprogramming of the immune microenvironment [7, 10]. At the same time, immunogenic cell death due to radiotherapy can promote the release of cytokines such as interferons, tumor necrosis factor-α, Interleukin-1 and Interleukin-6, etc [11]. Based on some of these findings, several clinical trials incorporating radiotherapy and immunotherapy have been designed (e.g., PACIFIC-4 NCT03833154 and ISABR NCT03148327).

An increase of radiotherapy efficacy will not only improve the outcomes of patients, but also gives patients more options, such as the opportunity to receive surgery for patients who are treated with neoadjuvant radiotherapy, and improve the effectiveness of the multidisciplinary synergistic treatment model. In addition, radiotherapy gives hope to patients who are unable to receive other treatment modalities. In the Netherlands, for example, with the use of stereotactic ablative body radiotherapy (SABR), the proportion of untreated elderly lung cancer patients is gradually decreasing and, accordingly, survival is increasing [12].

Therefore, with the development of innovative treatment techniques, options and combinations, radiotherapy continues its important role in the treatment of cancer.

1.2 Precision medicine in radiation therapy

While the value of radiotherapy is clear, it also presents challenges. If radiotherapy can be used wisely to maximize its value, it can improve patient prognosis and reduce patient pain and treatment costs. Conversely, if treatment decisions are not made appropriately, for example, if the estimated efficacy of the treatment is far from the actual outcome, it will not only fail to improve the prognosis, but also impact other interventions. Specifically, it may delay the intervention of other therapies or prevent the immediate use of systemic therapy due to side effects of radiation therapy [13]. Therefore, it is crucial to improve the targeting of radiotherapy, i.e., to precisely select patients suitable for radiotherapy and to implement tailored radiotherapy regimens.

Precision medicine is not a recent novel concept [14], but it is still relevant in modern times and has practical implications that require dedicated research to complete and refine its framework. In the field of radiotherapy, precision medicine is widely recognized and accepted, especially with the emergence of novel technologies, and known as precision radiotherapy [15]. The aim of precision radiotherapy is to give the optimal treatment regimen to each patient, based on individualized conditions. As shown in Figure 1, precision radiotherapy is achieved by combining it with other treatment modalities to develop a personalized care plan with appropriate radiotherapy techniques, ultimately improving the patient's prognosis and reducing the side effects of the treatment(s). One of the elements that make this concept possible is a variety of biomarkers that reflect information about a patient's tumor and/or normal organs and thus predict clinical endpoints such as survival and toxicities. [16].

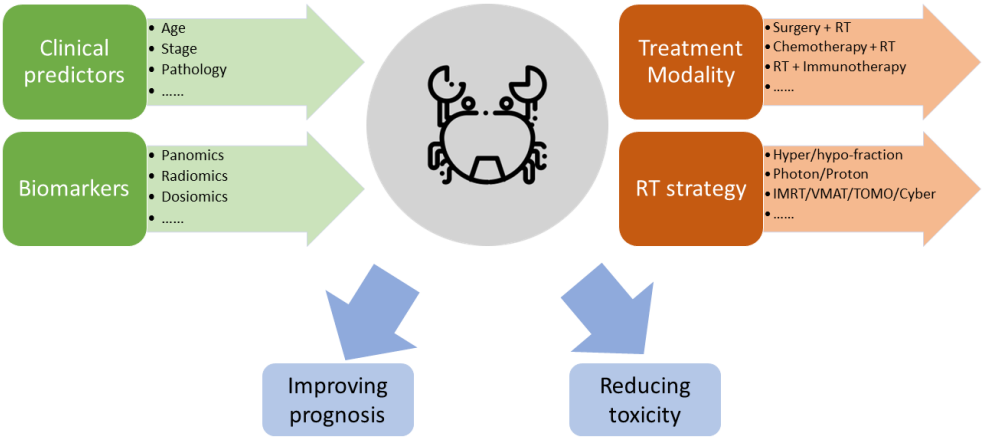


Figure 1. Precision radiotherapy requires predictors from different origins to help make clinical decisions and treatment strategies, and the ultimate goal of precision radiotherapy is to improve prognosis and reduce toxicity.

Panomics is an important class of biomarkers on which numerous researchers have focused their work [17, 18]. With the application of next-generation sequencing technology, many exciting genomics studies have emerged [19]. In addition, cytokines, immune microenvironment-related markers, proteome and metabolome have also been shown to have prognostic roles and guiding individualized treatment [20]. The estimated radiation dose to immune cells has also been demonstrated to correlate with the overall survival of patients treated with radiotherapy [21]. However, a limitation of these panomics biomarkers is that most of them are difficult to obtain, i.e., by invasive means and at high cost.

One of the main research questions of this thesis is: Is it possible to develop low-cost biomarkers that can be accessed quickly and non-invasively to assist in clinical decision making?

1.3 Radiomics and Image-based biomarkers

Radiomics, as a non-invasive method to extract quantitative features from medical images, was proposed in 2012 [22]. On the one hand, radiomics can reflect some of the same information as the semantic features obtained by physicians. On the other hand, radiomics contains some information that is not available to the naked eye. Radiomics are divided into two categories, one is handcrafted features, i.e., already predefined by mathematical formulas, and the other is deep learning features, i.e., features automatically extracted by models constructed from convolutional neural networks and other means, which have no fixed formula and definition [23]. The steps to build a radiomics model from a technical point of view are roughly divided into region of interest definition, preprocessing, feature extraction, and model building. Thus, radiomics, in contrast to panomics, contains not only image-based biomarkers, but also models, or so-called signatures, built on the biomarkers. [24]. As a result, radiomics is more accessible, reusable, cost effective, and in most cases does not require additional patient examinations. More importantly, it can be combined with traditional clinical predictors, and panomics, etc., without mutual exclusivity [25]. It is currently considered as a promising technology to assist/guide clinical decision making.

Improvements in radiotherapy techniques are always accompanied by advances in imaging [26, 27], and a large amount of imaging data is acquired throughout the management of radiotherapy patients. Medical imaging, such as diagnostic CT, MRI and PET, is included in the pre-treatment diagnosis of most patients. During radiotherapy, planning CT/MRI and cone beam CTs are obtained. After radiation therapy, patients are followed up with regular imaging examinations. Therefore, radiomics is a worthwhile research area for the field of radiotherapy.

There has been a large amount of radiomics studies in the field of radiotherapy [28]. In the case of lung cancer, for example, studies on radiomics cover almost every pathological type and every stage. As mentioned above in the definition of precision radiotherapy, most of the studies are aimed at improving prognosis and reducing the incidence of side effects [29, 30]. These inspiring studies give researchers confidence and demonstrate its potential for application. However, throughout these studies there are also some issues and challenges that need to be addressed. The first is the uneven standard of radiomics research, which is mainly due to the lack of a corresponding quality evaluation system [31, 32]. Therefore, there is a **need to develop methodological evaluation criteria for radiomics research** (Chapter 2) based on the existing quality assessment standards. And based on these methodological evaluation criteria, the published studies should be evaluated to have an objective assessment of the current stage of radiomics research. Most studies currently include a small amount of data (around 50-200 patients) [28, 33], and in our experience, we consider this lack of data to be a major obstacle to building clinical hypothesis models. As a result, real-world research does not really live up to the “big data” vision. The reasons for this come from a variety of sources, including inadequate data storage and management; the amount of data available for a particular clinical problem is drastically reduced after rigorous screening based on inclusion and exclusion criteria. A research question in this thesis is thus: Is it possible to **perform sample expansion / augmentation** (Chapter 3) by technical means to meet the data requirements of radiomics studies?

1.4 The need for radiomics and image-based biomarkers in predicting the prognosis and toxicity of radiotherapy

From the clinical perspective, most of the current radiomics studies are closely related to clinical needs and have practical application prospects [34], but there are still some details that are worth exploring. For early-stage NSCLC patients who do not wish to undergo surgery and for those who are medically inoperable, SBRT has become the standard of care. It is well tolerated and provides high rates of local control [35]. Nevertheless, distant failure in early-stage patients is common. Distant failure is highly correlated with poor prognosis, and for patients with distant failure, systemic therapy, such as chemotherapy or tyrosine kinase inhibitor (TKI)-targeted therapy, can help improve prognosis. Therefore, there is a need to **develop a biomarker to effectively predict distant failure in early-stage patients treated with SBRT** (Chapter 4) and to guide physicians on appropriate treatment for high-risk groups.

For locally-advanced lung cancer, curative radiotherapy is one of the main treatment modalities. However, there is a large variation in the survival of patients who receive radiotherapy. Therefore, the identification of **biomarkers predicting the prognosis of these patients is relevant and crucial for the radiotherapy field** (Chapter 5) with which more attention can be given to high-risk patients. On the other hand, because radiotherapy is a double-edged sword, radiation pneumonitis (RP) [36] is a major toxicity of lung cancer treated with radiotherapy. RP is a non-infectious pneumonia, induced by radiation, that reduces the quality of life of patients and can be fatal in severe cases. It is not uncommon for RP to occur, especially in patients with locally advanced lung cancer. Robust and reproducible **prediction models that could estimate the risk of developing RP after lung RT** (Chapter 6 and 7) would be of tangible clinical value. And for those patients at high risk of developing RP, prophylactic medication and active vigilance could be indicated.

1.5 Aim and outline of this thesis

The overall aim of this thesis is to use radiomics to assist in clinical decision-making regarding prognosis and toxicity.

Following this introduction, this thesis assesses the quality of published radiomics articles and presents a methodological assessment checklist (Chapter 2), introduces a data augmentation method based on a deep learning approach (Chapter 3). The predictive power of radiomics for lung cancer prognosis (Chapters 4-5) and radiotherapy-related toxicity (Chapters 6-7) are also explored and evaluated with prospective or/and multi-institutional datasets. Finally, I discuss the challenges and prospects of radiomics (Chapter 8). The outline of this thesis is summarized in **Table 1**.

Chapter 2 proposes an appraisal matrix with 13 items to assess the methodological quality of radiomics studies. Published studies are also evaluated, using esophageal cancer as an example.

Chapter 3 uses a dual discriminator super-resolution generative adversarial network to generate synthetic ground glass nodules that have the potential to become lung cancer. Radiomic features were extracted from both the generated and real nodules, and these features were compared.

Chapter 4 uses PET- and CT-based radiomic features to predict the risk of distant metastases in patients with early-stage lung cancer who underwent SABR.

Chapter 5 utilizes radiomics features extracted from both normal lung tissue, as well as tumor tissue for prognosis prediction of patients with locally advanced lung cancer.

Chapter 6 combines radiomics, dosiomics and clinical parameters to predict radiation pneumonitis and compares it with the benchmark models.

Chapter 7 uses CT images and radiation dose images to predict radiation pneumonitis and uses deep learning techniques to make the model applicable to groups with different dose patterns without the need for complex retraining.

Chapter 8 discusses the main findings of this thesis and reflections on the results. The limitations of the current radiomics research are described, and the future direction of radiomics in radiotherapy is prospected.

Table 1. The chapters in this thesis.

Section	Chapter	Title	Main finding
Introduction	Chapter 1	Introduction of the thesis	
Basic research in radiomics	Chapter 2	Methodological quality of machine learning-based quantitative imaging analysis studies in esophageal cancer: a systematic review of clinical outcome prediction after concurrent chemoradiotherapy	Methodological evaluation checklist is presented.
	Chapter 3	Generation of synthetic ground glass nodules using generative adversarial networks (GANs)	GAN method can generate synthetic images.
Biomarkers of prognosis	Chapter 4	A PET/CT radiomics model for predicting distant metastasis in early-stage non-small cell lung cancer patients treated with stereotactic body radiotherapy: A multicentric study	CT-based combined with PET-based radiomics can effectively predict DM.
	Chapter 5	Combining tumor radiomics features and whole-lung radiomics features to predict prognosis in locally advanced non-small cell lung cancer treated with curative radiotherapy	The role of lung tissue cannot be ignored when predicting OS in lung cancer by radiomics.

Biomarkers of toxicity	Chapter 6	Radiomics and dosiomics signature from whole lung predicts radiation pneumonitis: a model development study with prospective external validation and decision-curve analysis	Dosiomics performs better than the DVH metric in predicting RP.
	Chapter 7	Computed tomography and radiation dose images-based deep-learning model for predicting radiation pneumonitis in lung cancer patients after radiation therapy: A pilot study with external validation	Deep learning approach can help to apply the model to different cohorts.
Discussion	Chapter 8	Discussion and Future Perspectives	

Abbreviation: GAN, Generative Adversarial Network; DM, distant metastasis; OS, overall survival; DVH, dose-volume histogram; RP, radiation pneumonitis.

References

1. Delaney GP, Barton MB. Evidence-based estimates of the demand for radiotherapy. Clin Oncol (R Coll Radiol). 2015;27:70-6. doi:10.1016/j.clon.2014.10.005.

2. Pillay B, Wootten AC, Crowe H, Corcoran N, Tran B, Bowden P, et al. The impact of multidisciplinary team meetings on patient assessment, management and outcomes in oncology settings: A systematic review of the literature. Cancer Treat Rev. 2016;42:56-72. doi:10.1016/j.ctrv.2015.11.007.

3. Schaue D, McBride WH. Opportunities and challenges of radiotherapy for treating cancer. Nat Rev Clin Oncol. 2015;12:527-40. doi:10.1038/nrclinonc.2015.120.

4. Billiet C, Peeters S, Decaluwé H, Vansteenkiste J, Mebis J, Ruysscher D. Postoperative radiotherapy for lung cancer: Is it worth the controversy? Cancer Treat Rev. 2016;51:10-8. doi:10.1016/j.ctrv.2016.10.001.

5. Citrin DE. Recent Developments in Radiotherapy. N Engl J Med. 2017;377:1065-75. doi:10.1056/NEJMra1608986.

6. Bradley JD, Hu C, Komaki RR, Masters GA, Blumenschein GR, Schild SE, et al. Long-Term Results of NRG Oncology RTOG 0617: Standard- Versus High-Dose Chemoradiotherapy With or Without Cetuximab for Unresectable Stage III Non-Small-Cell Lung Cancer. J Clin Oncol. 2020;38:706-14. doi:10.1200/jco.19.01162.

7. Mondini M, Levy A, Meziani L, Milliat F, Deutsch E. Radiotherapy-immunotherapy combinations - perspectives and challenges. Mol Oncol. 2020;14:1529-37. doi:10.1002/1878-0261.12658.

8. Arina A, Gutiontov SI, Weichselbaum RR. Radiotherapy and Immunotherapy for Cancer: From “Systemic” to “Multisite”. Clin Cancer Res. 2020;26:2777-82. doi:10.1158/1078-0432.Ccr-19-2034.

9. Gray JE, Villegas A, Daniel D, Vicente D, Murakami S, Hui R, et al. Three-Year Overall Survival with Durvalumab after Chemoradiotherapy in Stage III NSCLC-Update from PACIFIC. J Thorac Oncol. 2020;15:288-93. doi:10.1016/j.jtho.2019.10.002.

10. Yu WD, Sun G, Li J, Xu J, Wang X. Mechanisms and therapeutic potentials of cancer immunotherapy in combination with radiotherapy and/or chemotherapy. Cancer Lett. 2019;452:66-70. doi:10.1016/j.canlet.2019.02.048.

11. Chen Y, Gao M, Huang Z, Yu J, Meng X. SBRT combined with PD-1/PD-L1 inhibitors in NSCLC treatment: a focus on the mechanisms, advances, and future challenges. J Hematol Oncol. 2020;13:105. doi:10.1186/s13045-020-00940-z.

12. Palma D, Visser O, Lagerwaard FJ, Belderbos J, Slotman BJ, Senan S. Impact of introducing stereotactic lung radiotherapy for elderly patients with stage I non-small-cell lung cancer: a population-based time-trend analysis. J Clin Oncol. 2010;28:5153-9. doi:10.1200/jco.2010.30.0731.

13. Barazzuol L, Coppes RP, van Luijk P. Prevention and treatment of radiotherapy-induced side effects. Mol Oncol. 2020;14:1538-54. doi:10.1002/1878-0261.12750.

14. Hodson R. Precision medicine. Nature. 2016;537:S49. doi:10.1038/537S49a.

15. Caudell JJ, Torres-Roca JF, Gillies RJ, Enderling H, Kim S, Rishi A, et al. The future of personalised radiotherapy for head and neck cancer. Lancet Oncol. 2017;18:e266-e73. doi:10.1016/s1470-2045(17)30252-8.

16. Califf RM. Biomarker definitions and their applications. Exp Biol Med (Maywood). 2018;243:213-21. doi:10.1177/1535370217750088.

17. Cheng F, Hong H, Yang S, Wei Y. Individualized network-based drug repositioning infrastructure for precision oncology in the panomics era. Brief Bioinform. 2017;18:682-97. doi:10.1093/bib/bbw051.

18. Hsieh JJ, Le V, Cao D, Cheng EH, Creighton CJ. Genomic classifications of renal cell carcinoma: a critical step towards the future application of personalized kidney cancer care with pan-omics precision. J Pathol. 2018;244:525-37. doi:10.1002/path.5022.

19. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155:27-38. doi:10.1016/j.cell.2013.09.006.

20. Meng L, Xu J, Ye Y, Wang Y, Luo S, Gong X. The Combination of Radiotherapy With Immunotherapy and Potential Predictive Biomarkers for Treatment of Non-Small Cell Lung Cancer Patients. Front Immunol. 2021;12:723609. doi:10.3389/fimmu.2021.723609.

21. Yin X, Luo J, Xu C, Meng C, Zhang J, Yu H, et al. Is a higher estimated dose of radiation to immune cells predictive of survival in patients with locally advanced non-small cell lung cancer treated with thoracic radiotherapy? Radiother Oncol. 2021;159:218-23. doi:10.1016/j.radonc.2021.03.026.

22. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441-6. doi:10.1016/j.ejca.2011.11.036.

23. Scapicchio C, Gabelloni M, Barucci A, Cioni D, Saba L, Neri E. A deep look into radiomics. Radiol Med. 2021;126:1296-311. doi:10.1007/s11547-021-01389-x.

24. Compter I, Verduin M, Shi Z, Woodruff HC, Smeenk RJ, Rozema T, et al. Deciphering the glioblastoma phenotype by computed tomography radiomics. Radiother Oncol. 2021;160:132-9. doi:10.1016/j.radonc.2021.05.002.

25. Jha AK, Mithun S, Purandare NC, Kumar R, Rangarajan V, Wee L, et al. Radiomics: a quantitative imaging biomarker in precision oncology. Nucl Med Commun. 2022;43:483-93. doi:10.1097/mnm.0000000000001543.

26. Tan Mbbs Mrcp Frcr Md LT, Tanderup Ph DK, Kirisits Ph DC, de Leeuw Ph DA, Nout Md Ph DR, Duke Mbbs Frcr S, et al. Image-guided Adaptive Radiotherapy in Cervical

Cancer. Semin Radiat Oncol. 2019;29:284-98. doi:10.1016/j.semradonc.2019.02.010.

27. Lou Y, Niu T, Jia X, Vela PA, Zhu L, Tannenbaum AR. Joint CT/CBCT deformable registration and CBCT enhancement for cancer radiotherapy. Med Image Anal. 2013;17:387-400. doi:10.1016/j.media.2013.01.005.

28. A T, L W, A D, R G. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. International journal of radiation oncology, biology, physics. 2018;102. doi:10.1016/j.ijrobp.2018.05.053.

29. Peng Z, Wang Y, Wang Y, Jiang S, Fan R, Zhang H, et al. Application of radiomics and machine learning in head and neck cancers. Int J Biol Sci. 2021;17:475-86. doi:10.7150/ijbs.55716.

30. Chiu HY, Chao HS, Chen YM. Application of Artificial Intelligence in Lung Cancer. Cancers (Basel). 2022;14. doi:10.3390/cancers14061370.

31. Shur JD, Doran SJ, Kumar S, Ap Dafydd D, Downey K, O'Connor JPB, et al. Radiomics in Oncology: A Practical Guide. Radiographics. 2021;41:1717-32. doi:10.1148/rg.2021210037.

32. Zhong J, Hu Y, Ge X, Xing Y, Ding D, Zhang G, et al. A systematic review of radiomics in chondrosarcoma: assessment of study quality and clinical value needs handy tools. Eur Radiol. 2022. doi:10.1007/s00330-022-09060-3.

33. Park CJ, Park YW, Ahn SS, Kim D, Kim EH, Kang SG, et al. Quality of Radiomics Research on Brain Metastasis: A Roadmap to Promote Clinical Translation. Korean J Radiol. 2022;23:77-88. doi:10.3348/kjr.2021.0421.

34. Gardin I, Grégoire V, Gibon D, Kirisli H, Pasquier D, Thariat J, et al. Radiomics: Principles and radiotherapy applications. Crit Rev Oncol Hematol. 2019;138:44-50. doi:10.1016/j.critrevonc.2019.03.015.

35. Ball D, Mai GT, Vinod S, Babington S, Ruben J, Kron T, et al. Stereotactic ablative radiotherapy versus standard radiotherapy in stage 1 non-small-cell lung cancer (TROG 09.02 CHISEL): a phase 3, open-label, randomised controlled trial.

36. Kong FM, Ritter T, Quint DJ, Senan S, Gaspar LE, Komaki RU, et al. Consideration of dose limits for organs at risk of thoracic radiotherapy: atlas for lung, proximal bronchial tree, esophagus, spinal cord, ribs, and brachial plexus. Int J Radiat Oncol Biol Phys. 2011;81:1442-57. doi:10.1016/j.ijrobp.2010.07.1977.

Chapter 2: Methodological Quality of Machine Learning-based Quantitative Imaging Analysis Studies in Esophageal Cancer: A Systematic Review of Clinical Outcome Prediction after Concurrent Chemoradiotherapy

Adapted from: Zhenwei Shi; **Zhen Zhang***; Zaiyi Liu; Lujun Zhao; Zhaoxiang Ye; Andre Dekker; Leonard Wee. Methodological Quality of Machine Learning-Based Quantitative Imaging Analysis Studies in Esophageal Cancer: A Systematic Review of Clinical Outcome Prediction after Concurrent Chemoradiotherapy. European Journal of Nuclear Medicine and Molecular Imaging 2021. https://doi.org/10.1007/s00259-021-05658-9.*

** indicates equal contributions*

Abstract

Purpose Studies based on machine learning-based quantitative imaging techniques have gained much interest in cancer research. The aim of this review is to critically appraise the existing machine learning-based quantitative imaging analysis studies predicting outcomes of esophageal cancer after concurrent chemoradiotherapy in accordance with PRISMA guidelines.

Methods A systematic review was conducted in accordance with PRISMA guidelines. The citation search was performed via PubMed and Embase Ovid databases for literature published before April 2021. From each full-text article, study characteristics and model information were summarized. We proposed an appraisal matrix with 13 items to assess the methodological quality of each study based on recommended best-practices pertaining to quality.

Results Out of 244 identified records, 37 studies met the inclusion criteria. Study endpoints included prognosis, treatment response, and toxicity after concurrent chemoradiotherapy with reported discrimination metrics in validation datasets between 0.6 and 0.9, with wide variation in quality. A total of 30 studies published within the last five years were evaluated for methodological quality and we found 11 studies with at least 6 "Good" item ratings.

Conclusion A substantial number of studies lacked prospective registration, external validation, model calibration, and support for use in clinic. To further improve the predictive power of machine learning-based models and translate into real clinical applications in cancer research, appropriate methodologies, prospective registration and multi-institution validation are recommended.

Keywords: Quantitative imaging analysis; Esophageal cancer; Concurrent chemoradiotherapy; Clinical outcomes; Methodological assessment

Introduction

Esophageal cancer (EC) is the seventh most common malignancy, and the sixth most common cause of cancer-related death worldwide [1]. Prognosis for EC patients remains poor to date, with a five-year survival chance of 20% [2]. Although the histopathology and disease characteristics differ between eastern and western countries due to genetic variations, concurrent chemoradiotherapy (CCRT) plays an important global role in the treatment of EC [3].

The CROSS trial was a landmark study that established the role of neoadjuvant chemoradiotherapy (nCRT), and laid the foundation of nCRT as the standard of care for resectable EC [4]. While CROSS demonstrated that nCRT improved average survival among EC patients and side-effect rates were acceptable, it remains clinically meaningful to select patients that will personally benefit from nCRT versus their probable side effects. Definitive chemoradiotherapy is standard of care for unresectable EC [5]. However, it remains difficult to predict individual outcomes (e.g., treatment response) of any type of CCRT due to tumor heterogeneity between subjects and complex tumor microenvironments within.

Technical advances in radiation delivery such as modulated radiotherapy, image-guidance and scanning proton beams have vastly improved target coverage and avoidance of adjacent healthy organs. It is practically impossible to entirely avoid some unintended damage to nearby organs, which results in radiotherapy complications. A way to predict treatment response and side effects at the earliest step of CCRT works hand in hand with radiotherapy technology and new drug therapies, and this is essential to guide individually personalized treatment, to improve the survival likelihood and to retain high quality of remaining life for EC patients.

The spatial and time heterogeneity of solid tumors at the genetic, protein, cellular, microenvironmental, tissue and organ levels makes it difficult to accurately and representatively characterize a tumor using only invasive sampling methods, such as pathology and molecular imaging examination. Quantitative analysis based on volumetric non-invasive imaging (i.e. radiomics [6-8]) suggests the attractive hypothesis of measuring whole-tumor heterogeneity in vivo. Radiomics makes it feasible to characterize whole-tumor heterogeneity and also monitor tumor evolution over time.

Radiomics requires large volumes of clinical imaging data to be converted into a vast number of numerical features with the assistance of computers, which can then be mined for clinically actionable insights using high-dimensionality machine learning methods. Radiomics includes features that are defined a priori by human operators (i.e “handcrafted” features) as well as purely data-driven features arising via end-to-end training of deep learning neural networks. A number of key studies and evidence syntheses have shown that radiomics has potential to recognize heterogeneity in primary tumors and/or lymph nodes in a variety of cancers with clinical applications for diagnosis and prognostication [9-12].

Within EC, radiomics is presently an active area of original research (e.g., in [13, 14]), but at time of writing there has been no comprehensive PRISMA-compliant (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) systematic review of radiomics specifically addressing methodological robustness and clinical relevance of radiomics for patients with EC treated by CCRT. In this systematic review, we present to the reader a cohesive critical appraisal of research up to date, and a summary of clinical relevance of ra-

diomics as a potential tool for predicting (i) treatment outcomes, (ii) longer term prognosis and (iii) CCRT treatment-related toxicity.

Methods

1. Eligibility Criteria

We conducted this systematic review from May to June 2021, in accordance with PRISMA guidelines [15]. In this study, we included only primary observational studies published between May 2011 and June 2021 using either handcrafted and/or deep learning-based radiomics features extracted from clinical imaging - specifically computed tomography (CT), magnetic resonance (MR) and positron-emission tomography (PET) - to develop clinical prediction models on human primary EC subjects treated by CCRT. Articles eligible for critical appraisal had to be published as full texts in peer-reviewed journals in the English language within the last 5 years.

2. Exclusion Criteria

Diagnostic accuracy studies evaluating tumor differentiation grade or the diagnosis of lesions were excluded. Studies that exclusively addressed modelling on non-radiomic features, such as only standardized uptake value (SUV), clinical parameters, and/or dosimetric parameters, were excluded. Clinical outcomes that were primarily associated with surgery alone, radiotherapy alone, or chemotherapy alone were excluded. Case reports, other (systematic) reviews, conference abstracts, editorials and expert opinion papers were also excluded.

3. Search Methods

An initial citation screening in PubMed and EMBASE electronic databases was performed on 9 May 2021. We used a search string containing Medical Subject Headings (MeSH) or Emtree terms for 'esophageal cancer' combined with other text words that related to outcomes, prediction, model, radiomics (including textural analyses and quantitative analyses), and artificial intelligence. The search filters used are provided in the Supplementary Material **Table S1**. Articles were also included for screening based on prior knowledge of the authors. We searched the reference section of reviewed papers for any additional articles that may have been missed in the electronic databases.

4. Selection Process

Two authors (Z.Z. and L.W.) worked independently on screening PubMed and EMBASE records, based on titles and abstracts alone. Candidate articles were combined then any disagreements were resolved by consensus; a third author (Z.S.) was available for adjudication but was not required. Full text of the candidate articles were obtained using an institutional journal subscription, and examined in detail for eligibility against the aforementioned criteria. Only full-text articles unanimously deemed eligible for review were then included for detailed data extraction and critical appraisal.

5. Data Extraction

Two authors (Z.S. and Z.Z.) independently performed extraction of publication details and clinical outcomes. From the eligible articles, information pertaining to general study char-

acteristics were extracted (author, publication year, primary cancer type, imaging protocol, treatment modality, sample size) together with radiomics feature-related descriptions (deep learning-based or/and handcrafted features, software used for feature extraction, and whether radiomics features were combined with non-radiomics predictors). Model characteristics and primary reported findings of the included studies were also extracted and summarized, which included use of retrospectively/prospectively collected patient personal data, the collaborating institution(s), sample sizes used to build the model, number of radiomics features initially considered versus that retained in the final model, type of model assessed, the reported performance metrics, and results of model calibration if given.

6. Methodological robustness

Classical evaluation tools such as Quality in Prognostic Studies (QUIPS) for prognostic studies [16], Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) for diagnostic tests [17], and Prediction model Risk Of Bias ASsessment Tool (PROBAST) [18] were not specifically designed for high-dimensional predictive modelling studies such as radiomics. Lambin et al. [19] proposed a radiomics quality score (RQS) that assigned "points" to various steps in radiomics modelling workflow, and such RQS evaluation approach has been previously used [20-24] in reviews. However, specialist evidence synthesis communities (such as the Cochrane Collaboration), advise that a single numerical score may not be appropriate to capture a complex question such as overall methodological robustness of a diagnostic/prognostic model. Other reviewers have also used Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) [25] type as a surrogate measure for quality, but it must be re-emphasized that TRIPOD is a model reporting guideline, not in fact a critical appraisal checklist.

In this work, we have applied an assessment metric guided by the RQS together with findings of other radiomics methodological evaluations [26, 27]. Due to the rapid changes in machine learning and radiomics expertise in the relevant scientific community, we limited the methodological quality appraisal to the included studies published within the past five years. The appraisal was initially performed independently by two authors (Z.S. and Z.Z.) then combined. Disagreements were resolved by consensus, and an experienced senior author (L.W.) adjudicated on differences of evaluation. Each methodological criterion was provided a consensus rating of "good", "moderate" or "poor", based on 13 specific quality criteria :

1. It would have been ideal if a detailed study protocol with its statistical analysis plan had been prospectively registered in an open access registry prior to commencement. Studies that used prospectively collected patient data was rated as "moderate" since the study plan would probably have been registered during internal ethical review. Absence of any of the above was deemed "poor".

2. For reproducibility and comparison between institutions, it is important to provide detailed information that documents the image acquisition conditions. Typical information might include scanner make/model, scan protocol, enhanced/unenhanced CT scans, tube voltage, tube current, slice thickness, voxel size, etc. appropriate to the imaging modality examined. Partial or incomplete information was rated "moderate", but its absence in text or supplemental was deemed "poor".

3. It is widely known that digital image preprocessing steps can strongly influence

the quantitative image analysis results that follow. Studies that give detailed information to reproduce the pre-processing steps (typically includes filters for de-noising, intensity normalization, voxel resampling, etc.). Partial or incomplete information was rated “moderate”, but its absence in text or supplemental was deemed “poor”.

4. The method by which the region of interest (ROI) for analysis has been defined can also influence the generalizability of radiomics models. For instance, automated or semi-automated delineation of organs may be more consistent than manual delineation. A “good” score was given for full information on ROI delineations, including review by experienced experts and/or any inter-observer sensitivity checks. Partial information or no information were scored “moderate” and “poor”, respectively.

5. Radiomics studies typically consider a massive number of features relative to the sample size and the event rate of the outcome of interest, therefore feature selection / dimensionality reduction steps are generally needed to reduce risk of overfitting. We deem that reproducibility and repeatability tests of feature stability, and/or unsupervised feature selection methods (such as principal components analysis or clustering), prior to applying supervised learning with the outcome of interest, would be “good”. Partial documentation or inadequately justified methods were deemed “moderate”, otherwise “poor” when there was a high risk of either over-fitting or false positive association.

6. Potential correlations should be examined between radiomics and non-radiomics (other biological) features, since this can identify possible confounders and justify the added value of imaging features. Adequeate checks for possible correlations are deemed “good”, insufficient or limited checks as “moderate”, or if such checks were not attempted then “poor”.

7. Since the general idea of a prognostic model is to permit stratification of patients, it is important for studies to provide clear justification for defining risk groups, including how risk thresholds and optimum operating points had been determined. Stratification based on clinical argumentation, or agnostically using median or standard cutoffs (e.g. class probability of 0.5) were deemed “good”. Use of optimally “tuned” cutoffs or deriving risk groups as part of the model optimization step can introduce some loss of robustness, and were thus deemed “moderate”. No justification or lack of documentation in this regard were scored as “poor”.

8. As emphasized by TRIPOD, model performance should be evaluated with an external validation cohort, ideally with fully independent researchers, scanners, delineations, etc. Model performance metrics with strong support in external validation (TRIPOD type III) would have been rated as “good”. Validation by non-random split from the training cohort (eg by time, location, or some other pre-treatment characteristic) or by multiple repeated random sampling (k-folds, bootstrapping) were rated “moderate”. However, one-time random sampling or no report of model validation at all were rated as “poor”.

9. Models utilizing radiomics features should be able to show added value when compared against, or combined with, clinical and/or non-radiomics models. We defined the presence of sufficient description about comparison with clinical/non-radiomics model or holistic models as “good”, inadequate comparison as “moderate”, and otherwise as “poor”.

10. Model performance should be reported in terms of appropriate discrimination met-

rics, such as c-index for time-to-event models and AUC for binary classification models. A study was deemed "good" if it reported discrimination metrics for training and test dataset (or other related metrics) together with confidence intervals and statistical significance. Partial information about discrimination was deemed "moderate", or if no information was provided then "poor".

11. As recommended in TRIPOD, model calibration should also be reported in addition to its discriminative performance. A "good" study provided a test of calibration or goodness-of-fit results, together with a calibration figure. Partial information about calibration was deemed "moderate", or if no calibration results were given then "poor".

12. For ease of implementation, studies should discuss the potential clinical utility of their model(s) and provide some justification for use, such decision curves analysis or cost-benefit analysis. We defined the presence of an estimated clinical utility as "good", partial or inadequate analysis as "moderate", and otherwise as "poor".

13. Studies should report parameters of their model(s) in ample detail to permit independent external validation. Those studies rated "good" provided the reader with regression coefficients for each feature or otherwise made it possible to calculate risk scores, such as making their model(s) accessible via an online repository or by providing a calculation aid (e.g. a nomogram). Studies that only reported features selected in the final model were deemed "moderate", however studies that did not provide adequate information to independently validate the model were rated "poor".

7. Objectives

The primary objective was to estimate the overall ability of radiomics models, or models containing some radiomics information, to predict clinical outcomes that are of particular clinical interest in CCRT for EC. This gives us a picture of the current status of clinical readiness of radiomics as a potential tool for clinical decision-making and/or possible incorporation of radiomics-powered models into holistic decision support systems. Secondly, we included a critical appraisal of reported model performance against the methodological robustness (i.e. internal validity) because this is key for understanding its clinical applicabilty, and such robustness informs the degree of wide generalizability (i.e. external validity) that might be expected from a reported model.

Results

1 Literature search results

A PRISMA flowchart diagram illustrating article selection is shown in Figure 1. A total of 384 records were identified based on the specified search terms (MEDLINE/PubMed n=196, EMBASE n=187, and one was found in the cited references of an included article). After duplicates removal, there were 245 articles available for screening. Applying the selection criteria led to 52 studies for full-text screening. At the end, a total of 37 articles were deemed eligible [28-64], including 30 articles within five years [28-38, 41-43, 45, 47-50, 52-54, 57-64]

2 Overall characteristics of included studies

Table 1 and Supplementary Material **Table S2** summarizes the general characteristics across all included studies. The majority (20 of 37) of studies combined both esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAD) patients. There were 13 studies conducted exclusively on ESCC patients but only two studies on EAD patients alone. Two other studies did not actually mention the histopathology type of the cohorts studied.

The majority of imaging modalities mentioned in the retrieved studies were PET (20/37) [28, 30, 34-40, 44-47, 49, 50, 52, 55, 56, 59, 61], CT (16/37) [29, 31-33, 41, 43, 48, 51, 53, 54, 57, 58, 60, 62-64], and one cone beam CT (CBCT) [42]. Although the search criteria included MRI, we did not locate any eligible study in our search.

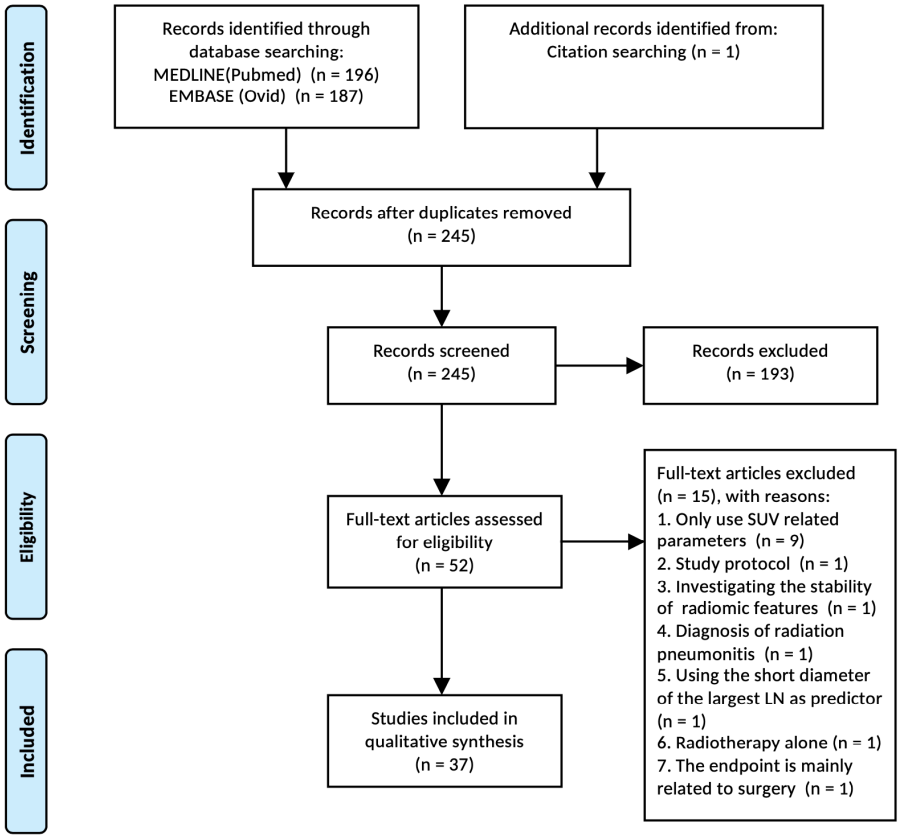


Figure 1. Flowchart of the literature search and study selection (PRISMA 2009 [65]).

More than half of the included studies (19/37) addressed nCRT [28-30, 33, 35, 38, 40, 43-47, 49, 52, 54, 56, 61-64]. The majority of patients included in 13 studies were treated specifically with radical CCRT [31, 32, 36, 39, 41, 42, 48, 50, 51, 53, 55, 58, 59]. In three studies, most patients were treated with CCRT, but the rest received a variety of different treatments depending on their situation [34, 57, 60]. There was one study that did not speci-

fy the intent of CCRT [37].

The number of patients reported in the included studies ranged from 20 [40, 44, 52, 56] up to 464 [60]. Three studies utilized deep learning [46, 53, 64] and all other studies used only handcrafted features with Cox proportional hazards, logistic regression (LR), linear regression, support vector machine (SVM) and random forest (RF) models.

There were a wide range of software tools used to extract radiomics features. The in-house codes were predominantly generated in Matlab and Python. The most commonly used [31, 33, 41, 42] free and open-source software package was 3D Slicer [66], which allowed for manual or semi-automatic ROI delineation followed by radiomics features extraction using its Radiomics [67] plug-in. Studies using Python and 3D Slicer were almost exclusively based on the pyradiomics library [67] developed by Griethuysen et al. Five studies investigated exclusively radiomics features [29, 32, 46, 53, 57], while the other studies examined a combination of radiomics with non-radiomics features (most commonly, clinical factors). In this review, classical PET features were defined as intensity-related metrics such as standardized uptake value (SUV), metabolic tumor volume (MTV), and total lesion glycolysis (TLG). There were 8, 7 and 10 studies that combined radiomics with clinical features [33, 41, 43, 47, 51, 54, 58, 60], classical PET features [39, 44, 52, 55, 56, 59, 61], and both clinical and classical PET features [30, 34-38, 40, 45, 49, 50], respectively. Among more recently published studies, three included genes as features [28, 63, 64], two included clinical factors with dosimetric features [42, 48], one included histopathologic features [62], and one used a combination of clinicopathological, dosimetric, and hematological features [31].

3 Overall characteristics of included studies

The model results from the included studies are summarized in **Table 2** and additional details added in Supplementary Material **Table S2**. Patient data were mostly retrospectively extracted (31/37). Only four studies re-analyzed prospectively collected data, which all originated in the CROSS clinical trial [35, 45, 47, 49]. Three studies used both prospective and retrospective data, where the prospective data were also re-analyzed from other clinical trials [35, 47, 63]. One study did not describe if the data used was retrospectively or prospectively derived [46].

There were few multi-institute studies in general. The majority of studies (27/37) were performed within a single institution. Nine studies incorporated data from two distinct institutes, and one study incorporated data from three distinct institutes.

Study endpoints were broadly classified into three categories: (1) prognosis (9/37), such as overall survival (OS), progression-free survival (PFS) and disease-free survival (DFS), (2) treatment response (20/37), such as prediction of complete/partial response after radical CCRT, and pathology complete response (pCR) after nCRT, and (3) others, such as prediction of lymph node status [47] and radiation pneumonitis (RP) [31, 42]. There were five studies that reported both prognosis and treatment response prediction [30, 32, 37, 50, 59].

The number of events of the included studies ranged from 9 [52] to 113 [34], and the number of radiomics features in the final model ranging from only one [60, 62] up to 40 [43]. Overall, the number of events were small relative to the number of selected features. The number of positive events from studies predicting treatment-related side effects were overall much smaller than those predicting prognosis, which was consistent with real-world in-

cidences.

The most frequently used model was Cox regression, followed by logistic regression. The most widely used machine learning approach was SVM (n=7) but there was high heterogeneity in mathematical procedures. The deep learning architectures used were artificial neural networks (ANN) in one study [53], and convolutional neural networks (CNN) in two studies [46, 64], respectively.

Model performance had been summarized according to different study endpoints. For prognosis, some studies grouped patients by clustering only. Studies that reported the discriminative performance of the models had c-indices ranging from 0.64 [60] to 0.875 [63], and AUCs ranging from 0.69 [43] to 0.918 [63] in the training set. As expected, the discriminative performance overall decreased in the validation/test cohort, with c-indices ranging from 0.57 [60]-0.719 [63] and AUCs between 0.61 [43, 60] to 0.805 [57] in the validation/test set.

For treatment response, reported AUCs were from 0.685 [28] to 1.0 [40] in training set but decreased overall in the validation/test sets (AUCs 0.6 [53] to 0.852 [29]). AUCs in the training and validation sets for the prediction of lymph node metastases study were 0.82 and 0.69 [47], respectively, and the AUCs in the validation set for the prediction of RP study were 0.921 [31] and 0.905 [42]. Except for RP, the validation set AUCs were roughly in the range of 0.6-0.8. Only six studies performed model calibration, four of which used the Hosmer-Lemeshow test for goodness of fit [28, 45, 47, 49].

Table 1. Summary of general study characteristics.

Ref.	Cancer type (recruitment period)	Imaging modality	Imaging acquisition settings	Treatment	Sample size	Type of features	Radiomics software	Non-radiomics cofactors
Xie: 2021 [63]	ESCC 2007-2016	CT	Inst 1: 120KVp, 200-400mA, 2.5mm slices; Inst 2: 120KVp, 200-300mA, 5mm	nCRT	65 (train) 41 (test)	HF	Pyradiomics	Genetic
Beukinga: 2021 [28]	ESCC and EAD 2010-2018	PET/ CT	Gaussian filter of 6.5 mm in full-width at half-maximum	nCRT	96 (ESCC: 88 EAD: 8)	HF	In-house (Matlab V2018a)	Clinical factors, HER2 and CD44
Hu: 2021 [64]	ESCC 2007-2018	CT	Same as Hu:2020	nCRT	161 (train) 70 (test)	HF and DLF	PyRadiomics (V2.1.2)	No
Wang: 2021 [31]	ESCC and EAD 2012-2018	CT	120KVp, 200mA, 3mm	dCCRT	200 (train, ESCC: 189, EAD: 11) 200 (val., ESCC: 195, EAD: 5)	HF	3D Slicer (V4.8.1)	Clinicopathological, dosimetrics, and hematological
Li: 2020 [36]	ESCC Train 2009-2013 Val. 2015-2018	PET/CT	Voxel size: 4 × 4 × 5 mm3	dCCRT	152 (train) 32 (val.)	HF	PyRadiomics (V2.0.1)	Clinical and classical PET
Xie: 2020 [58]	ESCC 2008-2014	CT	120kV, 180-280mA, 3mm	CCRT	57	HF	IBEX (V1.0β)	Clinical factors
Hu: 2020 [29]	ESCC 2007-2018	CT	120KV, 200-400mA 2.5mm (inst 1) 5mm (inst 2) voxel sizes: 1×1×5 mm3	nCRT	161 (train) 70 (test)	HF	PyRadiomics (V3.0)	No
Luo: 2020 [41]	ESCC 2013-2015	CT	120 kV, 120 mAs, 5 mm	dCCRT	160 (train) 66 (val.)	HF	3DSlicer (V4.10.2)	Clinical factors
Li: 2020 [54]	ESCC 2012-2019	CT	120kV/140kV, 140-300mA, 5 mm	nCRT	121	HF	IBEX	Clinical factors

Zhang: 2020 [47]	EAD 2010-2016	PET/ CT	120 kVp, 20-200 mA	surgery alone, neoadjuvant chemotherapy, and nCRT	190	HF	Matlab	Clinical factors
Du: 2020 [42]	ESCC 2017-2019	CBCT	125 kVp, 80mA, 13 ms, 680mAs, pixel size: 384 ×384, 2.5mm, half-fan CBCT	dCCRT or definitive radiotherapy	67 (train) 29 (val.)	HF	3D Slicer (V4.10.2)	Clinical and dosimetrics
Foley: 2019 [35]	ESCC and EAD 2010-2015	PET/ CT	Same as Foley:2018	Same as Foley:2018	46 (external val.)	HF	In-house (Matlab)	Clinical and classical PET
Xie: 2019 [57]	ESCC Train 2012-2016 Val. 2008-2011	CT	Inst 1: 120 kVp, 406 mAs, 3-5 mm; Inst 2: 120 kVp, 150 mAs, 3–8 mm; Voxel size: 1 × 1 × 5 mm3	dCCRT	87 (train) 46 (val.)	HF	In-house (Matlab 2015b)	No
Wang: 2019 [60]	ESCC Train 2012-2016 Val. 2004-2014	CT	120kV, 180-280mA, 3 mm	CCRT and RT alone	83 (train) 98+283 (val.)	HF	IBEX (V1.0β)	Clinical
Chen: 2019 [30]	ESCC 2011-2017	PET/ CT	PET scanner: 120KV, 12mA, 3.75mm	dCCRT	44	HF	CGITA	Clinical and classical PET
Yan: 2019 [32]	ESCC 2013-2017	CT	120kVp, 4mm	nCRT	32	HF	CUBETAB (Matlab V2017b)	None
Yang: 2019 [33]	ESCC 2012-2016	CT	120 kVp, pixel size: 1.46mm, 5 mm	nCRT	44 (train) 11 (test)	HF	3DSlicer (V4.8.1)	Clinical factors
Jin: 2019 [48]	ESCC, EAD, and Small cell 2012-2015	CT	120 kV, 180–280 mA, 3mm	CCRT	94 (ESCC: 92, EAD: 1, Small cell: 1)	HF	IBEX	Clinical and dosimetrics
Foley: 2018 [34]	ESCC and EAD	PET/ CT	PET: 120 kVp, 20-200 mA	Multiple treatments incl. nCRT and dCCRT	302 (train, ESCC: 65 EAD: 237)	HF	In-house (Matlab)	Clinical and classical PET

	Train 2010-2014 Val. 2014-2015				101 (val., ESCC: 79 EAD: 22)			
Larue: 2018 [43]	ESCC (n=46) and EAD (n=193) 2010-2016	CT	Inst 1: 120 kV, 2.5–5 mm; Inst 2: 120 kV, 1–3 mm Voxel size: 1 × 1 × 3 mm3	nCRT	165 (train) 74 (val.)	HF	In-house (Matlab)	Clinical
Beukinga: 2018 [49]	ESCC and EAD 2014-2017	PET/CT	80–120 kV, 20–35 mAs, 5 mm	nCRT	73 (ESCC: 8, EAD: 65)	HF	In-house (Matlab V2014b)	Clinical and classical PET
Riyahi: 2018 [52]	ESCC and EAD 2006-2009	PET/ CT	Same as Tan:2013	Same as Tan:2013	Same as Tan:2013	HF	Elastix and ITK toolbox	Classical PET features
Paul: 2017 [37]	n.r.	PET/ CT	Voxel size: 4 × 4 × 2mm3	CCRT	65	HF	n.r.	Clinical and classical PET
Desbordes: 2017 [50]	ESCC and EAD 2006-2013	PET/ CT	Voxel size: 4 × 4× 2 mm3	CCRT	65 (ESCC: 57 EAD: 8)	HF	n.r.	Clinical and classical PET
Nakajo: 2017 [59]	n.r. 2011-2013	PET/ CT	120 kV, 35-100 mA, 3.75 mm	CCRT	52	HF	In-house (Python)	classical PET features
Beukinga: 2017 [45]	ESCC and EAD 2009-2016	PET/ CT	PET: 0.98 × 0.98 mm, 2 mm; CT: 0.98 × 0.98 mm, 3 mm	nCRT	97 (ESCC: 9, EAD: 88)	HF	n.r.	Clinical and classical PET
Wakatsuki: 2017 [62]	ESCC and EAD 2008-2015	CT	120kV, 5 mm	nCRT	50 (ESCC: 46, EAD: 4)	HF	Unnamed	Clinical and histopathologic
Hou: 2017 [53]	ESCC 2015-2016	CT	120 kV, 200–250 mAs, 2.5-3 mm, pixel size: 0.97 × 0.97 mm	dCCRT	37 (train) 12 (test)	HF	In-house (Matlab 2015a)	No
Yip: 2016 [61]	ESCC and EAD	PET/ CT	n.r.	nCRT	45 (ESCC: 1, EAD: 44)	HF	CGITA	Classical PET features
Rossum: 2016	EAD 2006-2013	PET/ CT	CT: 120 kV, 300 mA, 3.75 mm, voxel size: 5.47 x 5.47 x 3.27 mm	nCRT	217	HF	IBEX	Clinical and classical PET

[38]								
Ypsilantis: 2015 [46]	ESCC and EAD n.r.	PET/CT	3.27 mm, pixel size: 4.7 × 4.7 mm.	nCRT	107 (ESCC: 20, EAD: 86, Undefined: 1)	HF/DLF	n.r.	No
Yip: 2014 [51]	ESCC and EAD 2005-2008	CT	120 kV, 180–280 mA, 3-5 mm	dCCRT	36 (ESCC: 26 EAD: 9 Not specified:1)	HF	TexRAD	Clinical
Zhang: 2014 [40]	ESCC and EAD 2006-2009	PET/ CT	Same as Tan:2013	nCRT	20 (ESCC: 3, EAD: 17)	HF	n.r.	Clinical and classical PET
Tan: 2013 [44]	ESCC and EAD 2006-2009	PET/ CT	120 kV, 200 mA, 0.98 × 0.98 × 4 mm3 (CT) 4 × 4 × 4 mm3 (PET)	nCRT	20 (ESCC: 3, EAD: 17)	HF	n.r.	Classical PET features
Hatt: 2013 [55]	ESCC and EAD 2004-2008	PET/ CT	120kV, 100mAs (CT) PET voxel size: 4 × 4 × 4 mm3	CCRT	50 (ESCC: 36, EAD: 14)	HF	n.r.	Classical PET features
Tan: 2013 [56]	ESCC and EAD 2006-2009	PET/ CT	Same as Tan:2013	nCRT	20 (ESCC: 3 EAD: 17)	HF	ITK	Classical PET features
Tixier: 2011 [39]	ESCC and EAD 2003-2008	PET/ CT	n.r.	CCRT	41 (ESCC: 31 EAD: 10)	HF	n.r.	Classical PET features

Abbreviations used in the table – n.r.: not reported; val.: validation; ESCC: esophageal squamous cell carcinoma; EAD: esophageal adenocarcinoma; nCRT: neoadjuvant chemoradiotherapy; CCRT: concurrent chemoradiotherapy; dCCRT: definitive concurrent chemoradiotherapy; RT: radiotherapy; CT: computed tomography; CBCT: cone-beam computed tomography; HF: handcrafted features; DLF: deep learning-based features.

Table 2. Summary of radiomics-based prediction model characteristics described in included studies.

Ref.	Data type	# of institution(s)	Predicted outcome(s)	# of events/# of samples	# of features (considered / in final model)	Type of model	Reported performance	Model calibration tested
Xie: 2021 [63]	R + P	2	DFS	Train: 21/28 Int. validation: 24/37 External test: 13/41	2553/8	Cox	(train, validation and external test) AUC= 0.912, 0.852, and 0.769; C-index= 0.869, 0.812, and 0.719	Yes
Beukinga: 2021 [28]	R	1	pCR after nCRT	Group 1: 21/96 Group 2: 9/43	101/2	LR	AUC = 0.685 and 0.857 (Best of group 1 and group 2)	Yes
Hu: 2021 [64]	R	2	pCR after nCRT	Train: 74/161 Test: 31/70	Handcrafted features: 851/7 Handcrafted combined with deep learning-based: n.r./14	SVM	Handcrafted model: AUC=0.822, and 0.725 (train and test) Deep learning-based: AUC=0.807-0.901, and 0.635-0.805 (train and test)	Yes
Wang: 2021 [31]	R	2	RP	Train: 45/200 Val.: 41/200	850/24	Linear regression	C-index= 0.975, and 0.921 (internal and external val.)	Yes
Li: 2020 [36]	R	2	OS, DFS, LC	n.r./184	DFS: 105/3 OS: 105/4 LC: 105/4	Cox	Clustering of OS: p<0.0001	No
Xie: 2020 [58]	R	1	OS	1-year survival: 43/57	16/4	Cox	1-year and 2-year survival: AUC=0.79	No
Hu: 2020 [29]	R	2	pCR after nCRT	Train: 74/161 Test: 31/70	Intratumoral: 1208/16 Peritumoral: 1036/8 Combined model: 7 (intra) and 6 (peri)	8 different types of models	Combined model AUC=0.906, and 0.852 (train and test)	Yes
Luo: 2020 [41]	R	1	CR after CCRT	Train: 56/160 Val.: 22/66	851/7	LASSO-LR	AUC=0.844, and 0.807 (train and val.)	No
Li: 2020 [54]	R	1	pCR after nCRT	51/121	405/18	LR	AUC = 0.84 (val.)	Yes
Zhang: 2020 [47]	R + P	2	Clinical lymph node staging	Train: 75/130 Val.: 35/60	154/9	LR	AUC=0.82, and 0.69 (train and val.)	Yes
Du: 2020 [42]	R	1	RP	39/96	851/2	LR	AUC = 0.836, and 0.905 (train and val.)	Yes
Foley: 2019 [35]	R + P	2	OS	External val.: 26/46	16/3	Cox	X2 = 1.27, df = 3, p = 0.74 (Kaplan-Meier)	Yes

Xie: 2019 [57]	R	2	OS	Train: 26/87 Val.: 9/46	548/7	Cox	AUC=0.811 (Train) AUC=0.805 (Val.)	No
Wang: 2019 [60]	R	3	OS PFS	Train: 23/83, Val.1: 18/98, Val.2: 53/283 Train: 21/83, Val.1: 8/98, Val.2 36/283	1/1	Cox	(Train, Val. 1 and 2) OS: C-index= 0.64, 0.61, and 0.58 PFS: C-index= 0.66, 0.60, and 0.57	No
Chen: 2019 [30]	R	1	pCR after nCRT, DFS, OS	nCRT response: 17/42	nCRT response 23/1	n.r.	Clustering response to nCRT: p=0.009	No
Yan: 2019 [32]	R	1	CR after RT survival	CR: 22/32	CR: 10/4 Survival: 10/2	n.r.	RT response: P <0.0001 Survival: r = 0.9917, P = 0.0001	No
Yang: 2019 [33]	R	1	pCR after nCRT	Train: 19/44 Test: 4/11	1030/5 (Model 1), 6 (Model 2/3)	LR	Model 1(bin size=32): 0.86, and 0.79 (train and test)	No
Jin: 2019 [48]	R	1	response to CCRT	58/94	42/n.r.	SVM, XGBoost	AUC=0.689	No
Foley: 2018 [34]	R	1	OS	Train: 70/302 Test: 43/101	16/3	Cox	X2 143.14, df 3, p < 0.001(Train) X2 20.621, df 3, p < 0.001(Val.)	No
Larue: 2018 [43]	R	2	OS	Train: 67/165 Val.: 25/74	1049/40	RF	AUC= 0.69 (Train) AUC = 0.61 (Val.)	No
Beukinga: 2018 [49]	P	1	pCR after nCRT	16/73	113/6	LASSO-LR	AUC=0.82 and 0.81 (train and val.)	Yes
Riyahi: 2018 [52]	R	1	pCR/mRD after nCRT	9/20	664/2	SVM-LASSO	AUC=0.94±0.05	No
Paul: 2017 [37]	R	1	CR after CCRT, OS	CR: 41/65 OS: 16/65	CR: 45/9 OS: 45/8	RF	CR: AUC=0.823±0.032 OS: AUC= 0.750±0.108	No
Desbordes: 2017 [50]	R	1	CR after CCRT, 3-years OS	CR: 41/65 OS: 24/65	45/1	RF	CR: AUC = 0.836±0.105 OS: AUC=0.822±0.059	No
Nakajo: 2017 [59]	R	1	CR/RP after CCRT, PFS, OS	CR: 18/52	CR 6/2 PFS and OS 6/0	Cox	CR: AUC=0.75 PFS and OS: P <0.001	No
Beukinga: 2017 [45]	P	1	pCR after nCRT	19/97	140/20	LR	AUC=0.78, and 0.74 (train and val.)	Yes
Wakatsuki: 2017 [62]	R	1	response to nCRT	17/50	1/1 CT number	LR	AUC=0.73, P=0.009	No

Hou: 2017 [53]	R	1	CR/PR after CCRT	Train: 26/37 Test: 7/12	SVM: 214/9 ANN: 214/7	SVM, ANN	ANN: accuracy=0.972, and 0.917; AUC=0.927, and 0.800 (train and test) SVM: accuracy=0.891, and 0.667; AUC=0.818, and 0.600 (train and test)	No
Yip: 2016 [61]	R	1	response to nCRT	30/45	3/3	n.r.	AUC = 0.72–0.78	No
Rossum: 2016 [38]	R	1	pCR after nCRT	59/217	78/9	LR	c-index=0.82 (apparent) c-index=0.77 (corrected)	Yes
Ypsilantis: 2015 [46]	n.r.	1	response to nCRT	38/107	85/n.r.	LR, gradient boosting, RF, SVM, CNN	Accuracy: 73.4±5.3	No
Yip: 2014 [51]	R	1	OS	5/36	6/4	Cox	AUC=0.802	No
Zhang: 2014 [40]	R	1	pCR/mRD after nCRT	9/20	137/14	SVM, LR	AUC=1 (no misclassifications)	No
Tan: 2013 [44]	R	1	pCR/mRD after nCRT	9/20	16+19/2+16	n.r.	Texture feature: AUC=0.83, p=0.01; histogram distances: AUC=0.78-0.89, p=0.04	No
Hatt: 2013 [55]	R	1	CR/PR after CCRT	36/50	9/9	n.r.	(best) AUC=0.90	No
Tan: 2013 [56]	R	1	pCR/mRD after nCRT	10/20	33/2	n.r.	(best) AUC=0.85	No
Tixier: 2011 [39]	R	1	CR/PR after CCRT	CR: 9/41 PR: 21/41	38/4	n.r.	Sensitivity: 76%-92% Specificity: 56%-91%	No

Abbreviations used in the table – #: number; R: Retrospective; P: Prospective; OS: overall survival; DFS: disease-free survival; PFS: progression-free survival; LC: local control; pCR: complete pathologic response; mRD: microscopic residual disease; SVM: support vector machine; RF: Random Forest; RT: radiotherapy; CR: complete responders; PR: partial responders; LASSO: least absolute shrinkage and selection operator; LR: logistic regression; XGBoost: extreme gradient boosting; ANN: artificial neural network; CNN: convolutional neural network; AUC: Area Under the Receiver Operating Characteristic Curve; RT: radiotherapy; nCRT: neoadjuvant chemoradiotherapy; CCRT: Concurrent chemoradiotherapy; RP: radiation pneumonitis.

4 Methodological quality of the included studies

Given the rapid advances in AI tools and radiomics expertise, we restricted the assessment of methodological quality of recent radiomics studies published in the last five years [28-38, 41-43, 45, 47-50, 52-54, 57-64]. **Table 3** provides an overview of the distribution of methodological quality and reporting completeness of 30 recent studies. A detailed report of quality assessment by the authors has been provided in Supplementary Material **Table S3**.

No study had been prospectively registered prior to commencement of the radiomics analysis. Among the 13 methodological items considered, around one-third of the studies reported essential details about image acquisition settings (12/30 rated good), digital image preprocessing (only 7/30 rated good) and how ROIs were derived (11/30 rated good).

In terms of feature selection, 11/30 studies evaluated repeatability/reproducibility of individual features and/or performed well-justified dimensionality reduction prior to fitting the final model. Ten studies tested the relationship between radiomics and non-radiomic features; out of which 4 showed an association between radiomic features and PET uptake measures [36, 50, 59, 61], another 4 showed the relationship between radiomics and gene expression [29, 62-64], and the next 2 evaluated correlation between radiomics and clinical features [57, 60].

For elements related to reporting model performance, discrimination metrics in training and validation, with confidence intervals, were mostly reported well (16/30 studies), but fewer studies also included a check for model calibration (12/30 studies). Half (15/30 studies) defined clinically-appropriate risk groupings and four studies used median [32, 58] or quartiles [34, 35] as risk group cut-offs, but two studies did not specify how risk groups were obtained [36, 60]. A few (5/30 studies) used ROC curves to obtain optimally-tuned cut-offs (eg Youden index).

For model validation, we found 10/30 studies used multi-institutional data, and 9/30 used internal cross-validation with some form of random splits of data, of which 5/30 studies used bootstrap methods ranging from 1000 to 20,000 replicates.

In regards to clinical impact, relatively few studies (8/30) estimated the clinical impact of their models, including use of decision curve analysis. Only 3 studies reported on all of model discrimination, model performance and clinical utility in the same time [31, 42, 63]. The majority of radiomics studies (22/30) had been compared against non-radiomics models and/or constructed combined models.

As for documentation of the final prognostic model to a degree that permitted independent external validation, only 16/30 studies were rated as good. One study failed to reported on the features selected in the final model. However, none of these 30 studies made their models or analysis code available for download from an electronic repository.

We further observed that methodological aspects among recent studies for predicting prognosis was generally somewhat better than for studies aiming to predict treatment response. Eleven studies were rated “good” for at least 6 out of 13 assessment items, whereas five studies of PFS or/and OS [35, 36, 57, 60, 63], four studies predicted treatment response (pCR after nCRT) [29, 38, 54, 64], and two studies predicting RP [31, 42] were of similar ratings. The best rating among these studies was scored “good” for 11 out of 13 items [64].

Figure 2 visually summarises the headline reported discrimination metric (AUC or c-indices) with the number of methodological items rated "good" in this review. Additionally, we have colour-coded the dots to correspond to the TRIPOD type of study. A small number of methodologically strong studies near the top of the figure suggest a discriminative performance around 0.8 to 0.92 for radiomics prognostic models in EC, followed by a wider scatter of performance metrics for models of lower methodological rigour ranging from 0.61 up to 0.94. Interestingly, this overview found no models with a discriminative index lower than 0.6. The highest reported discrimination metric however coincides with a study of questionable methodological robustness. Overlaid above this, there is a clear trend of TRIPOD type 3 or 4 sudy designs obtaining higher methodological robustness ratings than TRIPOD type 1B, 2A or 2B, with TRIPOD type 1A study designs tending towards the lower methodological ratings. A detailed description of different types of prediction model studies covered by TRIPOD statement can be found in Reference [68].

Figure 2 Reported AUC/C-index of the included studies with number of good items were classified by Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Type 1a: Development only; Type 1b: Development and validation using resampling; Type 2a: Random split-sample development and validation; Type 2b: Non-random split-sample development and validation; Type 3: Development and validation using separate data; Type 4: Validation only.

Table 3. Assessment of methodological quality of included studies.

Number	Reference	Prospective registration	Imaging protocol	Image pre-processing	Segmentation method	Repeatability, reproducibility, and dimensionality	Correlations with non-radiomics biomarkers	Justification of risk groupings	Validation method	Compare to non-radiomics features or build Holistic	Discrimination statistics	Model calibration	Estimation of clinical utility	information for external validation	Number of items rated good
1.	Xie et al., 2021 [63]	🔴	🟢	🔴	🟡	🟢	🟢	🟡	🟢	🟢	🟢	🟢	🟢	🟡	8
2.	Beukinga et al., 2021 [28]	🔴	🔴	🟡	🟢	🟡	🔴	🔴	🟡	🟢	🔴	🟢	🔴	🟡	3
3.	Hu et al., 2021 [64]	🔴	🟢	🟢	🟢	🟢	🟢	🟡	🟢	🟢	🟢	🟢	🟢	🟢	11
4.	Wang et al., 2021 [31]	🔴	🟢	🔴	🟢	🟢	🔴	🟡	🟢	🟢	🟢	🟢	🟢	🟢	9
5.	Li Yimin et al., 2020 [36]	🔴	🟢	🟢	🟡	🟡	🟢	🟡	🟢	🟢	🔴	🔴	🔴	🟢	6
6.	Xie et al., 2020 [58]	🔴	🟡	🟡	🟢	🔴	🔴	🟢	🔴	🔴	🟢	🔴	🔴	🟢	4
7.	Hu et al., 2020 [29]	🔴	🟢	🟢	🟢	🟢	🟢	🟡	🟢	🟢	🟢	🟢	🟢	🟡	10
8.	Luo et al., 2020 [41]	🔴	🟡	🔴	🟡	🟢	🔴	🔴	🔴	🟢	🟡	🔴	🟢	🟢	4
9.	Li Yue et al., 2020 [54]	🔴	🟢	🔴	🟢	🟢	🟡	🔴	🟡	🟢	🔴	🟢	🟢	🟢	7
10.	Zhang et al., 2020 [47]	🔴	🟢	🟡	🟡	🟡	🔴	🔴	🟢	🟢	🟢	🟢	🔴	🟡	5
11.	Du et al., 2020 [42]	🔴	🟡	🔴	🟢	🟢	🔴	🔴	🟡	🟢	🟢	🟢	🟢	🟢	7
12.	Foley et al., 2019 [35]	🔴	🟢	🟡	🟡	🟡	🔴	🟢	🟢	🟢	🔴	🟢	🔴	🟢	6
13.	Xie et al., 2019 [57]	🔴	🟢	🟡	🟡	🟡	🟢	🟡	🟢	🟢	🟢	🔴	🔴	🟢	6
14.	Wang et al., 2019 [60]	🔴	🟢	🔴	🟢	🟢	🟢	🟡	🟢	🟢	🟢	🔴	🔴	🟢	8
15.	Chen et al., 2019 [30]	🔴	🟡	🔴	🔴	🔴	🔴	🟡	🔴	🔴	🔴	🔴	🔴	🟢	1
16.	Yan et al., 2019 [32]	🔴	🔴	🔴	🔴	🔴	🔴	🟢	🔴	🔴	🔴	🔴	🔴	🟢	2
17.	Yang et al., 2019 [33]	🔴	🟡	🔴	🟡	🟡	🔴	🔴	🔴	🔴	🟡	🔴	🔴	🟢	1
18.	Jin et al., 2019 [48]	🔴	🟡	🔴	🟡	🔴	🔴	🔴	🟡	🟢	🟡	🔴	🔴	🟡	1
19.	Foley et al., 2018 [34]	🔴	🟡	🟡	🟡	🟡	🔴	🟢	🟡	🟢	🔴	🔴	🔴	🟢	3
20.	Larue et al., 2018 [43]	🔴	🟢	🟢	🟡	🟡	🔴	🔴	🟢	🟢	🟢	🔴	🔴	🟡	5
21.	Beukinga et al., 2018 [49]	🔴	🟡	🟢	🟡	🟢	🔴	🔴	🟡	🟢	🟢	🟢	🔴	🟡	5
22.	Riyahi et al., 2018 [52]	🔴	🟡	🔴	🟡	🟡	🔴	🔴	🔴	🟢	🟡	🔴	🔴	🟡	1
23.	Paul et al., 2017 [37]	🔴	🔴	🔴	🔴	🟡	🔴	🔴	🔴	🔴	🟢	🔴	🔴	🟡	1
24.	Desbordes et al., 2017 [50]	🔴	🔴	🟢	🟡	🟡	🟢	🔴	🟡	🟢	🟢	🔴	🔴	🟡	4
25.	Nakajo et al., 2017 [59]	🔴	🟡	🟡	🟢	🔴	🟢	🟡	🔴	🟢	🟢	🔴	🔴	🟡	4
26.	Beukinga et al., 2017 [45]	🔴	🟡	🔴	🟢	🟢	🔴	🔴	🟡	🟢	🟢	🟢	🔴	🔴	5
27.	Wakatsuki et al., 2017 [62]	🔴	🟡	🔴	🔴	🔴	🟢	🟡	🔴	🔴	🟡	🔴	🔴	🟢	2
28.	Hou et al., 2017 [53]	🔴	🟡	🟡	🟢	🟡	🔴	🟡	🔴	🔴	🟢	🔴	🔴	🟡	2
29.	Yip et al., 2016 [61]	🔴	🔴	🔴	🔴	🟡	🟢	🔴	🔴	🟡	🔴	🔴	🔴	🟡	1
30.	Rossum et al., 2016 [38]	🔴	🟢	🟢	🟡	🟢	🔴	🔴	🟡	🟢	🔴	🟢	🟢	🟢	7

Red circle: Poor rating, Yellow circle: Moderate rating, Green circle: Good rating.

Discussion

This systematic review summarized the basic characteristics and the reported results of radiomics studies predicting clinical outcomes after CCRT in EC, and assessed the methodological quality of recent studies. The included studies focused on the prediction of treatment response and side effects to neoadjuvant and definitive CCRT, and prognosis. Prediction models were constructed by using either handcrafted or deep learning-based radiomics features. Although a few methodologically robust studies have reported promising results and have demonstrated the potential to be adopted as clinical practice tools, the methodological quality of a sizable number of studies remains suboptimal. Future studies have significant room for improvement in terms of more complete reporting of essential details of the modelling work, more robust methods in construction of the model and better documentation of the final model such that independent external validation can be easily performed.

The results of this review showed that more and more researchers are investigating radiomics for prediction of nCRT response in EC. Most of these studies used pCR as an endpoint, with AUC ranging from 0.74 [45] to 0.857 [28]. However, one of the most significant shortcomings is lack of independent validation. We think that more attention should be given to testing the wider generalizability of the models through independent external validation. In addition, the difference in radiotherapy and chemotherapy regimens used in studies will also affect the probability of achieving pCR. Although some studies have combined clinical parameters with radiomics, the effect of different treatment regimens on the predictive power of the final model has not yet been investigated in detail.

Li et al. [54] demonstrated radiomics combined with clinical factors has a superior discriminative performance and a better goodness-of-fit than the clinical model. According to Van et al. [38], the addition of comprehensive PET features improves the predictive power of the model compared to using only clinical features. Based on the results of the studies included in this review, it can be concluded that the predictive power of a multidimensional predictive model is usually higher than that of a predictive model built using a single type of data.

Hu et al. [29] showed that peritumoral CT handcrafted features were less robust than the intratumoral features, and the predictive power of the model could be improved by combining peritumoral and intratumoral features. This study also included a radiogenomics analysis to explain the association of peritumoral tissue with pCR from the perspective of immune microenvironment. This result gives us an indication that the definition of ROI should be further explored. Furthermore, Hu et al. [64] conducted a deep learning study that used the same cohort of data to extract features by using six CNN models with AUCs in the range of 0.635-0.805, which demonstrated that deep learning-based radiomics also have the ability to predict the response to nCRT.

Three other studies defined endpoints as greater than 30% reduction of tumor [48], Mandard grades 1-3 [62], and downstaging [61] and obtained moderate predictive efficacy (AUC range was 0.689-0.78). We can see that a radiomics-based model can screen out not only the patients who are very sensitive to nCRT, which refers to those who can achieve pCR, but also the patients who have partial remission.

In countries such as China and Japan, clinical guidelines recommend concurrent chemoradiotherapy as the standard of care, but fewer patients in these countries receive this type

of treatment in clinical practice compared to Western countries. The reason for this may be related to the different tolerances and responses to side effects in different ethnic groups [69]. However, it might also be related to genetics, since a number of studies [70-72] revealed a correlation between gene single nucleotide polymorphism and the intrinsic radiosensitivity of lung to radiation. Therefore, if rare side effects associated with concurrent chemoradiotherapy of the esophagus can be accurately predicted, it may be additionally helpful to improve the treatment outcome and the quality of patient survival, as well as to assist in clinical decision making.

Accurately predicting patient prognosis is still a challenging task, and some studies have used radiomics for predicting endpoints such as OS, PFS, and DFS, but the results vary widely, with C-index/AUC ranging from 0.57 [60] to 0.822 [50]. These studies used retrospective data, and one of the most fundamental problems is that the accuracy of follow-up with prognosis as an endpoint cannot always be obtained. In general, the current studies for prognostic prediction are pilot investigations, and adding more dimensions such as clinical parameters and genetic information can improve the predictive power of model.

With our 13-point methodological assessment criteria, we must emphasise that we are not proposing that some models are intrinsically "better" or "worse". The primary purpose of the critical appraisal was to understand which of these reported model results have a high likelihood of being successful reproduced independently elsewhere, and thus have higher change of wide clinical generalizability. Both reproducibility and generalizability are essential aspects of our estimation of methodological robustness.

It would have been ideal if data collection and a statistical analysis protocol of radiomics modelling studies could have been prospectively registered, but there is presently no widely held consensus on where the such protocols or modelling studies might be registered in advance. We recommend that biomedical modelling registries (e.g., AIMe registry [73]) should be given more attention by the radiomics community, so that there exists an opportunity for collaboration, review and advice for improvement prior to commencing a radiomics study.

The reviewed studies paid attention to imaging settings, ROI definition, discrimination metrics and comparison of radiomics with non-radiomics predictors, however relatively few studies gave the same degree of attentiveness to : (i) documenting image pre-processing steps if any were used, (ii) clearly defining and justifying the clinical relevance of risk groupings, (iii) testing model calibration and (iv) estimating the clinical impact of the model, for example by decision curve analysis. We recommend that additional attention be paid to the aforementioned aspects by future researchers and journal editors.

Independent validation remains one of the key areas in which future radiomics modelling studies in EC could be significantly improved; our review found that the vast majority (27/30 studies) comprised solely of single-institutional datasets. Reporting of selected features in the final model together with regression coefficients would aid reproducibility testing of such models. In cases where a regression model has not been used, we recommend that models should be made openly accessible to download, or an online calculator of risk scores should be provided, to allow other researchers to independently externally validate using new datasets.

Adoption of standards and guidelines are expected to have an overall positive effect on

widespread generalizability and external validity. If an option for prospective image collection for radiomics study exists, we recommend fully standardized image acquisition and reconstruction guidelines such as the EANM Research Limited (EARL) [74], but we also acknowledge that (for the present time) the vast majority of images available for radiomics study consists of retrospectively extracted data from routine care procedures. In addition to standardizing radiomics feature definitions, the Imaging Biomarker Standardization Initiative (IBSI) [75] advises reporting of patient handling, image acquisition, image pre-processing, feature extraction, and model building, hence we also recommend this when reporting on radiomics analyses.

Studies reviewed were consistent such that the event rate was low compared to the number of possible model parameters considered (before feature selection/dimensionality reduction). This was especially true for models with treatment side-effects as the primary outcome. Increasing the sample size and synthetically enhancing data diversity are two intuitive approaches that may be considered in future. A growing number of domain generalization techniques are emerging from the deep learning field, such as domain adaptation [76] and meta-learning [77] that could assist the latter approach. However, the more immediate solution remains the former, and an option may be to make multi-institutional data publicly accessible in a centralized repository such as The Cancer Imaging Archive (TCIA). Alternatively, privacy-preserving federated learning [78] (also known as distributed learning) may be a feasible solution that for modelling on private data between institutions without physically exchanging individual patient data. Federated learning has been shown to be feasible in the radiomics domain [79, 80], and also for EC in particular [81].

Based on a small number of methodologically robust studies, we estimated the state of the art prognostic performance for radiomics models in EC to be in the ballpark of 0.85. There was indeed a correlation between our methodological assessment items with TRIPOD type of study, which is in agreement with a systematic review in lung cancer [25]. While we noted no studies published with a discriminative index below 0.60, we cannot at the present moment conclude whether or not this is a sign of publication bias; to effectively do this, we would need a prospective registry of modelling studies, as mentioned previously. This has been the widely adopted standard for epidemiological clinical studies (such as randomized controlled trials) as a means of incentivizing research transparency and detecting the presence of publication bias. Hence, we re-iterate our recommendation that the community should come to a consensus about a prospective registry for biomedical modelling studies.

Only a small number of studies at the present time addressed deep learning-based radiomics, however we would expect this number to grow rapidly in future. Different studies suggest that discriminative performance of deep learning models are superior to models based only on handcrafted features, however it remains difficult to interpret the significance of deep learning features when applied to a specific clinical case. Explainable and interpretable deep learning is presently an active area of technical development, and we have seen some use of “attention mapping” (e.g., Grad-CAM [82]) to indicate which region of the image appears to influence the discrimination strongly. Additionally research is also required to determine the relationship between image-based features and biological processes that may underpin the observed clinical outcomes.

We may note a number of limitations of the current systematic review that could potentially be addressed in some future work. First, we were not able to perform a quantative

meta-analysis due to the high heterogeneity of the mathematical procedures, even among related types of clinical outcome. Instead, we attempted a visual synthesis of reported model performance versus methodological robustness and TRIPOD study design (see Figure 2). Secondly, we may have been able to detect more studies by searching in grey literature for non-peer reviewed work, however we did not expect studies of high methodological quality to appear from those sources. On the other hand, it may have been possible to detect works where the model discriminative performance was between 0.5 to 0.6, whereas anything below 0.6 appears to be absent in our eligible articles. Thirdly, while we made our best possible attempt at evaluating methodological procedure with an objective criteria, independent raters and then combined consensus, some residual amount of subjectivity and debatable result of assessment may still persist; we have provided additional detailed notes in the supplementary material regarding methodology and tried to make our evaluations as transparent as possible. Lastly, we introduced some inclusion bias by only allowing full-text articles in the English language. This was done for the purely pragmatic reason that all authors of this review understood English, and that such selected material will be accessible / understandable to readers of the present review, should they wish to inspect the individual papers by themselves.

Conclusions

We summarized the available studies applying radiomics in predicting clinical outcomes of esophageal cancer patients who received concurrent chemoradiotherapy. Furthermore, the methodological quality of the included studies were analyzed to further improve the predictive power of radiomics and unlock the process of translation to clinical applications. Due to the limitations of inappropriate methodologies, incomplete and unclear reporting of information in radiomics model development and validation phases, the clinical application of radiomics has been impeded. The current systematic review pointed out these issues and provided our recommendations to increase generalization, biological interpretation, and clinical utility of a radiomics model.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: a cancer journal for clinicians. 2021;71:209-49. doi:10.3322/caac.21660.

2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA: a cancer journal for clinicians. 2020;70:7-30. doi:10.3322/caac.21590.

3. Moaven O, Wang TN. Combined Modality Therapy for Management of Esophageal Cancer: Current Approach Based on Experiences from East and West. Surg Clin North Am. 2019;99:479-99. doi:10.1016/j.suc.2019.02.004.

4. van Hagen P, Hulshof MCCM, van Lanschot JJB, Steyerberg EW, van Berge Henegouwen MI, Wijnhoven BPL, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. N Engl J Med. 2012;366:2074-84. doi:10.1056/NEJMoa1112088.

5. Watanabe M, Otake R, Kozuki R, Toihata T, Takahashi K, Okamura A, et al. Recent progress in multidisciplinary treatment for patients with esophageal cancer. Surg Today. 2020;50:12-20. doi:10.1007/s00595-019-01878-7.

6. Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, et al. Introduction to Radiomics. Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine. 2020;61:488-95. doi:10.2967/jnumed.118.222893.

7. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. European Journal of Cancer (Oxford, England: 1990). 2012;48:441-6. doi:10.1016/j.ejca.2011.11.036.

8. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magnetic Resonance Imaging. 2012;30:1234-48. doi:10.1016/j.mri.2012.06.010.

9. Wang H, Wang L, Lee EH, Zheng J, Zhang W, Halabi S, et al. Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures. Eur J Nucl Med Mol Imaging. 2021;48:1478-86. doi:10.1007/s00259-020-05075-4.

10. Park HJ, Park B, Lee SS. Radiomics and Deep Learning: Hepatic Applications. Korean J Radiol. 2020;21:387-401. doi:10.3348/kjr.2019.0752.

11. Avanzo M, Stancanello J, Pirrone G, Sartor G. Radiomics and deep learning in lung cancer. Strahlenther Onkol. 2020;196:879-87. doi:10.1007/s00066-020-01625-9.

12. Bibault J-E, Giraud P, Housset M, Durdux C, Taieb J, Berger A, et al. Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. Sci Rep. 2018;8:12611. doi:10.1038/s41598-018-30657-6.

13. Xie C-Y, Pang C-L, Chan B, Wong EY-Y, Dou Q, Vardhanabhuti V. Machine Learning and Radiomics Applications in Esophageal Cancers Using Non-Invasive Imag-

ing Methods-A Critical Review of Literature. Cancers (Basel). 2021;13. doi:10.3390/cancers13102469.

14. Sah B-R, Owczarczyk K, Siddique M, Cook GJR, Goh V. Radiomics in esophageal and gastric cancer. Abdominal Radiology (New York). 2019;44:2048-58. doi:10.1007/s00261-018-1724-8.

15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.

16. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med. 2013;158:280-6. doi:10.7326/0003-4819-158-4-201302190-00009.

17. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155:529-36. doi:10.7326/0003-4819-155-8-201110180-00009.

18. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170:51-8. doi:10.7326/M18-1376.

19. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14:749-62. doi:10.1038/nrclinonc.2017.141.

20. Zhong J, Hu Y, Si L, Jia G, Xing Y, Zhang H, et al. A systematic review of radiomics in osteosarcoma: utilizing radiomics quality score as a tool promoting clinical translation. Eur Radiol. 2021;31:1526-35. doi:10.1007/s00330-020-07221-w.

21. Park JE, Kim D, Kim HS, Park SY, Kim JY, Cho SJ, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol. 2020;30:523-36. doi:10.1007/s00330-019-06360-z.

22. Spadarella G, Calareso G, Garanzini E, Ugga L, Cuocolo A, Cuocolo R. MRI based radiomics in nasopharyngeal cancer: Systematic review and perspectives using radiomic quality score (RQS) assessment. Eur J Radiol. 2021;140:109744. doi:10.1016/j.ejrad.2021.109744.

23. Wang H, Zhou Y, Li L, Hou W, Ma X, Tian R. Current status and quality of radiomics studies in lymphoma: a systematic review. Eur Radiol. 2020;30:6228-40. doi:10.1007/s00330-020-06927-1.

24. Sanduleanu S, Woodruff HC, de Jong EEC, van Timmeren JE, Jochems A, Dubois L, et al. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2018;127:349-60. doi:10.1016/j.radonc.2018.03.033.

25. Fornacon-Wood I, Faivre-Finn C, O'Connor JPB, Price GJ. Radiomics as a per-

sonalized medicine tool in lung cancer: Separating the hope from the hype. Lung Cancer. 2020;146:197-208. doi:10.1016/j.lungcan.2020.05.028.

26. Zhang C, de A. F. Fonseca L, Shi Z, Zhu C, Dekker A, Bermejo I, et al. Systematic review of radiomic biomarkers for predicting immune checkpoint inhibitor treatment outcomes. Methods. 2021;188:61-72. doi:10.1016/j.ymeth.2020.11.005.

27. A T, L W, A D, R G. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. International journal of radiation oncology, biology, physics. 2018;102. doi:10.1016/j.ijrobp.2018.05.053.

28. Beukinga RJ, Wang D, Karrenbeld A, Dijksterhuis WPM, Faber H, Burgerhof JGM, et al. Addition of HER2 and CD44 to 18F-FDG PET-based clinico-radiomic models enhances prediction of neoadjuvant chemoradiotherapy response in esophageal cancer. Eur Radiol. 2021;31:3306-14. doi:10.1007/s00330-020-07439-8.

29. Hu Y, Xie C, Yang H, Ho JWK, Wen J, Han L, et al. Assessment of Intratumoral and Peritumoral Computed Tomography Radiomics for Predicting Pathological Complete Response to Neoadjuvant Chemoradiation in Patients With Esophageal Squamous Cell Carcinoma. JAMA Netw Open. 2020;3:e2015927. doi:10.1001/jamanetworkopen.2020.15927.

30. Chen Y-H, Lue K-H, Chu S-C, Chang B-S, Wang L-Y, Liu D-W, et al. Combining the radiomic features and traditional parameters of 18F-FDG PET with clinical profiles to improve prognostic stratification in patients with esophageal squamous cell carcinoma treated with neoadjuvant chemoradiotherapy and surgery. Ann Nucl Med. 2019;33:657-70. doi:10.1007/s12149-019-01380-7.

31. Wang L, Gao Z, Li C, Sun L, Li J, Yu J, et al. Computed tomography-based delta-radiomics analysis for discriminating radiation pneumonitis in patients with esophageal cancer after radiation therapy. International Journal of Radiation Oncology, Biology, Physics. 2021. doi:10.1016/j.ijrobp.2021.04.047.

32. Yan Z, Zhang J, Long H, Sun X, Li D, Tang T, et al. Correlation of CT texture changes with treatment response during radiation therapy for esophageal cancer: An exploratory study. PLoS ONE. 2019;14:e0223140. doi:10.1371/journal.pone.0223140.

33. Yang Z, He B, Zhuang X, Gao X, Wang D, Li M, et al. CT-based radiomic signatures for prediction of pathologic complete response in esophageal squamous cell carcinoma after neoadjuvant chemoradiotherapy. Journal of Radiation Research. 2019;60:538-45. doi:10.1093/jrr/rrz027.

34. Foley KG, Hills RK, Berthon B, Marshall C, Parkinson C, Lewis WG, et al. Development and validation of a prognostic model incorporating texture analysis derived from standardised segmentation of PET in patients with oesophageal cancer. Eur Radiol. 2018;28:428-36. doi:10.1007/s00330-017-4973-y.

35. Foley KG, Shi Z, Whybra P, Kalendralis P, Larue R, Berbee M, et al. External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer. Radiotherapy and Oncology. 2019;133:205-12. doi:10.1016/j.radonc.2018.10.033.

36. Li Y, Beck M, Päßler T, Lili C, Hua W, Mai HD, et al. A FDG-PET radiomics

signature detects esophageal squamous cell carcinoma patients who do not benefit from chemoradiation. Sci Rep. 2020;10:17671. doi:10.1038/s41598-020-74701-w.

37. Paul D, Su R, Romain M, Sébastien V, Pierre V, Isabelle G. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. Computerized Medical Imaging and Graphics. 2017;60:42-9. doi:10.1016/j.compmedimag.2016.12.002.

38. van Rossum PSN, Fried DV, Zhang L, Hofstetter WL, van Vulpen M, Meijer GJ, et al. The Incremental Value of Subjective and Quantitative Assessment of 18F-FDG PET for the Prediction of Pathologic Complete Response to Preoperative Chemoradiotherapy in Esophageal Cancer. Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine. 2016;57:691-700. doi:10.2967/jnumed.115.163766.

39. Tixier F, Le Rest CC, Hatt M, Albarghach N, Pradier O, Metges JP, et al. Intratumor Heterogeneity Characterized by Textural Features on Baseline 18F-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer. Journal of Nuclear Medicine. 2011;52:369-78. doi:10.2967/jnumed.110.082404.

40. Zhang H, Tan S, Chen W, Kligerman S, Kim G, D'Souza WD, et al. Modeling Pathologic Response of Esophageal Cancer to Chemoradiation Therapy Using Spatial-Temporal 18F-FDG PET Features, Clinical Parameters, and Demographics. International Journal of Radiation Oncology*Biology*Physics. 2014;88:195-203. doi:10.1016/j.ijrobp.2013.09.037.

41. Luo H-S, Huang S-F, Xu H-Y, Li X-Y, Wu S-X, Wu D-H. A nomogram based on pretreatment CT radiomics features for predicting complete response to chemoradiotherapy in patients with esophageal squamous cell cancer. Radiation Oncology. 2020;15:249. doi:10.1186/s13014-020-01692-3.

42. Du F, Tang N, Cui Y, Wang W, Zhang Y, Li Z, et al. A Novel Nomogram Model Based on Cone-Beam CT Radiomics Analysis Technology for Predicting Radiation Pneumonitis in Esophageal Cancer Patients Undergoing Radiotherapy. Front Oncol. 2020;10:596013. doi:10.3389/fonc.2020.596013.

43. Larue RTHM, Klaassen R, Jochems A, Leijenaar RTH, Hulshof MCCM, van Berge Henegouwen MI, et al. Pre-treatment CT radiomics to predict 3-year overall survival following chemoradiotherapy of esophageal cancer. Acta Oncologica. 2018;57:1475-81. doi :10.1080/0284186X.2018.1486039.

44. Tan S, Zhang H, Zhang Y, Chen W, D'Souza WD, Lu W. Predicting pathologic tumor response to chemoradiotherapy with histogram distances characterizing longitudinal changes in F-FDG uptake patterns: Predicting pathologic tumor response with F-FDG histogram distances. Med Phys. 2013;40:101707. doi:10.1118/1.4820445.

45. Beukinga RJ, Hulshoff JB, van Dijk LV, Muijs CT, Burgerhof JGM, Kats-Ugurlu G, et al. Predicting Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer with Textural Features Derived from Pretreatment F-FDG PET/CT Imaging. Journal of Nuclear Medicine. 2017;58:723-9. doi:10.2967/jnumed.116.180299.

46. Ypsilantis P-P, Siddique M, Sohn H-M, Davies A, Cook G, Goh V, et al. Predicting

Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. PLoS ONE. 2015;10:e0137036. doi:10.1371/journal.pone.0137036.

47. Zhang C, Shi Z, Kalendralis P, Whybra P, Parkinson C, Berbee M, et al. Prediction of lymph node metastases using pre-treatment PET radiomics of the primary tumour in esophageal adenocarcinoma: an external validation study. 2020.

48. Jin X, Zheng X, Chen D, Jin J, Zhu G, Deng X, et al. Prediction of response after chemoradiation for esophageal cancer using a combination of dosimetry and CT radiomics. Eur Radiol. 2019;29:6080-8. doi:10.1007/s00330-019-06193-w.

49. Beukinga RJ, Hulshoff JB, Mul VEM, Noordzij W, Kats-Ugurlu G, Slart RHJA, et al. Prediction of Response to Neoadjuvant Chemotherapy and Radiation Therapy with Baseline and Restaging ¹⁸ F-FDG PET Imaging Biomarkers in Patients with Esophageal Cancer. Radiology. 2018;287:983-92. doi:10.1148/radiol.2018172229.

50. Desbordes P, Ruan S, Modzelewski R, Pineau P, Vauclin S, Gouel P, et al. Predictive value of initial FDG-PET features for treatment response and survival in esophageal cancer patients treated with chemo-radiation therapy using a random forest classifier. PLoS ONE. 2017;12:e0173208. doi:10.1371/journal.pone.0173208.

51. Yip C, Landau D, Kozarski R, Ganeshan B, Thomas R, Michaelidou A, et al. Primary Esophageal Cancer: Heterogeneity as Potential Prognostic Biomarker in Patients Treated with Definitive Chemotherapy and Radiation Therapy. Radiology. 2014;270:141-8. doi:10.1148/radiol.13122869.

52. Riyahi S, Choi W, Liu C-J, Zhong H, Wu AJ, Mechalakos JG, et al. Quantifying local tumor morphological changes with Jacobian map for prediction of pathologic tumor response to chemo-radiotherapy in locally advanced esophageal cancer. Phys Med Biol. 2018;63:145020. doi:10.1088/1361-6560/aacd22.

53. Hou Z, Ren W, Li S, Liu J, Sun Y, Yan J, et al. Radiomic analysis in contrast-enhanced CT: predict treatment response to chemoradiotherapy in esophageal carcinoma. Oncotarget. 2017;8:104444-54. doi:10.18632/oncotarget.22304.

54. Li Y, Liu J, Li H-X, Cai X-W, Li Z-G, Ye X-D, et al. Radiomics Signature Facilitates Organ-Saving Strategy in Patients With Esophageal Squamous Cell Cancer Receiving Neoadjuvant Chemoradiotherapy. Front Oncol. 2020;10:615167. doi:10.3389/fonc.2020.615167.

55. Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. Eur J Nucl Med Mol Imaging. 2013;40:1662-71. doi:10.1007/s00259-013-2486-8.

56. Tan S, Kligerman S, Chen W, Lu M, Kim G, Feigenberg S, et al. Spatial-Temporal [18F]FDG-PET Features for Predicting Pathologic Response of Esophageal Cancer to Neoadjuvant Chemoradiation Therapy. International Journal of Radiation Oncology*Biology*-Physics. 2013;85:1375-82. doi:10.1016/j.ijrobp.2012.10.017.

57. Xie C, Yang P, Zhang X, Xu L, Wang X, Li X, et al. Sub-region based radiomics

analysis for survival prediction in oesophageal tumours treated by definitive concurrent chemoradiotherapy. EBioMedicine. 2019;44:289-97. doi:10.1016/j.ebiom.2019.05.023.

58. Xie Y, Wang Q, Cao B, Lv J, Wang Y, Wu L, et al. Textural features based enhanced contrast CT images predicts prognosis to concurrent chemoradiotherapy in stage III esophageal squamous cell cancer. CBM. 2020;27:325-33. doi:10.3233/CBM-190586.

59. Nakajo M, Jinguji M, Nakabeppu Y, Nakajo M, Higashi R, Fukukura Y, et al. Texture analysis of 18F-FDG PET/CT to predict tumour response and prognosis of patients with esophageal cancer treated by chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2017;44:206-14. doi:10.1007/s00259-016-3506-2.

60. Wang Q, Cao B, Chen J, Li C, Tan L, Zhang W, et al. Tumor Compactness based on CT to predict prognosis after multimodal treatment for esophageal squamous cell carcinoma. Sci Rep. 2019;9:10497. doi:10.1038/s41598-019-46899-x.

61. Yip SSF, Coroller TP, Sanford NN, Huynh E, Mamon H, Aerts HJWL, et al. Use of registration-based contour propagation in texture analysis for esophageal cancer pathologic response prediction. Phys Med Biol. 2016;61:906-22. doi:10.1088/0031-9155/61/2/906.

62. Wakatsuki K, Matsumoto S, Migita K, Ito M, Kunishige T, Nakade H, et al. Usefulness of computed tomography density of a tumor in predicting the response of advanced esophageal cancer to preoperative chemotherapy. Surgery. 2017;162:823-35. doi:10.1016/j.surg.2017.06.003.

63. Xie C-Y, Hu Y-H, Ho JW-K, Han L-J, Yang H, Wen J, et al. Using Genomics Feature Selection Method in Radiomics Pipeline Improves Prognostication Performance in Locally Advanced Esophageal Squamous Cell Carcinoma-A Pilot Study. Cancers (Basel). 2021;13. doi:10.3390/cancers13092145.

64. Hu Y, Xie C, Yang H, Ho JWK, Wen J, Han L, et al. Computed tomography-based deep-learning prediction of neoadjuvant chemoradiotherapy treatment response in esophageal squamous cell carcinoma. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;154:6-13. doi:10.1016/j.radonc.2020.09.014.

65. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS medicine. 2009;6:e1000097. doi:10.1371/journal.pmed.1000097.

66. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magnetic Resonance Imaging. 2012;30:1323-41. doi:10.1016/j.mri.2012.05.001.

67. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77:e104-e7. doi:10.1158/0008-5472.CAN-17-0339.

68. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015;162:W1-

W73. doi:10.7326/M14-0698.

69. Faehling M, Schulz C, Laack H, Wolff T, Rückert A, Reck M, et al. PACIFIC subgroup analysis: pneumonitis in stage III, unresectable NSCLC patients treated with durvalumab vs. placebo after CRT. Pneumologie. 2019;73:P272.

70. Wen J, Liu H, Wang Q, Liu Z, Li Y, Xiong H, et al. Genetic variants of the LIN28B gene predict severe radiation pneumonitis in patients with non-small cell lung cancer treated with definitive radiation therapy. European Journal of Cancer (Oxford, England: 1990). 2014;50:1706-16. doi:10.1016/j.ejca.2014.03.008.

71. Pu X, Wang L, Chang JY, Hildebrandt MaT, Ye Y, Lu C, et al. Inflammation-related genetic variants predict toxicity following definitive radiotherapy for lung cancer. Clin Pharmacol Ther. 2014;96:609-15. doi:10.1038/clpt.2014.154.

72. Pang Q, Wei Q, Xu T, Yuan X, Lopez Guerra JL, Levy LB, et al. Functional promoter variant rs2868371 of HSPB1 is associated with risk of radiation pneumonitis after chemoradiation for non-small cell lung cancer. International Journal of Radiation Oncology, Biology, Physics. 2013;85:1332-9. doi:10.1016/j.ijrobp.2012.10.011.

73. Matschinske J, Alcaraz N, Benis A, Golebiewski M, Grimm DG, Heumos L, et al. The AIMe registry for artificial intelligence in biomedical research. Nat Methods. 2021;18:1128-31. doi:10.1038/s41592-021-01241-0.

74. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging. 2015;42:328-54. doi:10.1007/s00259-014-2961-x.

75. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295:328-38. doi:10.1148/radiol.2020191145.

76. Balaji Y, Sankaranarayanan S, Chellappa R. MetaReg: Towards Domain Generalization using Meta-Regularization. NeurIPS; 2018.

77. Dou Q, Castro DC, Kamnitsas K, Glocker B. Domain Generalization via Model-Agnostic Learning of Semantic Features. arXiv:191013580 [cs]. 2019.

78. McMahan B, Moore E, Ramage D, Hampson S, Arcas BAy. Communication-Efficient Learning of Deep Networks from Decentralized Data. Artificial Intelligence and Statistics: PMLR; 2017. p. 1273-82.

79. Shi Z, Zhovannik I, Traverso A, Dankers FJWM, Deist TM, Kalendralis P, et al. Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. Sci Data. 2019;6:218. doi:10.1038/s41597-019-0241-0.

80. Bogowicz M, Jochems A, Deist TM, Tanadini-Lang S, Huang SH, Chan B, et al. Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer. Sci Rep. 2020;10:4542. doi:10.1038/s41598-020-61297-4.

81. Shi Z, Foley KG, Pablo de Mey J, Spezi E, Whybra P, Crosby T, et al. External

Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patients. Front Oncol. 2019;9:1411. doi:10.3389/fonc.2019.01411.

82. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 618-26.

Supplementary Materials

Table S1: Systematic search

Table S2: Detailed image acquisition and model characteristics

Table S3: Assessment of methodological quality of included studies

Supplementary Table 1: Systematic search

a) Medline

(((“Validat”[All Fields] OR ((“predict”[All Fields] OR “predictabilities”[All Fields] OR “predictability”[All Fields] OR “predictable”[All Fields] OR “predictably”[All Fields] OR “predicted”[All Fields] OR “predicting”[All Fields] OR “prediction”[All Fields] OR “predictions”[All Fields] OR “predictive”[All Fields] OR “predictively”[All Fields] OR “predictiveness”[All Fields] OR “predictives”[All Fields] OR “predictivities”[All Fields] OR “predictivity”[All Fields] OR “predicts”[All Fields]) AND “ti”[All Fields]) OR “Rule”[All Fields] OR ((“predict”[All Fields] OR “predictabilities”[All Fields] OR “predictability”[All Fields] OR “predictable”[All Fields] OR “predictably”[All Fields] OR “predicted”[All Fields] OR “predicting”[All Fields] OR “prediction”[All Fields] OR “predictions”[All Fields] OR “predictive”[All Fields] OR “predictively”[All Fields] OR “predictiveness”[All Fields] OR “predictives”[All Fields] OR “predictivities”[All Fields] OR “predictivity”[All Fields] OR “predicts”[All Fields]) AND (“outcome”[All Fields] OR “outcomes”[All Fields] OR (“risk”[MeSH Terms] OR “risk”[All Fields]) OR (“model”[All Fields] OR “model s”[All Fields] OR “modeled”[All Fields] OR “modeler”[All Fields] OR “modeler s”[All Fields] OR “modelers”[All Fields] OR “modeling”[All Fields] OR “modelings”[All Fields] OR “modelization”[All Fields] OR “modelizations”[All Fields] OR “modelize”[All Fields] OR “modelized”[All Fields] OR “modelled”[All Fields] OR “modeller”[All Fields] OR “modellers”[All Fields] OR “modelling”[All Fields] OR “modellings”[All Fields] OR “models”[All Fields]))) OR ((“history”[MeSH Terms] OR “history”[All Fields] OR “histories”[All Fields] OR “history”[MeSH Subheading] OR (“variabilities”[All Fields] OR “variability”[All Fields] OR “variable”[All Fields] OR “variable s”[All Fields] OR “variables”[All Fields] OR “variably”[All Fields]) OR (“criteria s”[All Fields] OR “criterias”[All Fields] OR “standards”[MeSH Subheading] OR “standards”[All Fields] OR “criteria”[All Fields]) OR “Scor”[All Fields] OR (“characteristic”[All Fields] OR “characteristics”[All Fields]) OR (“diagnosis”[MeSH Subheading] OR “diagnosis”[All Fields] OR “findings”[All Fields] OR “diagnosis”[MeSH Terms] OR “finds”[All Fields] OR “signs and symptoms”[MeSH Terms] OR (“signs”[All Fields] AND “symptoms”[All Fields]) OR “signs and symptoms”[All Fields] OR “finding”[All Fields]) OR (“factor”[All Fields] OR “factor s”[All Fields] OR “factors”[All Fields])) AND (“predict”[All Fields] OR “predictabilities”[All Fields] OR “predictability”[All Fields] OR “predictable”[All Fields] OR “predictably”[All Fields] OR “predicted”[All Fields] OR “predicting”[All Fields] OR “prediction”[All Fields] OR “predictions”[All Fields] OR “predictive”[All Fields] OR “predictively”[All Fields] OR “predictiveness”[All Fields] OR “predictives”[All Fields] OR “predictivities”[All Fields] OR “predictivity”[All Fields] OR “predicts”[All Fields] OR (“model”[All Fields] OR “model s”[All Fields] OR “modeled”[All Fields] OR “model-

er”[All Fields] OR “modeler s”[All Fields] OR “modelers”[All Fields] OR “modeling”[All Fields] OR “modelings”[All Fields] OR “modelization”[All Fields] OR “modelizations”[All Fields] OR “modelize”[All Fields] OR “modelized”[All Fields] OR “modelled”[All Fields] OR “modeller”[All Fields] OR “modellers”[All Fields] OR “modelling”[All Fields] OR “modellings”[All Fields] OR “models”[All Fields]) OR (“decision”[All Fields] OR “decision s”[All Fields] OR “decisions”[All Fields] OR “decisive”[All Fields] OR “decisively”[All Fields]) OR “Identif”[All Fields] OR “Prognos”[All Fields])) OR ((“decision”[All Fields] OR “decision s”[All Fields] OR “decisions”[All Fields] OR “decisive”[All Fields] OR “decisively”[All Fields]) AND (“model”[All Fields] OR “model s”[All Fields] OR “modeled”[All Fields] OR “modeler”[All Fields] OR “modeler s”[All Fields] OR “modelers”[All Fields] OR “modeling”[All Fields] OR “modelings”[All Fields] OR “modelization”[All Fields] OR “modelizations”[All Fields] OR “modelize”[All Fields] OR “modelized”[All Fields] OR “modelled”[All Fields] OR “modeller”[All Fields] OR “modellers”[All Fields] OR “modelling”[All Fields] OR “modellings”[All Fields] OR “models”[All Fields] OR (“ambulatory care facilities”[MeSH Terms] OR (“ambulatory”[All Fields] AND “care”[All Fields] AND “facilities”[All Fields]) OR “ambulatory care facilities”[All Fields] OR “clinic”[All Fields] OR “clinic s”[All Fields] OR “clinical”[All Fields] OR “clinically”[All Fields] OR “clinicals”[All Fields] OR “clinics”[All Fields]) OR (“logistic models”[MeSH Terms] OR (“logistic”[All Fields] AND “models”[All Fields]) OR “logistic models”[All Fields]))) OR ((“prognostic”[All Fields] OR “prognostical”[All Fields] OR “prognostically”[All Fields] OR “prognosticate”[All Fields] OR “prognosticated”[All Fields] OR “prognosticates”[All Fields] OR “prognosticating”[All Fields] OR “prognostication”[All Fields] OR “prognostications”[All Fields] OR “prognosticator”[All Fields] OR “prognosticators”[All Fields] OR “prognostics”[All Fields]) AND (“history”[MeSH Terms] OR “history”[All Fields] OR “histories”[All Fields] OR “history”[MeSH Subheading] OR (“variabilities”[All Fields] OR “variability”[All Fields] OR “variable”[All Fields] OR “variable s”[All Fields] OR “variables”[All Fields] OR “variably”[All Fields]) OR (“criteria s”[All Fields] OR “criterias”[All Fields] OR “standards”[MeSH Subheading] OR “standards”[All Fields] OR “criteria”[All Fields]) OR “Scor”[All Fields] OR (“characteristic”[All Fields] OR “characteristics”[All Fields]) OR (“diagnosis”[MeSH Subheading] OR “diagnosis”[All Fields] OR “findings”[All Fields] OR “diagnosis”[MeSH Terms] OR “finds”[All Fields] OR “signs and symptoms”[MeSH Terms] OR (“signs”[All Fields] AND “symptoms”[All Fields]) OR “signs and symptoms”[All Fields] OR “finding”[All Fields]) OR (“factor”[All Fields] OR “factor s”[All Fields] OR “factors”[All Fields]) OR (“model”[All Fields] OR “model s”[All Fields] OR “modeled”[All Fields] OR “modeler”[All Fields] OR “modeler s”[All Fields] OR “modelers”[All Fields] OR “modeling”[All Fields] OR “modelings”[All Fields] OR “modelization”[All Fields] OR “modelizations”[All Fields] OR “modelize”[All Fields] OR “modelized”[All Fields] OR “modelled”[All Fields] OR “modeller”[All Fields] OR “modellers”[All Fields] OR “modelling”[All Fields] OR “modellings”[All Fields] OR “models”[All Fields]))) OR (“stratification”[All Fields] OR “stratifications”[All Fields] OR (“roc curve”[MeSH Terms] OR (“roc”[All Fields] AND “curve”[All Fields]) OR “roc curve”[All Fields]) OR (“discriminabilities”[All Fields] OR “discriminability”[All Fields] OR “discriminable”[All Fields] OR “discriminably”[All Fields] OR “discriminance”[All Fields] OR “discriminant”[All Fields] OR “discriminants”[All Fields] OR “discriminate”[All Fields] OR “discriminated”[All Fields] OR “discriminates”[All Fields] OR “discriminating”[All Fields] OR “discrimination, psychological”[MeSH Terms] OR (“discrimination”[All Fields] AND “psychological”[All Fields]) OR “psychological discrimination”[All Fields] OR “discrimination”[All Fields] OR “discrimi-

nations"[All Fields] OR "discriminative"[All Fields] OR "discriminatively"[All Fields] OR "discriminator"[All Fields] OR "discriminators"[All Fields]) OR ("discriminabilities"[All Fields] OR "discriminability"[All Fields] OR "discriminable"[All Fields] OR "discriminably"[All Fields] OR "discriminance"[All Fields] OR "discriminant"[All Fields] OR "discriminants"[All Fields] OR "discriminate"[All Fields] OR "discriminated"[All Fields] OR "discriminates"[All Fields] OR "discriminating"[All Fields] OR "discrimination, psychological"[MeSH Terms] OR ("discrimination"[All Fields] AND "psychological"[All Fields]) OR "psychological discrimination"[All Fields] OR "discrimination"[All Fields] OR "discriminations"[All Fields] OR "discriminative"[All Fields] OR "discriminatively"[All Fields] OR "discriminator"[All Fields] OR "discriminators"[All Fields]) OR ("c basel"[Journal] AND ("statistic"[All Fields] OR "statistic s"[All Fields] OR "statistical"[All Fields] OR "statistically"[All Fields] OR "statistics"[All Fields])) OR ("c basel"[Journal] AND ("statistic"[All Fields] OR "statistic s"[All Fields] OR "statistical"[All Fields] OR "statistically"[All Fields] OR "statistics"[All Fields])) OR ("area under curve"[MeSH Terms] OR ("area"[All Fields] AND "under"[All Fields] AND "curve"[All Fields]) OR "area under curve"[All Fields] OR ("area"[All Fields] AND "under"[All Fields] AND "curve"[All Fields]) OR "area under the curve"[All Fields]) OR "auc"[All Fields] OR ("calibrant"[All Fields] OR "calibrants"[All Fields] OR "calibrate"[All Fields] OR "calibrated"[All Fields] OR "calibrates"[All Fields] OR "calibrating"[All Fields] OR "calibration"[MeSH Terms] OR "calibration"[All Fields] OR "calibrations"[All Fields] OR "calibrator"[All Fields] OR "calibrators"[All Fields]) OR ("indicate"[All Fields] OR "indicated"[All Fields] OR "indicates"[All Fields] OR "indicating"[All Fields] OR "indicative"[All Fields] OR "indicatives"[All Fields] OR "indicators and reagents"[Pharmacological Action] OR "indicators and reagents"[MeSH Terms] OR ("indicators"[All Fields] AND "reagents"[All Fields]) OR "indicators and reagents"[All Fields] OR "indicator"[All Fields] OR "indicators"[All Fields] OR "indice"[All Fields] OR "indices"[All Fields]) OR ("algorithm s"[All Fields] OR "algorithmic"[All Fields] OR "algorithmically"[All Fields] OR "algorithmics"[All Fields] OR "algorithmization"[All Fields] OR "algorithms"[MeSH Terms] OR "algorithms"[All Fields] OR "algorithm"[All Fields]) OR ("multivariable"[All Fields] OR "multivariables"[All Fields] OR "multivariably"[All Fields] OR "multivariance"[All Fields] OR "multivariant"[All Fields] OR "multivariate"[All Fields] OR "multivariated"[All Fields] OR "multivariately"[All Fields] OR "multivariates"[All Fields] OR "multivariative"[All Fields])) OR ("predict*"[Title/Abstract] OR "predictive value of tests"[MeSH Terms] OR "scor*"[Title/Abstract] OR "observ*"[Title/Abstract] OR "observer variation"[MeSH Terms])) AND ((esophageal cancer[MeSH Terms]) OR (esophageal cancers[MeSH Terms]))) AND ((((((computed tomography[Text Word]) OR (CT[Text Word])) OR (magnetic resonance imaging[Text Word])) OR (MR*[Text Word])) OR (positron emission tomography[Text Word])) OR (PET[Text Word]))) AND ((((((((((radiomic[Text Word]) OR (radiomics[Text Word])) OR (textur*[Text Word])) OR (quantitative[Text Word])) OR (artificial intelligence[Text Word])) OR (AI[Text Word])) OR (deep learning[Text Word])) OR (shape[Text Word])) OR (feature[Text Word])) OR (features[Text Word]))
Filters: Humans, English

b) Embase

1. ('esophageal cancer'/exp AND 'computed tomography'/exp OR 'ct' OR 'magnetic resonance imaging'/exp OR 'mr*' OR 'pos-itron emission tomography' OR 'pet') AND 'radiomic' OR 'radiomics' OR 'textur*' OR 'quantitative' OR 'artificial in-telligence' OR 'ai'

OR ‘deep learning’/exp OR ‘shape’ OR ‘feature’ OR ‘features’:ta,ab,ti

2. ‘validat’ OR ((‘predict’ OR ‘predictabilities’ OR ‘predictability’ OR ‘predictable’ OR ‘predictably’ OR ‘predicted’ OR ‘pre-dicting’ OR ‘prediction’ OR ‘predictions’ OR ‘predictive’ OR ‘predictively’ OR ‘predictiveness’ OR ‘predictives’ OR ‘predictivi-ties’ OR ‘predictivity’ OR ‘predicts’) AND ‘ti’) OR ‘rule’ OR ((‘predict’:ta,ab,ti OR ‘predictabilities’:ta,ab,ti OR ‘predictability’:ta,ab,ti OR ‘predictable’:ta,ab,ti OR ‘predictably’:ta,ab,ti OR ‘predicted’:ta,ab,ti OR ‘predicting’:ta,ab,ti OR ‘prediction’:ta,ab,ti OR ‘predictions’:ta,ab,ti OR ‘predictive’:ta,ab,ti OR ‘predictively’:ta,ab,ti OR ‘predic-tiveness’:ta,ab,ti OR ‘predictives’:ta,ab,ti OR ‘predictivities’:ta,ab,ti OR ‘predictivity’:ta,ab,ti OR ‘predicts’:ta,ab,ti) AND (‘outcome’:ta,ab,ti OR ‘outcomes’:ta,ab,ti OR ‘risk’:ta,ab,ti OR ‘model’:ta,ab,ti OR ‘model s’:ta,ab,ti OR ‘modeled’:ta,ab,ti OR ‘modeler’:ta,ab,ti OR ‘modeler s’:ta,ab,ti OR ‘modelers’:ta,ab,ti OR ‘modeling’:ta,ab,ti OR ‘modelings’:ta,ab,ti OR ‘modelization’:ta,ab,ti OR ‘modelizations’:ta,ab,ti OR ‘modelize’:ta,ab,ti OR ‘modelized’:ta,ab,ti OR ‘modelled’:ta,ab,ti OR ‘modeller’:ta,ab,ti OR ‘modellers’:ta,ab,ti OR ‘modelling’:ta,ab,ti OR ‘modellings’:ta,ab,ti OR ‘models’:ta,ab,ti))

3. #1 AND #2 AND ([article]/lim OR [article in press]/lim OR [data papers]/lim) AND [english]/lim

Supplementary Table 2: Detailed image acquisition and model characteristics

Number	Reference	Image acquisition scanner and PET data acquisition time	Details of treatment modality	Detailed report performance
1.	Xie et al., 2021	Institution 1: Aquilion TSX-101A (Toshiba) or Discovery 750 HD (GE) Institution 2: Discovery VCT, GE Healthcare	All patients were treated with nCRT followed by surgery	AUC= 0.912 and 0.918; C-index=0.869 and 0.875 (nomogram 1 and 2 in the Training set) AUC=0.852 and 0.810; C-index=0.812 and 0.757 (nomogram 1 and 2 in the Internal test set) AUC= 0.769 and 0.724; C-index=0.719 and 0.668 (nomogram 1 and 2 in the External test set)
2.	Beukinga et al., 2021	Biograph mCT-64 PET/CT (Siemens) Images were acquired sixty minutes after tracer injection	All patients were treated with nCRT followed by surgery	AUC = 0.685 (Best of group 1) AUC = 0.857 (Best of group 2)
3.	Hu et al., 2021	Same as Hu et al., 2020	Same as Hu et al., 2020	Handcrafted model: AUC=0.822 (training) AUC/C-index=0.725, accuracy=67.1% (test) Deep learning-based: AUC=0.807-0.901 (training) AUC=0.635-0.805(test)
4.	Wang et al., 2021	Philips Brilliance Big Bore CT scanner	All patients treated with definitive CCRT	C-index=0.975 (0.953–0.996, 95% CI, internal validation) C-index=0.921 (0.876–0.966, 95% CI, external validation)
5.	Li et al., 2020	PET scanner: institution 1: Discovery STE (GE) Data acquisition started 67 ± 22 min (range 50–140 min) after injection of 142–548 MBq FDG institution 2: Gemini TF 16 Astonish (Philips) Data acquisition started 71 ± 9 min (range 60–86 min) after injection of 236–248 MBq FDG	(Training) Patients received definitive CCRT (Validation) 22 patients received definitive CCRT while 10 patients received preoperative CCRT	Clustering of OS: $p<0.0001$
6.	Xie et al., 2020	Contrast CT: 16–detector row CT scanner (PHILIPS)	All patients were treated with CCRT	1-year and 2-year survival: AUC=0.79
7.	Hu et al., 2020	Contrast-enhanced CT	All patients underwent nCRT followed by surgery	Intratumoral model AUC=0.881(training) AUC=0.730 (95%CI,0.609-0.850, test)

		scanner: Aquilion TSX-101A (Toshiba) Discovery VCT, GE Healthcare		Peritumoral model AUC=0.895 (training) AUC=0.734 (95%CI,0.614-0.854, test) Combined model AUC=0.906 (training) AUC=0.852 (95%CI,0.753-0.951, test)
8.	Luo et al., 2020	GE Lightspeed 64-slice spiral CT	All patients were treated with definitive CCRT	AUC=0.844 (95% CI 0.779–0.897, training) AUC=0.807 (95% CI 0.691–0.894, validation)
9.	Li et al., 2020	A variety of CT scanners	All patients were treated with nCRT followed by surgery	AUC = 0.84 (validation)
10.	Zhang et al., 2020	PET scanner: GE 690 scanner Uptake time was 90min.	(Training) 130 patients receiving either surgery alone, neoadjuvant chemotherapy, or nCRT followed by surgery (Validation) 60 patients who underwent nCRT	AUC=0.82 (95% CI 0.74–0.89, training) AUC=0.69 (95% CI 0.54–0.82, validation)
11.	Du et al., 2020	On-board imager (OBI) system mounted on the Varian Trilogy medical linear accelerator	Patients received definitive CCRT or definitive radiotherapy	AUC=0.836 (0.700–0.918, 95% CI, training) AUC=0.905 (0.799–1.000, 95% CI, validation)
12.	Foley et al., 2019	Same as Foley et al., 2018	Training and internal validation same as Foley et al., 2018 (External validation) All patients treated with nCRT	X2=1.27, df=3, p=0.74 (Kaplan-Meier)
13.	Xie et al., 2019	Institution 1: Brilliance Big Bore CT scanner (Philips) Institution 2: LightSpeed Pro 16 CT (GE)	The majority of patients received definitive CCRT, for patients with advanced ages or poor performance status, radiotherapy alone was delivered to these patients.	Results of 1-year and 2-year survival are omitted in this table 3-year survival: AUC=0.811 (95%CI, 0.670–0.952, training) AUC=0.805 (95%CI ,0.638–0.973, validation)
14.	Wang et al., 2019	Institution 1(Training set): contrast CT 16 detector row CT scanner (PHILIPS)	Most patients (237) were treated with CCRT, others (227) were treated with radiotherapy alone	OS: C-index= 0.64 (0.55–0.73,95% CI, training) C-index= 0.6073 (0.53-0.68,95% CI, validation 1) C-index= 0.58 (0.54-0.62,95% CI, validation 2) PFS: C-index=0.66(0.58–0.74,95% CI, training) C-index=0.60(0.54-0.67,95% CI, validation 1) C-index= 0.57(0.53-0.61,95% CI, validation 2)
15.	Chen et al., 2019	PET scanner: GE Discovery ST PET/CT unit	All patients underwent nCRT followed by surgery	Clustering response to nCRT: p=0.009

		PET images were obtained between 40 and 60 min after injection of 18F-FDG (400 MBq)		
16.	Yang et al., 2019	Non-enhanced CT scanner: Philips Brilliance CT Big Bore Oncology Configuration, Cleveland, OH	All patients underwent nCRT followed by surgery	Model 1(bin size=32): 0.86 (95% CI, 0.74-0.98, training) 0.79 (95% CI, 0.48-1.00, test) Model 2(bin size=64): 0.84 (95% CI, 0.72-0.95, training) 0.75 (95% CI, 0.42-1.00, test) Model 3(bin size=128): 0.84 (95% CI, 0.72-0.96, training) 0.71 (95% CI, 0.38-1.00, test)
17.	Yan et al., 2019	CT scanner: CT-on-rails (CTVision; Siemens) during daily CT-guided IGRT.	All patients were received radiotherapy, of which 2 patients were preoperative Radiotherapy, 42 patients were CCRT	RT response: coarseness P <0.0001, STD P=0.0007, entropy P=0.0003, strength P <0.0001 Survival: coarseness r=0.9572, P=0.0027, strength r = 0.9917, P=0.0001
18.	Jin et al., 2019	Contrast CT: 16-detector row (Brilliance, Phillips)	All patients were received CCRT	AUC=0.689
19.	Foley et al., 2018	PET scanner: GE 690 scanner Uptake time was 90min	Patients underwent surgery alone, neoadjuvant chemotherapy or nCRT prior to surgery, definitive CCRT, or palliative therapy.	X2 143.14, df 3, p < 0.001(Training) X2 20.621, df 3, p < 0.001(validation)
20.	Larue et al., 2018	Institution 1: General Electric LightSpeed RT16 (General Electric), Philips PQ5000 or Philips Gemini TF (Philips) Institution 2: Siemens SOMATOM Sensation Open CT or Siemens Biograph 40 PET/CT scanner	All patients treated with nCRT followed by surgery	AUC= 0.69 (95% CI 0.61–0.77, Training) AUC=0.61 (95% CI 0.47–0.75, Validation)
21.	Beukinga et al., 2018	Biograph mCT-64 PET/CT (Siemens) Images were acquired sixty minutes after tracer injection	All patients were treated with nCRT followed by surgery	AUC=0.82(Training) AUC=0.81(validation)
22.	Riyahi et al., 2018	Same as Tan et al., 2013	Same as Tan et al., 2013	Sensitivity=94.4±0.08%, Specificity=91.8±0.06%, Accuracy=94.0±0.05%, AUC=0.94±0.05

23. Paul et al., 2017	PET/CT Biograph Sensation 16 (Siemens).	All patients were treated with CCRT	Response to treatment: AUC=0.823±0.032 OS: AUC=0.750±0.108
24. Desbordes et al., 2017	PET scanner: Biograph1 Sensation 16 Hi-Rez device (Siemens) Images were acquired 60 (±10) minutes after tracer injection	All patients were treated by CCRT, 14 patients followed by surgery	Response: AUC=0.836±0.105 OS: AUC=0.822±0.059
25. Nakajo et al., 2017	Discovery 600 M PET/CT system (GE) Images were acquired sixty minutes after tracer injection	All patients were treated with CCRT	Response to treatment: AUC=0.75 PFS and OS: P <0.001
26. Beukinga et al., 2017	Biograph mCT 4-64 PET/CT, Siemens. Images were acquired sixty minutes after tracer injection. CT scanner: Somatom Sensation 16 or 64, Siemens.	All patients were treated with nCRT followed by surgery	AUC=0.78 (Training) AUC=0.74 (Validation)
27. Wakatsuki et al., 2017	Enhanced CT: dual-source CT scanner (SOMATOM)	All patients were treated with nCRT followed by surgery	AUC=0.73, P=0.009
28. Hou et al., 2017	Enhanced CT: Philips Brilliance 6	All patients were treated with definitive CCRT	ANN: accuracy=0.972, AUC=0.927 (Training) accuracy=0.917, AUC=0.800(Testing) SVM: accuracy=0.891, AUC=0.818(Training) accuracy=0.667, AUC=0.600(Testing)
29. Yip et al., 2016	PET scanner: GE Discovery or Siemen Biograph (Siemens) Images were acquired 65 minutes after tracer injection	All patients were treated with nCRT followed by surgery	AUC = 0.72–0.78
30. Rossum et al., 2016	PET/CT system (Discovery RX, ST, STE, or HR; GE) Images were acquired 60-90 minutes after tracer injection. Enhanced CT	All patients were treated with nCRT followed by surgery	c-index=0.82(95% CI 0.75–0.88) (apparent) c-index=0.77 (95% CI 0.70–0.83) (corrected)
31. Ypsilantis et al., 2015	Scanner: n.r.	All patients were treated with nCRT	Sensitivity :80.7±11.5Specificity: 81.6±9.2 Accuracy: 73.4±5.3

32.	Yip et al., 2014	Contrast CT: 16–, 128–, or 256–detector row CT scanner (Phillips)	All patients were treated with definitive CCRT	AUC=0.802
33.	Zhang et al., 2014	Same as Tan et al., 2013	All patients were treated with nCRT followed by surgery	AUC=1 (no misclassifications)
34.	Tan et al., 2013	16-slice Gemini PET/CT scanner (Philips) Images were acquired sixty minutes after tracer injection	All patients were treated with nCRT followed by surgery	Texture feature: AUC=0.83, p=0.01 Bin-to-bin and cross-bin histogram distances: AUC=0.78-0.89, p=0.04
35.	Hatt et al., 2013	Philips GEMINI PET/CT scanner Images were acquired sixty minutes after tracer injection	All patients treated with exclusive CCRT	(best) AUC=0.90
36.	Tan et al., 2013	Same as Tan et al., 2013	All patients were treated with nCRT followed by surgery	(best) AUC=0.85
37.	Tixier et al., 2011	Gemini PET/CT scanner (Philips). Images were acquired on average 54 min after injection	All patients were treated with exclusive CCRT	Sensitivity: 76%-92% Specificity: 56%-91%

Abbreviations used in the table - nCRT: neoadjuvant chemoradiotherapy; CCRT: concurrent chemoradiotherapy.

Supplementary Table 3: Assessment of methodological quality of included studies

Number	Ref	Prospective registration	Imaging protocol	Image pre-processing	Segmentation method	Repeatability, reproducibility, and dimensionality reduction	Correlations with non-handcrafted biomarkers	Justification of risk groupings	Validation method	Comparison to clinical/non-imaging model or Holistic model	Discrimination statistics	Model calibration	Estimation of clinical utility	External validation availability
1.	Xie *et al.*, 2021	**Poor**	**Good**: Several scanners from two institutions, main details provided in suppl	**Poor**: Only resample in suppl	**Moderate**: ROI was manually delineated, but no details about checking	**Good**: Inter-observer, ComBat method for feature harmonization, overlapped genes were used as a filter for the selection of radiomics features (Pearson), univariate analysis, LASSO	**Good**: Correlations with genes and tumor volume	**Moderate**: The cut-off points for the nomograms were determined by Youden Index	**Good**: External validation was performed in a completely independent group of patients, and internal validation	**Good**: Building holistic model combined with genes	**Good**: C-index, AUC, Time-dependent AUC, Kaplan-Meier curve	**Good**: Calibration plots for nomogram models 1 and 2 of training, internal test, external test set	**Good**: Decision curve	**Moderate**: Rad-score based Nomogram is presented, without the coefficients of the features.
2.	Beukinga *et al.*, 2021	**Poor**	**Poor**: One scanner, only details of PET	**Moderate**: Only resampled and normalized	**Good**: Delineated manually after reaching consensus	**Moderate**: No check of either repeatability or	**Poor**: No	**Poor**: No details about cut-off	**Moderate**: Lacking an independent test or validation cohort;	**Good**: Building holistic model combined with clinical	**Poor**: Only AUC	**Good**: Calibration plots	**Poor**: No	**Moderate**: The features without coefficients are given. No

			acquisition, but no details of CT imaging protocol		between 3 collaborating researchers	reproducibility; univariable logistic regression analysis and LASSO to select features			internally validated by a bootstrap approach with 20000 repetitions	factors, HER2 and CD44				Nomogram or online code or model.
3.	Hu *et al.*, 2021	**Poor**	**Good**: Two scanners, details provided in supplementary table	**Good**: details including resampling, filter, and reconstruction	**Good**: Two senior radiologists respectively delineated the ROI	**Good**: Multi-user delineations to choose features with reproducibility; Pearson correlation coefficient, decision tree, and wrapper method to find features	**Good**: The association of radiomics features and corresponding pathophysiological features (Radiogenomics)	**Moderate**: Cut-offs were determined by Youden index, diagnostic possibilities calculated by cut-off were performed in suppl.	**Good**: External validation was performed in a completely independent group of patients (Guangzhou and Hong Kong)	**Good**: Compared with the clinical model; to elucidate the pathophysiological association with the radiomics signatures (radiogenomics)	**Good**: AUC (95%CI), accuracy, sensitivity, specificity, positive predictive value, negative predictive value; No details of cross-validation or bootstrapping	**Good**: Calibration curve in suppl. But no slope, intercept, P-value was reported.	**Good**: Decision curve in suppl.	**Good**: Pretrained model and selected handcrafted features are presented. Model available online.
4.	Wang *et al.*, 2021	**Poor**	**Good**: Scanners from two institutions, main details provided in the article	**Poor**: No details on digital filters or resampling, intensity discretization reported.	**Good**: semiautomatic segmentation method; verified by a senior radiologist	**Good**: 2-months interval from two-time point to calculate features, intra-class correlatio	**Poor**: No	**Moderate**: cut-off values of parameters were determined using receiver operating characteristics (ROC)	**Good**: External validation was performed in a completely independent group of patients	**Good**: Building holistic model combined with clinicopathological, dosimetric, and hematologic	**Good**: C-index, risk classification	**Good**: Calibration plots for nomogram model	**Good**: Decision curve	**Good**: Nomogram and coefficients of features are provided.

						n coefficient to collected reproducible features; LASSO		curve		al indicators				
5.	Li Yimin *et al.*, 2020	**Poor**: No prospective registration	**Good**: Two scanners, details provided in the article	**Good**: details including resampling, filter, and reconstruction	**Moderate**: Semi-automatic delineation, no details of how many people checked ROI	**Moderate**: No check of either repeatability or reproducibility of features; features selected by repeated Lasso-Cox regression 100 times	**Good**: tested against MTV	**Moderate**: Cut-offs were shown by Kaplan-Meier estimates	**Good**: External validation was performed in a completely independent group of patients (Xiamen and Berlin)	**Good**: Compared with MTV and SUV model in suppl. (failed to identify patients with high or low risk for local recurrence)	**Poor**: Internal and external validated by Kaplan-Meier estimates	**Poor**: No calibration	**Poor**: No decision curve or cost-benefit analysis	**Good**: The selected features and weightings are given.
6.	Xie *et al.*, 2020	**Poor**	**Moderate**: Only one scanner, main details provided in reference	**Moderate**: Four filters with different widths; but no details of resampling and intensity discretization	**Good**: The delineation was performed manually with consensus between three radiation oncologists, and refined by an additional thresholding procedure	**Poor**: No check of either repeatability or reproducibility of features; Cox proportional hazards model were performed on each	**Poor**: no correlation testing against non-radiomics features.	**Good**: Median as cut-offs, which were presented in the Figures	**Poor**: Lacking an independent test or validation cohort.	**Poor**: No	**Good**: AUC, P-value; Kaplan-Meier curve	**Poor**: No	**Poor**: No	**Good**: The selected features and coefficients are provided.

					that excluded pixels less than -50HU	texture parameter								
7.	Hu *et al.*, 2020	**Poor**	**Good**: Two scanners, details provided in supplementary table	**Good**: details including resampling, filter, and reconstruction	**Good**: Two senior radiologists respectively delineated the ROI	**Good**: Multi-user delineations to choose features with reproducibility; Pearson correlation coefficient, decision tree, and wrapper method to find features	**Good**: The association of radiomics features and corresponding pathophysiological features (Radiogenomics)	**Moderate**: Cut-offs were determined by Youden index, diagnostic possibilities calculated by cut-off were performed in suppl.	**Good**: External validation was performed in a completely independent group of patients (Guangzhou and Hong Kong)	**Good**: Compared with the clinical model; to elucidate the pathophysiological association with the radiomics signatures (radiogenomics)	**Good**: AUC (95%CI), accuracy, sensitivity, specificity, positive predictive value, negative predictive value; No details of cross-validation or bootstrapping	**Good**: Calibration curve in suppl. But no slope, intercept, P-value was reported.	**Good**: Decision curve in suppl.	**Moderate**: Only selected features are given, no coefficients.
8.	Luo *et al.*, 2020	**Poor**	**Moderate**: Only one scanner, details provided in the article	**Poor**: Only resample information	**Moderate**: Details of segmentation are in the article. The same observer after two months later repeated the tumor segmentation to evaluate the reproducibility	**Good**: Intra-observer and inter-observer intra-class correlation coefficient analysis; LASSO logistic regression to identify	**Poor**: No	**Poor**: No details about cut-off	**Poor**: 66 cases were allocated into the testing set from the same institution	**Good**: Building holistic model combined with clinical factors	**Moderate**: AUC, 95%CI	**Poor**: No	**Good**: Decision curve	**Good**: Rad-score-based Nomogram is provided, with coefficients of features.

						the optimal features								
9.	Li Yue *et al.*, 2020	**Poor**	**Good**: Several scanners from one institution, main details provided in suppl	**Poor**: only reconstruction	**Good**: ROI was manually Delineated by two expert radiation oncologists	**Good**: Inter-observer, test-retest; univariate logistic regression and LASSO	**Moderate**: None of the selected radiomics feature is correlated with the analyzed clinical features	Poor: No details about cut-off	**Moderate**: Lacking an independent test or validation cohort; internally validated by a bootstrap approach with 2000 repetitions	**Good**: Building holistic model combined with clinical factors	**Poor**: Only AUC	**Good**: Calibration plots	**Good**: Decision curve	**Good**: Rad-score-based Nomogram is provided, with coefficients of features.
10.	Zhang *et al.*, 2020	**Poor**	**Good**: Five scanners, details provided in suppl.	**Moderate**: Details of resampling and reconstruction, but no details of filter and intensity discretization	**Moderate**: GTV for radiotherapy planning were used as ROI, but no details of the checking	**Moderate**: No check of repeatability; Pearson correlation was used; Recursive Feature Elimination (RFE) and LASSO was applied to select optimal predictors	**Poor**: no correlation testing against non-radiomics features.	**Poor**: no risk group analysis.	**Good**: External validation was performed in a completely independent group of patients	**Good**: Combined and compared with models developed with clinical variables	**Good**: AUC, accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV); 2000 stratified bootstrap	**Good**: Calibration plots for three models	**Poor**: No	**Moderate**: The selected features are provided. No details of calculating Rad-score.
11.	Du *et al.*, 2020	**Poor**	**Moderate**: Only one scanner, details provided	**Poor**: No details on digital filters or resampling,	**Good**: semiautomatic segmentation method; verified by a	**Good**: Spearman rank correlation test to delete the	**Poor**: No	**Poor**: No details about cut-off	**Moderate**: Lacking an independent test; 50 iterations of 10-fold	**Good**: Building holistic model combined with clinical	**Good**: AUC (95%CI), accuracy, sensitivity, specificity, 50 iterations	**Good**: Calibration plots for nomogram model	**Good**: Decision curve	**Good**: Rad-score-based Nomogram is provided, with coefficients

			in the article	intensity discretization reported.	senior radiologist	features with correlations greater than 0.9; univariate analysis and LASSO feature selection			nested cross-validation, random sampling was conducted	features and dosimetric parameters	of 10-fold nested cross-validation			of features.
12.	Foley *et al.*, 2019	**Poor**	**Good**: Five scanners, details provided in suppl.	**Moderate**: Derails of STAGE set, but no details of CROSS set.	**Moderate**: Details of segmentation are in the article. But no details of checking	**Moderate**: No check of either repeatability or reproducibility of features; Features from previous knowledge	**Poor**: no correlation testing against non-radiomics features.	**Good**: Patients were separated into quartiles	**Good**: External validation was performed in a completely independent group of patients	**Good**: Compared with the model developed with clinical factors	**Poor**: No details about discrimination metrics	**Good**: Calibration slope and P-value	**Poor**: No decision curve or cost-benefit analysis and no detailed instructions on how to use prognostic model	**Good**: The selected features with coefficients are provided.
13.	Xie *et al.*, 2019	**Poor**	**Good**: Two scanners, details provided in the article	**Moderate**: Details other than filters are in the article	**Moderate**: For the training set, GTV for radiotherapy planning were used as ROI, but no details of the checking; For the testing set, details provided in the article	**Moderate**: Pearson correlation method; LASSO to select features; lack inter-observer test or test-retest	**Good**: Spearman rank correlation test was used to assess the correlation of radiomics features with clinical factors and CNAs	**Moderate**: Youden index was used to select the cut-off value	**Good**: External validation was performed in a completely independent group of patients	**Good**: compared with models based on conventional prognostic factors, clinical stage, and ECOG PS	**Good**: C-index, ROC, Kaplan-Meier, and Log-rank	**Poor**: No	**Poor**: No decision curve or cost-benefit analysis and no detailed instructions on how to use prognostic model	**Good**: Rad-score-based Nomogram is provided, with coefficients of features.
14.	Wang *et*	**Poor**	**Good**:	**Poor**: No	**Good**:	**Good**:	**Good**:	**Moderate**:	**Good**:	**Good**:	**Good**: C-	**Poor**: No	**Poor**: No	**Good**:

	al., 2019		Several scanners from four institutions, main details provided in suppl	details on digital filters or resampling, intensity discretization reported.	Delineations were performed by three radiation oncologists in each center	Stability and intro-observer were tested in suppl; only one feature, uni/multivariate analysis were used	Correlations with tumor length or TNM staging	Cut-off was identified using x-tile software	External validation was performed in three completely independent groups of patients	Compared with models based on clinical factors	index, 95%CI; Cox models were used to test the risk model based on compactness(feature) was able to independently predict OS and PFS			Features and cut-off points.
15.	Chen *et al.*, 2019	**Poor**	**Moderate**: Only one scanner, main details provided in the article	**Poor**: Only reconstruction information	**Poor**: Delineation based on images and SUV (cut-off value 2.5), but no description of how many doctors checked ROI	**Poor**: Four radiomics features, no check of repeatability, reproducibility, or dimensionality reduction of features	**Poor**: no correlation testing against non-radiomics features.	**Moderate**: Youden index was used to select the cut-off value that was used to develop a scoring system for predicting OS.	**Poor**: Only 16 cases were allocated into the testing set from the same institution	**Poor**: No	**Poor**: Only univariate and multivariate analyses of factors associated with outcomes (HR and P-value)	**Poor**: No	**Poor**: No	**Good**: The details of calculating the risk score are provided.
16.	Yan *et al.*, 2019	**Poor**	**Poor**: One scanner, no details on pixel spacing	**Poor**: No details on digital filters or resampling, intensity discretization reported.	**Poor**: No details on segmentation method or indication of how many doctors to check ROI	**Poor**: No check of either repeatability or reproducibility of features; no method of reduction of features	**Poor**: no correlation testing against non-radiomics features.	**Good**: The median change of only one feature as cut-off, two cohorts divided by cut-off were analyzed by KM analysis (P-value)	**Poor**: Lacking an independent test or validation cohort.	**Poor**: No	**Poor**: Only correlation between outcomes and features (r and P-value)	**Poor**: No	**Poor**: No	**Good**: Features are provided, and the cut-off point of one feature is given.

17.	Yang *et al.*, 2019	**Poor**	**Moderate**: Only one scanner, main details provided in the article	**Poor**: Details of filters are in the article. No details on intensity discretization were reported. No resample of the voxel size was used.	**Moderate**: Details of segmentation are in the article. But no details of checking	**Moderate**: No check of either repeatability or reproducibility of features; LASSO with 10-fold was applied to select optimal predictors	**Poor**: no correlation testing against non-radiomics features.	**Poor**: no risk group analysis.	**Poor**: Only 11 cases were allocated into the testing set from the same institution	**Poor**: No	**Moderate**: AUC, 95%CI	**Poor**: No	**Poor**: No	**Good**: The selected features with coefficients are provided.
18.	Jin *et al.*, 2019	**Poor**	**Moderate**: Only one scanner, main details provided in the article	**Poor**: No details on digital filters or resampling, intensity discretization reported.	**Moderate**: GTV for radiotherapy planning were used as ROI, but no details of the checking	**Poor**: No check of either repeatability or reproducibility of features; PCA was applied before training classifier	**Poor**: no correlation testing against non-radiomics features.	**Poor**: no risk group analysis.	**Moderate**: Only 24 cases were allocated into the testing set from the same institution; randomly partition was performed ten times	**Good**: Building model combined with dosimetric parameters	**Moderate**: Accuracy, AUC; the partition was performed ten times, 10-fold cross-validation was performed on the training set	**Poor**: No	**Poor**: No	**Moderate**: Only selected features are given, no coefficients.
19.	Foley *et al.*, 2018	**Poor**	**Moderate**: Only one scanner, main details provided in the article	**Moderate** Details other than filters are in the article	**Moderate**: Details of segmentation are in the article. But no details of checking	**Moderate**: No check of either repeatability or reproducibility of features; Features from	**Poor**: no correlation testing against non-radiomics features.	**Good**: Patients were separated into quartiles	**Moderate**: 101 samples set acquired from the same institution, but different times were allocated into the internal	**Good**: Compared with clinical factors, SUV and MTV and build a holistic model	**Poor**: Log-rank test evaluated significant differences in OS	**Poor**: No	**Poor**: No decision curve or cost-benefit analysis and no detailed instructions on how to use prognostic model	**Good**: The selected features with coefficients are provided.

						previous knowledge			validation set and the model was validated using KM plots.					
20.	Larue *et al.*, 2018	**Poor**	**Good**: Five scanners, details provided in suppl.	**Good**: details in the article and suppl.	**Moderate**: GTV for radiotherapy planning were used as ROI, but no details of the checking	**Moderate**: No check of repeatability and reproducibility. Selecting the 40 most important predictors and then these features were used as input for RF model	**Poor**: no correlation testing against non-radiomics features.	**Poor**: No details about cut-off	**Good**: External validation was performed in a completely independent group of patients	**Good**: Compared with the model developed with clinical factors	**Good**: AUC, 95%CI, and Kaplan-Meier curve	**Poor**: No	**Poor**: No decision curve or cost-benefit analysis and no detailed instructions on how to use the prediction model	**Moderate**: Only features are provided without coefficients. No Nomogram or online code or model.
21.	Beukinga *et al.*, 2018	**Poor**	**Moderate**: Only one scanner, main details provided in the article	**Good**: details in the article	**Moderate**: ROI was manually delineated by an expert radiation oncologist, but no details about checking of other doctors	**Good**: Details of intra-class correlation coefficient and Akaike Information Criterion. LASSO to select features.	**Poor**: no correlation testing against non-radiomics features.	**Poor**: No details about cut-off	**Moderate**: Lacking an independent test. Only internal validation by bootstrap resampling with 20000 replicates.	**Good**: Compared with clinical factors, and SUV and build a holistic model	**Good**: ROC and discrimination slope	**Good**: Calibration slope and intercept	**Poor**: No	**Moderate**: Only features are provided without coefficients. No Nomogram or online code or model.
22.	Riyahi *et*	**Poor**	**Moderat**	**Poor**: No	**Moderate**:	**Moderat**	**Poor**: no	**Poor**: No	**Poor**:	**Good**:	**Moderate**:	**Poor**: No	**Poor**: No	**Moderate**:

	al., 2018		e: Only one scanner, main details provided in reference	details on digital filters or resampling, intensity discretization reported.	Details of segmentation are in the article. But no details of checking	e: Used a pair-wise correlation cutoff, and selected features by SVM coupled with LASSO	correlation testing against non-radiomics features.	details about cut-off	Lacking an independent test or validation cohort.	Compared and combined with PET/CT features model	AUC, accuracy, sensitivity, specificity			Only features.
23.	Paul *et al.*, 2017	**Poor**	**Poor**: Only scanner and only voxel size were provided	**Poor**: No details on digital filters or resampling, intensity discretization reported.	**Poor**: No details on segmentation method or indication of how many doctors to check ROI	**Moderate**: No check of repeatability. Spearman's rank correlation analysis to eliminate correlated features; Genetic Algorithm based on Random Forest and LASSO to select features	**Poor**: no correlation testing against non-radiomics features.	**Poor**: no risk group analysis.	**Poor**: Lacking an independent test or validation cohort.	**Poor**: No	**Good**: Cross-validation method; AUC (95%CI), accuracy, sensitivity, specificity, positive predictive value, the negative predictive value	**Poor**: No	**Poor**: No	**Moderate**: Only features.
24.	Desbordes *et al.*, 2017	**Poor**	**Poor**: Only scanner and no details of scan protocols	**Good**: details in the article	**Moderate**: Details of segmentation are in the article. But no details of checking	**Moderate**: Spearman correlation rank analysis;	**Good**: Correlation with SUV and MTV	**Poor**: No details about cut-off	**Moderate**: Lacking an independent test. Only randomly divided the database	**Good**: Compared with models based on clinical factors, SUV and	**Good**: AUC, accuracy, sensitivity, specificity; Kaplan-Meier	**Poor**: No	**Poor**: No	**Moderate**: Only features.

						Random Forest to select features			into training and test set, and repeated 10 times	MTV	survival curves			
25.	Nakajo *et al.*, 2017	**Poor**	**Moderate**: Only one scanner, main details provided in reference	**Moderate**: Reconstruction and resampling of the intensity of FDG uptake represented in the article	**Good**: ROI was manually set, tumor boundaries were then automatically contoured. The focal uptake in the primary lesion was visually interpreted by three doctors	**Poor**: No check of either repeatability or reproducibility of features; univariate and multivariate analysis of features	**Good:** Correlation with PET features	**Moderate**: Cut-offs by using ROC, presented in TableS2; K-M curves of different patients divided by cut-off	**Poor**: Lacking an independent test or validation cohort.	**Good**: compared with models based on clinical factors	**Good**: AUC, accuracy, sensitivity, specificity; Kaplan-Meier survival curves	**Poor**: No	**Poor**: No	**Moderate**: Only features.
26.	Beukinga *et al.*, 2017	**Poor**	**Moderate**: Only one scanner, main details provided in the article	**Poor**: No details on digital filters or resampling, intensity discretization reported.	**Good**: The delineation was performed manually with consensus between three radiation oncologists	**Good**: The stability of features was evaluated with the intra-class correlation (ICC); potential predictors that met the Akaike Information Criterion (AIC)	**Poor**: no correlation testing against non-radiomics features.	**Poor**: No details about cut-off	**Moderate**: Lacking an independent test or validation cohort; internally validated by a bootstrap approach with 2000 repetitions	**Good**: Compared with the model developed with clinical factors	**Good**: AUC and discrimination slope; internally validated by bootstrap with 2000 repetitions	**Good**: Calibration slope and intercept	**Poor**: No	**Poor**: No exact features and coefficients or final model are given.

						were considered significant and selected by LASSO								
27.	Wakatsuki *et al.*, 2017	**Poor**	**Moderate**: Only one scanner, main details provided in the article	**Poor**: No details on digital filters or resampling, intensity discretization reported.	**Poor**: No details about segmentation method and checking	**Poor**: No check of either repeatability or reproducibility; CT number of a primary tumor as a feature, no need to do dimensionality reduction	**Good**: Correlation with Ki-67, P53, and CK5/6 expression	**Moderate**: According ROC to define cut-off, and used it to calculate Kaplan-Meier	**Poor**: Lacking an independent test or validation cohort.	**Poor**: No	**Moderate**: AUC and Kaplan-Meier	**Poor**: No	**Poor**: No	**Good**: Cut-off point of the feature.
28.	Hou *et al.*, 2017	**Poor**	**Moderate**: Only one scanner, details provided in the article	**Moderate**: Only resampled and normalized	**Good**: Manually delineated by two doctors, and reviewed by the third doctor	**Moderate**: Intra-class correlation coefficient was used to quantify reproducibility; Kruskal-Wallis test and wrapper-based feature	**Poor**: No	**Moderate**: According to ROC to define cut-off	**Poor**: Only 12 cases were allocated into the testing set from the same institution	**Poor**: No	**Good**: AUC, accuracy, sensitivity, specificity, positive predictive value, negative predictive value; No details of cross-validation or bootstrapping	**Poor**: No	**Poor**: No	**Moderate**: Only features are provided without coefficients. No Nomogram or online code or model.

						selection method to select features								
29.	Yip *et al.*, 2016	**Poor**	**Poor**: Two scanners, but no details of imaging protocol	**Poor**: Only reconstruction of PET data	**Poor**: No details about segmentation method and checking	**Moderate**: Three features were chosen due to a previous report predicting potential	**Good**: Correlation with MTV and tumor volume	**Poor**: No details about cut-off	**Poor**: Lacking an independent test or validation cohort.	**Moderate**: Only AUC of MTV was reported, but no figures	**Poor**: AUC of models based on different algorithm propagated contours	**Poor**: No	**Poor**: No	**Moderate**: Only features.
30.	Rossum *et al.*, 2016	**Poor**	**Good**: Several scanners, main details provided in suppl	**Good**: Details were provided in suppl	**Moderate**: Semiautomatic delineation method, followed by 1 interpreter	**Good**: Test-retest (ICC), Spearman rank correlation; univariable and multivariable analysis to select features; all these details provided in suppl	**Poor**: No	**Poor**: No details about cut-off	**Moderate**: Lacking an external test. Only internal validation by bootstrap method with 1000 repetitions	**Good**: Compared with models based on clinical factors, and different PET features	**Poor**: Apparent and corrected C-indices	**Good**: Calibration plot for all four models	**Good**: Decision-curve	**Good**: The selected features with coefficients are provided.

Chapter 3: Dual discriminator Super-Resolution Generative Adversarial Network-based synthetic GGO nodule image augmentation

Adapted from: Zhixiang Wang; **Zhen Zhang*;** Ying Feng; Lizza E. L. Hendriks; Razvan L. Miclea; Hester Gietema; Janna Schoenmaekers; Andre Dekker, Leonard Wee; Alberto Traverso. Generation of Synthetic Ground Glass Nodules Using Generative Adversarial Networks (GANs). Eur Radiol Exp 2022, 6 (1), 59. https://doi.org/10.1186/s41747-022-00311-y.*

** indicates equal contributions*

Abstract

Background Data shortage is a common challenge in developing computer-aided diagnosis systems. We developed a generative adversarial network (GAN) model to generate synthetic lung lesions mimicking ground glass nodules (GGNs).

Methods We used 216 computed tomography images with 340 GGNs from the Lung Image Database Consortium and Image Database Resource Initiative database. A GAN model retrieving information from the whole image and the GGN region was built. The generated samples were evaluated with visual Turing test performed by four experienced radiologists or pulmonologists. Radiomic features were compared between real and synthetic nodules. Performances were evaluated by area under the curve (AUC) at receiver operating characteristic analysis. In addition, we trained a classification model (ResNet) to investigate whether the synthetic GGNs can improve the performances algorithm and how performances changed as a function of labelled data used in training.

Results Of 51 synthetic GGNs, 19 (37%) were classified as real by clinicians. Of 93 radiomic features, 58 (62.4%) showed no significant difference between synthetic and real GGNs ($p \geq 0.052$). The discrimination performances of physicians (AUC 0.68) and radiomics (AUC 0.66) were similar, with no-significantly different ($p = 0.23$), but clinicians achieved a better accuracy (AUC 0.74) than radiomics (AUC 0.62) ($p < 0.001$). The classification model trained on datasets with synthetic data performed better than models without the addition of synthetic data.

Conclusions GAN has promising potential for generating GGNs. Through similar AUC, clinicians achieved better ability to diagnose whether the data is synthetic than radiomics.

Keywords: Deep learning, Tomography (x-ray computed), Lung, Neural networks (computer), Solitary pulmonary nodule

Introduction

Artificial intelligence is a rapidly developing field including many applications in computer vision, such as deep learning (DL) and machine learning methods for the segmentation [1] and the classification [2] of anatomical structures and abnormalities in standard of care diagnostic imaging. A strong effort is dedicated to the implementation of these methods as computer-aided diagnosis (CAD) tools to reduce the time burden of clinical tasks and improve radiologists' detection accuracy. For lung cancer screening, the number of CAD systems to automatically identify the presence of pulmonary nodules has exponentially increased in the last 10 years. DL methods have shown an increased detection accuracy for all the types of pulmonary nodules (solid, part solid, ground glass opacities) compared to traditional machine learning methods in low-dose screening computed tomography (CT) scans [3, 4].

The success of developing robust and widely applicable deep learning-based CAD systems relies on the availability of a large amount of curated and annotated data. However, annotating data consistently has a cost and is dependent on radiologists' time and availability. Even when large amount of data is collected for training DL networks, the problem of class imbalance may exist. The class imbalance problem refers to some labels (classes) being more frequent than others. Due to this unbalance, the DL network will learn better how to classify the more frequent samples, with degraded performances for the minority class(es) [5]. In the specific case of pulmonary nodule detection, ground glass nodules (GGN), although accounting for only 2.7 to 4.4% of all nodules, are malignant in 63% of the cases [6].

Next to classical statistical methods such as SMOTE (synthetic minority oversampling technique), researchers have investigated more advanced methods for generating synthetic samples of original data, to increase and balance the original sample size of the training dataset. Recently, generative adversarial networks (GANs) have been proposed as a method to generate synthetic images to improve the existing oversampling techniques [7]. GANs, which are DL algorithms based on game theory, have been applied to several computer vision tasks such as image denoising, reconstruction, and, as mentioned, synthetic data generation [8, 9]. Briefly, GANs consists of two competing actors: a generator and a discriminator. They are used to generate synthetic images/samples and "judge" the quality of the generated images, respectively. The equilibrium is reached when the synthetic (i.e., fake) samples cannot be distinguished from the real distribution [10].

While many studies demonstrated the potential of GANs to generate synthetic images, the generated images/samples have not been evaluated by radiologists, and this limits the acceptance and use of GANs in a clinical setting. In fact, generated images/samples should be representative of the "real" population. However, by only focusing on evaluating at the "human-level" appropriateness of synthetic samples, it is not possible to draw any conclusion whether the introduction of synthetic samples in the training samples will improve the detection performances of CAD systems. In principle, it is expected that adding as many synthetic samples as possible to the original data will lead to a CAD system with better detection performances. It is important to notice that generating synthetic samples via GAN is in itself a learning procedure, where the original data is used to train the networks to generate the synthetic samples. The ratio between original data available and the quality of generated samples is not clear yet.

In this study, we investigated the following research questions:

Is it possible to use a GAN model to generate synthetic GGNs on low-dose screening CT scans that are undisguisable by clinicians from the real samples?

How much labelled data is needed to generate synthetic GGNs of sufficient quality to train a CAD for pulmonary nodule detection achieving the same level of performance of a large amount of labelled data?

To answer these questions, we developed an optimized GAN model with dual discriminators to generate GGNs.

Methods

1. Study population

A total of 216 subjects were selected from The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) database for this study [11]. In this database, the nodules were classified into five grades by four radiologists: 1 = ground glass opacity (GGO1); 2 = intermediate between 2 and 3; 3 = part solid; 4 = intermediate between 4 and 5; 5 = solid. We chose 340 GGN nodules of grades 1 or 2 that were annotated by at least two radiologists for our study. To ensure data quality, further confirmation was performed by a radiologist (author Z.Z.), with 5 years of experience in lung CT, to verify that all the nodules were GGNs.

2. Image preprocessing

In the preprocessing methods, first, the two-dimensional slices with annotation as GGN from the CT volume were extracted. Second, in order to avoid interference from external tissues of the lung, we first cropped the lungs from the tissue and background with a seed-filling algorithm, which starts from an inner point of the polygon area and draws points with the given grey level from inside to outside until the boundary is found. Third, the cropped images were padded by 0 in the background to keep every image having the same sizes (512 × 512) in the dataset. Fourth, we normalized the data to the range 0–1, as is the standard practice in computer vision. Fifth, we erased the nodules from the original position and saved them as region of interest (ROI) for the training set. In general, each training batch contained two images: the original image as the target image, which serves as the ground truth for the generator (as shown in Figs. 1 and 2), and another image is the input image, in which stripped the nodule area, i.e., the ROI region was processed as blank for the input image. As shown in Fig. 1, the network generates the nodule from the input image. In addition, after generation, there are two discriminators (whole image discriminator and ROI discriminator) to evaluate the quality of the whole image and the ROI where the nodule is.

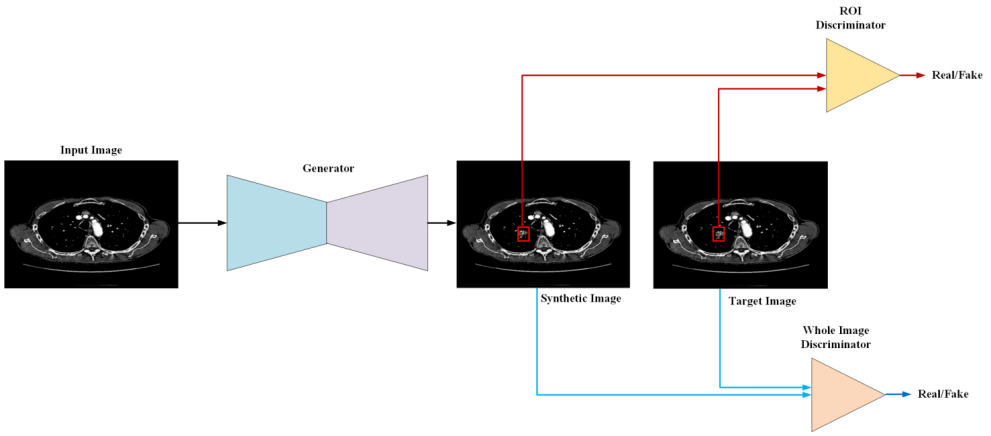


Figure 1. The pipeline for training the model. First, the generator synthetizes ground glass nodules from the background according to the input image. Second, the region of interest (ROI) discriminator (red line) and the whole image discriminator (blue line) extract features from the ROI and whole image to classify the synthetic image and the target whether the synthetic image is real.

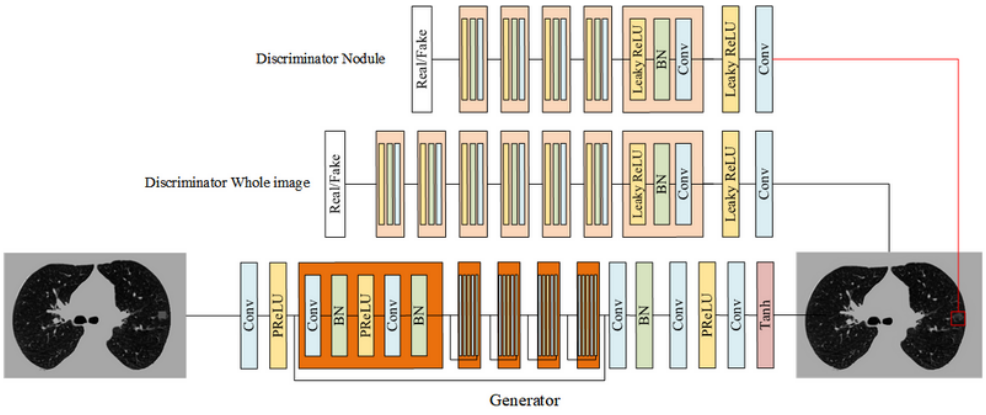


Figure 2. The structure of the network. The generator creates the synthetic ground glass nodule at the position where the mask in the input. The generator is composed of convolutional layers with a kernel size of 3 × 3, the batch normalization, and the "parametric rectified linear unit" (PReLU) activation function. The discriminator was composed of convolutional layers with a kernel size of 3 × 3, the batch normalization, and the leaky PReLU activation function.

3. Construction of the DL model

The super-resolution generative adversarial network (SRGAN) was used as the backbone of the generator [12]. SRGAN compares the features difference in the model between a pair of data and train the discriminators to improve the realism of the recovered images. Both the whole image discriminator and ROI discriminator are based on a ResNet [13] which is

a widely used classical classification networks combined by residual blocks with different input sizes and depths of the network. The structure of the network is shown in Fig. 2. For training the network, the loss function was as follows:

$$L_{D2SRGAN} = (L_{ssim} + L_{adversarial})_{whole\ image} + (L_{ssim} + L_{adversarial})_{ROI\ image} \quad (1)$$

$$L_{adversarial} = \sum_{n=1}^{N} -logD(G(x)) \quad (2)$$

$$L_{ssim}(x, y) = 1 - \frac{(2\mu_x\mu_y + C_1) + (\sigma_{xy} + C_2)}{(\mu_x{}^2\mu_y{}^2 + C_1)(\sigma_x{}^2\sigma_y{}^2 + C_2)} \quad (3)$$

The Lssim can be used to compare the similarity between two images. In this loss function, the whole image is separated into two parts to calculate the loss function respective. G and D represent the generator and discriminator, x is the input of the generator. μx and μy represent the average of input x,y respectively. σx and σy represent the standard deviation of input x,y respectively. σxy is the covariance of x and y. C1 and C2 are constants to avoid system errors caused by the denominator being zero.

All images were loaded with an unchanged original size of 512 × 512. The input size of the discriminator for the whole image and the ROI image were 512 × 512 and 32 × 32, respectively. An Adam optimizer was used to train both the generator and the discriminator with a learning rate of 0.0001. This model was trained using an NVIDIA Tesla V100 SXM2 32 GB graphics processing unit.

4. Evaluation of model performance

We evaluated the model performance using both subjective (visual Turing test, VTT) and objective (radiomics) approaches. VTT is an assessment method that evaluation the ability of a human or doctors to identify attributes and relationships from images [14]. Subjective evaluations were performed by two radiologists (authors R.M. and H.G.) and two pulmonologists (authors L.H. and J.S.), who all had more than 5 years of experience in lung CT imaging and on a daily basis evaluate chest CT scans. One hundred images (50 real and 50 synthetic GGNs) were divided into four batches and converted to a DICOM (Digital Imaging and COmmunications in Medicine) file with 25 slices of images, and each physician was randomly assigned to one of these batches. The physicians categorized the real and synthetic GGNs into four classes based on this categorical scale: confidently fake, leaning fake, leaning real, and confidently real.

To perform an objective evaluation, radiomic features were calculated from the original and generated data. Radiomics refers to the extraction of quantitative information from medical images by computing the statistical, morphological, and texture features. The following feature categories were extracted using the open source Pyradiomics package (version 3.0.1) with default values: first order statistics (n = 18), grey level co-occurrence matrix (n = 24),

grey level dependence matrix (n = 14), grey level run length matrix (n = 16), grey level size zone matrix (n = 16), and neighbouring grey tone difference matrix (n = 5) [15–17].

The Kolmogorov–Smirnov test was used for the analysis of whether the distribution of radiomics features were similar between the real and synthetic GGNs. We considered significant p values lower than 0.05.

The results of the subjective and objective evaluations were analysed using the area under the curve (AUC) at receiver operating characteristic analysis. For the subjective evaluation, we took into consideration the VTT results. For the objective evaluation, to compare the classification ability of radiomics and radiologist, a logistic regression model was build based on radiomic features to classify both real and synthetic GGNs. The same dataset was used for the physician evaluations and the radiomics logistic regression model, with a four-fold cross-validation.

In addition, we also investigated whether the synthetic GGNs can improve the performance of a CAD algorithm trained for recognizing GGNs from all types of nodules in the LIDC-IDRI dataset and how the performance changed as a function of labelled data used in the training.

As a CAD, we used a basic ResNet as the DL classification network with a cross-entropy loss function. First, we separated the dataset into 10 training subsets and an independent test set. We trained the classification network on portions of the original data ranging from 10 to 100% of the real data and we separately inferred on the test set. Then, we trained the classification network on the original data added systematic data generated by the GAN network trained in 10% to 100% of real data.

Results

Examples of synthetic GGNs generated in different parts of the lungs with different surrounding tissues are shown in Fig.3. Nodules classified as fake (Fig.3b) show more unnatural characteristics in terms of intensity and morphology than nodules classified as “real” (Fig.3a); specifically, “fake nodules” have very high fixed grey values and regular shapes such as rectangles.

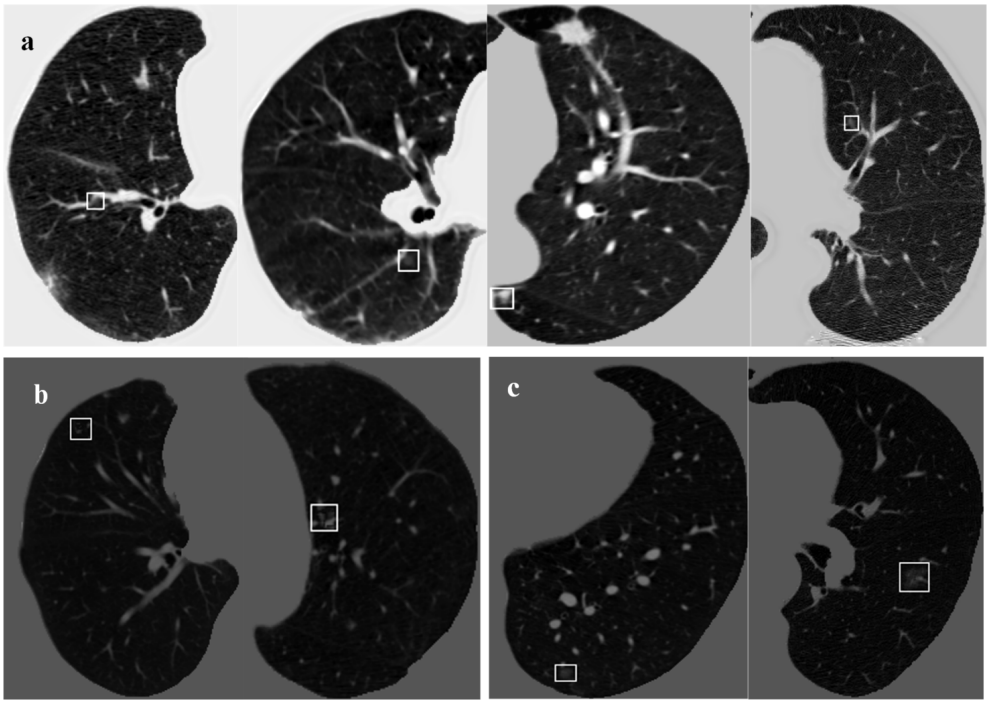


Figure 3. Examples of synthetic ground glass nodules (GGNs), the GGNs were categorised by physicians to four categories: confidently fake, leaning fake, leaning real, and confidently real. a Synthetic GGNs classified as "real" by clinicians. b Synthetic GGNs with less convincing generated lesions (classified as "leaning fake"). c A real GGNs in the original LIDC-IDRI dataset.

1. VTT results

Figure 4 presents the combination of the classification results for the four clinicians: of 51 synthetic GGNs, 19 (37%) were classified as real by clinicians, 8/51 (16%) were classified as confidently real, and 11/51 (22%) were classified as leaning real.

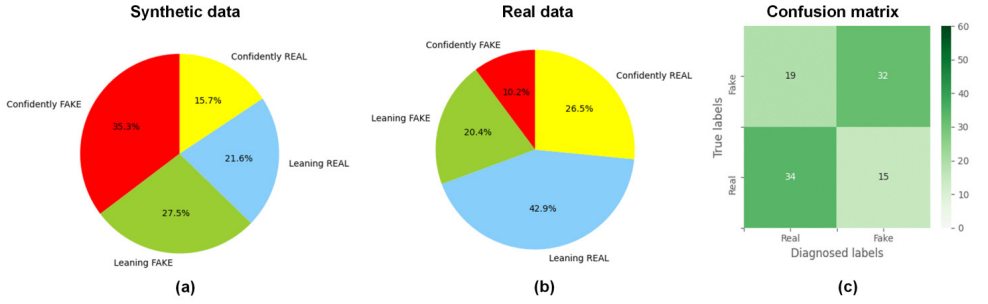


Figure 4. Visual Turing test results. a, b Prediction distribution in synthetic and real ground glass nodules. c Confusion matrix for the prediction.

2. Radiomics

Of a total of 93 features, 58 (62.4%) showed no significant difference ($p \geq 0.052$) between synthetic and real GGNs, and the detailed results are provided in Table 1. Figure 5a shows the comparison of the distribution of radiomic features between real and synthetic GGNs, the histogram shows the counts of specific feature values, and the differences (p-values) in the extracted radiomic features between real and synthetic GGNs were calculated. The receiver operating characteristic curves constructed based on the results of VVT by clinicians and logistic regression model developed by radiomics features are shown in Fig. 5b. We observed a similar classification performance of clinicians (0.68) and radiomics (0.66), with no-significantly different ($p = 0.23$). However, the clinicians achieve significant great performance accuracy around 0.74, better than the 0.62 radiomics accuracy ($p < 0.001$). The clinicians achieve better ability to diagnosis whether the data is synthetic than radiomics.

Table 1. Comparison between real and deep learning-generated radiomic features (p-values according to the Kolmogorov-Smirnov test)

Class	Feature name	p-value
Grey level co-occurrence matrix (GLCM)	Inverse difference moment	0.984
Grey level size zone matrix (GLSZM)	zone percentage	0.935
Grey level dependence matrix (GLDM)	Small dependence emphasis	0.933
Grey level co-occurrence matrix (GLCM)	Inverse difference	0.926
First order	Robust mean absolute deviation	0.903
GLSZM	Small area low grey level emphasis	0.860
Grey level run length matrix (GLRLM)	Run percentage	0.827
GLRLM	high grey level run emphasis	0.729
GLSZM	Grey level non-uniformity normalised	0.697
GLRLM	Long run emphasis	0.676
GLCM	Sum entropy	0.658
GLRLM	Long run high grey level emphasis	0.652
GLRLM	Run entropy	0.652
First order	Entropy	0.643
GLCM	Inverse variance	0.616
GLRLM	Short run high grey level emphasis	0.582
GLDM	high grey level emphasis	0.574
GLCM	Joint energy	0.570
GLCM	Joint entropy	0.570
GLRLM	Run length non-uniformity normalised	0.570
GLRLM	Short run emphasis	0.570
First order	90 percentile	0.541
GLDM	Small dependence low grey level emphasis	0.512

First order	Interquartile range	0.498
GLCM	Inverse difference normalised	0.456
GLDM	Large dependence emphasis	0.450
GLDM	dependence variance	0.445
GLSZM	Low grey level zone emphasis	0.445
First order	Mean absolute deviation	0.414
GLCM	Autocorrelation	0.407
GLDM	Dependence non-uniformity normalised	0.403
First order	Mean	0.389
GLRLM	Run variance	0.375
GLRLM	Grey level non-uniformity normalised	0.324
GLCM	Maximum probability	0.307
Neighbouring grey tone difference matrix (NGTDM)	Strength	0.272
GLCM	Cluster tendency	0.267
GLCM	Inverse difference moment normalised	0.264
GLDM	dependence entropy	0.261
GLRLM	Short run low grey level emphasis	0.227
First order	Minimum	0.212
GLSZM	Large area high grey level emphasis	0.202
First order	Root mean squared	0.186
GLSZM	Large area emphasis	0.178
GLDM	Grey level variance	0.170
GLCM	Joint average	0,160
GLCM	Sum average	0,160
First order	uniformity	0,133
GLDM	Small dependence high grey level emphasis	0,124
GLSZM	Zone variance	0,119
First order	Variance	0,108
GLCM	Sum squares	0,108
GLSZM	High grey level zone emphasis	0,105
GLDM	Large dependence low grey level emphasis	0.082
GLSZM	Size zone non-uniformity normalised	0.074
GLSZM	Small area emphasis	0.073
GLSZM	Large area low grey level emphasis	0.069
GLRLM	Grey level variance	0.066
GLCM	Informational measure of correlation 2	0.052

GLRLM	Low grey level run emphasis	0.045
GLSZM	Small area high grey level emphasis	0.044
GLCM	Cluster prominence	0.022
GLSZM	Grey level variance	0.021
NGTDM	Contrast	0.020
First order	10th percentile	0.015
GLDM	Low grey level emphasis	0.014
GLCM	Difference entropy	0.011
GLSZM	Zone entropy	0.010
GLRLM	Long run low grey level emphasis	0.008
GLCM	Informational measure of correlation 1	0.006
GLCM	Difference average	0.005
GLCM	Maximal correlation coefficient	0.005
GLDM	Large dependence high grey level emphasis	0.003
First order	Maximum	0.002
GLCM	Cluster shade	0.002
First order	Range	0.001
First order	Median	0.000
GLCM	Contrast	0.000
GLDM	Dependence non-uniformity	0.000
GLSZM	Size zone non-uniformity	0.000
NGTDM	Busyness	0.000
GLCM	Correlation	0.000
GLSZM	Grey level non-uniformity	0.000
NGTDM	Complexity	0.000
GLCM	Difference variance	0.000
NGTDM	Coarseness	0.000
First order	Skewness	0.000
First order	Energy	0.000
First order	Total energy	0.000
First order	Kurtosis	0.000
GLRLM	Run length non-uniformity	0.000
GLDM	Grey level non-uniformity	0.000
GLRLM	Grey level non-uniformity	0.000

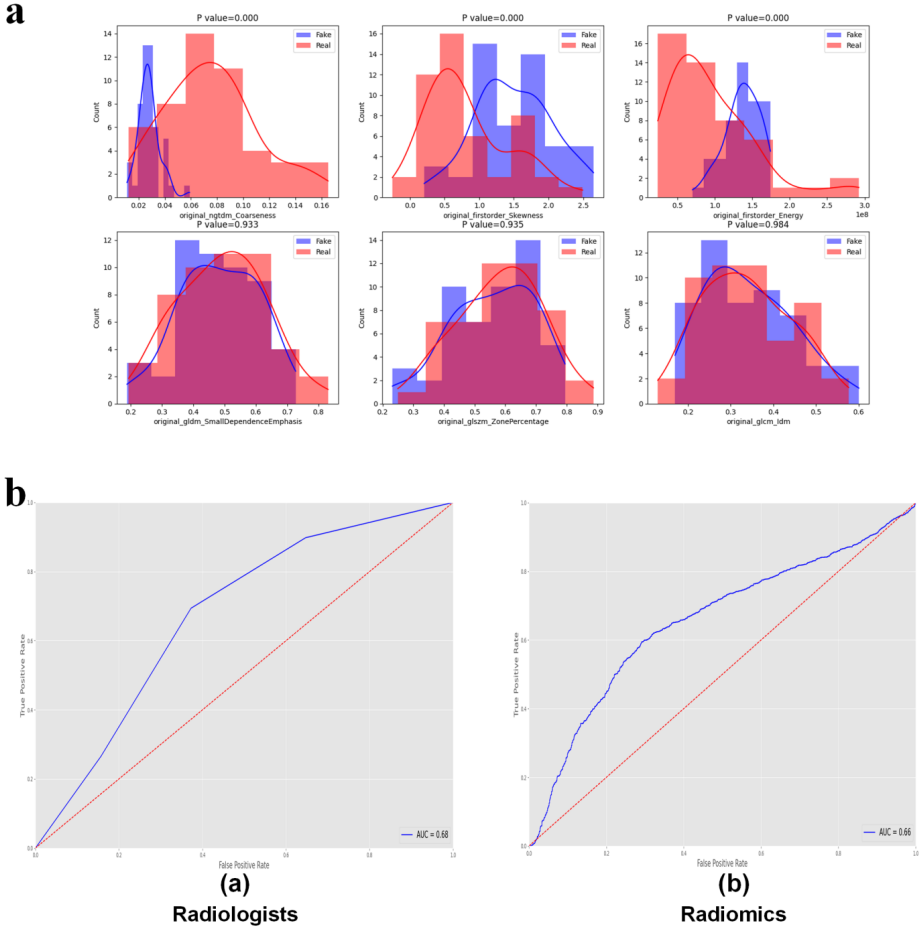


Figure 5. a. Examples for the comparison of radiomics features distribution between real and fake ground glass nodules (GGNs). The comparison of radiomics features distribution extracted from synthetic and real images with minimum three p-values shows in the upper row. The comparison of radiomics features distribution extracted from synthetic and real images with maximum three p-values shows in the lower row. **b., c.** Receiver operating characteristic curve of the prediction of distinguishing real and fake GGNs. by radiologists (**a**) and by the logistic regression model (**b**).

3. DL classification network

The results of the DL classification network trained using decreasing portions of the dataset are shown in Fig.6. When the dataset is 90%, the precision (i.e., positive predictive value) was similar between the two groups. How-ever, when the dataset decreased to 50%, the performance of the real data only group significantly decreased. On the other hand, synthetic GGNs can increase precision in training the DL network. When the sample decreased to 10%, the real data has better performance than by adding synthetic data. From Fig.6b, the

recall (i.e., sensitivity) of GGN was decreasing when decreasing the dataset both in real data only and real data with GAN groups. However, in most cases, models trained on datasets with synthetic data performed better than models without the addition of synthetic data.

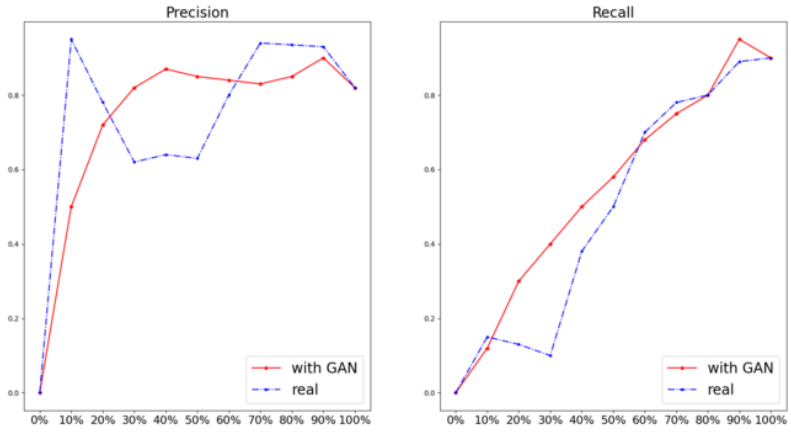


Figure 6. Comparison precision (i.e., positive predictive value) and recall (i.e., sensitivity) between real and added synthetic dataset in different percentages of the training set. The blue and the red lines present the performance of the deep learning classification model trained by real data and the real data plus synthetic data, respectively. The horizontal axis label is the percentage of training data in the dataset. The vertical axis label is the score of precision and the recall with the range from 0 to 1.

Discussion

In the present study, we applied a GAN-based model with double discriminators to generate GGN in low-dose CT scans. We benchmarked the performance of the model using a qualitative (VTT with clinicians) and a quantitative approach (radiomics).

To our knowledge, only one previous study proposed the use of GANs to generate lung lesions and performed a VTT [18], which showed that 67% and 100% of the fake nodules were marked as real by two radiologists, respectively. Differences exist between this study and our study: in the VVT of the cited study [16], the radiologists reviewed the generated lesions, but the surrounding tis-sues or the entire lungs were not included in the field of view. Moreover, the surrounding tissues and the lung background that has relationship with nodules were not considered when training and generating the nodules. Conversely, we generated GGNs from the whole lung to use the anatomical dependence with the background tis-sue [19]. However, the relatively small size of our study compared to the previous research [18] probably influenced the results of the visual Turing test.

Based on our VTT evaluation, we have shown that GAN-generated lung lesions have the potential to be very consistent with real lesions. This gives us the opportunity to use GAN-generated data to solve real-world problems, such as using the generated data to train and test junior doctors, especially for hospitals that do not have large cohort datasets, long-

time established picture archiving and communication systems, as privacy-preserving synthetic open datasets for research purposes.

More than half of the radiomic features were not statistically different between DL-generated and real nodules, proving that the generated GGNs are acquiring or learning detailed features from the real sample. Furthermore, these consistent radiomic features cover all classes, which could support the conclusion that the proposed approach mimics different aspects of real nodules. Conversely, one third of the features in this study showed significant differences in the distribution between the generated and real GGNs. Based on the radiomics results and the clinicians' opinion, we think that the low complexity of the generated GGNs is the main reason for the discrepancy between the generated and real GGNs. For example, the p-value of the radiomic features coarseness (which can measure the spatial change rate) and complexity (which can measure the non-uniformity of local grey levels) between real and synthetic GGNs are close to 0, supporting our hypothesis. We hypothesize the following explanations: (i) the data source is derived from public databases that have low resolution and lots of noise, and (ii) we did not optimize the training process by specifically including radiomics features in the loss function.

Based on the radiomics results, we built a "radiomics physician" to discriminate between real and generated GGNs, which interestingly is generally consistent with the discriminatory ability of real physicians. It is worth noting that the "radiomics physician" model was trained based on a sample of 100 cases, and the physicians have more than 5 years of experience. Overall, it is a challenging task to discriminate between real and generated GGNs for "radiomics physicians" and real physicians.

Finally, we wanted to test how data augmentation with GAN will affect the detection accuracy of a CAD system. Figure 6 shows that adding synthetic GGNs to the original dataset improves the performance of our DL CAD system. However, there was no significant contribution when the size of the training dataset is under 10% and over 70% of the original sample size. We hypothesize that when the training data is under 10%, there is an insufficient number of samples to train the GAN. A GAN trained on only a few samples cannot synthesize the rich diversity and complexity of real GGNs. Based on the results (Fig. 6), we conclude that the performance of the DL model increases with the sample size in certain ranges of real data samples. However, as shown in Fig. 6, the performance of the DL model cannot be improved after a threshold value larger than the sample size, which is the plateau of the model. Specifically, for effective dataset size to train a GAN, around 50% of training data which include around 100 samples of GGN has the biggest increase in accuracy of the classification model when synthetic GGN are added. Overall, from our experiment, we found that:

1. Synthetic data has the ability to increase the performance of a DL model unless only a few training samples can be used;

2. From the perspective of cost and effectiveness, around 100 samples are sufficient to develop a GAN model that can generate realistic GGNs to significant improve the performance of the detection GGN model.

This study has some limitations. First, we used a public dataset for training the model, but we want to extend the work to other datasets. In future studies, we will add high-resolution data from our center for model enhancement. Second, we only focused on GGNs, because

of their lower incidence compared to other types of nodules. However, the dimension and density variation of the included GGNs is limited, which has the potential risk of obtaining optimistic radiomic assessment results. We will perform transfer learning to generate lung nodules and tumours in the future based on the model in this study. Furthermore, the diagnosis of malignant GGN is a challenging task for clinical practice. However, in this study, we did not generate benign or malignant GGN. To address this issue, we are collecting data from the real world with follow-up endpoints and trying to generate qualitative GGN, especially malignant GGN.

Third, we generated only two-dimensional samples. However, generating three-dimensional (3D) images is costly for model training, first, because 3D GANs have a larger number of parameters which need more training data and also have a significantly higher requirement in hardware when the input data has large scale such as CT images. In the future work, we will consider the model compression to decrease the requirement of hardware and the size of dataset for training the 3D GAN. We tried to perform our visual Turing tests by getting closer as much as possible to a real clinical scenario. Nevertheless, it was out of the scope of this study to integrate our DL models within the clinical workstations available to our radiologists. As proof-of-concept, we proposed to our radiologists the generated and real pulmonary nodules as two-dimensional axial CT images in the standard lung window. Future work will include the production of the generated nodules in standard DICOM formats in all the 3D projections. We are also investigating the possibility to invest in the development of a cloud-based platform to homogenize visual Turing tests for similar experiments. In addition, we did not evaluate the morphological features between the generated and real GGNs.

Fourth, we have not discussed the trend of data requirement for different task, such as what happens when the quality of data is decreased, how many data points need to be added when the target size us increased, and whether different sources such as CT and magnetic resonance imaging influence the dataset requirements. In the future work, we will design experiments to figure out the connection between the data requirement and different tasks.

Fifth, according to the results of the radiomics part, there are still considerable differences between the real and generated GGO, and more than one third of the radiomic feature values were different, which may be a reflection that the GAN method proposed in this study is not optimal. Based on this result, there is still much potential for improvement of our algorithm, with a particular focus on improving the level of complexity of the textures.

Sixth, we did not conduct interobserver and intra-observer testing and the degree of disagreement between different readers was not assessed. On the other hand, in our experience, the differences between the readers (physicians) included in this study were limited to the same broad category, i.e., real or fake. For example, nodules labelled as “confidently real” by one physician have the possibility of being labelled as “leaning real” instead of “confidently/leaning fake” by other physicians.

Finally, despite the GANs are an elegant data generation mechanism gaining more and more popularity in the medical field, most of them still present a high level of complexity compared for example to traditional DL algorithms such as convolutional neural networks. For example, there is no consensus on the most appropriate metric to be used to stop the training at the best point (global minimum of the loss function). This will sometimes lead to a

not satisfactory quality of the generated data. Especially when dealing with medical images, the risk of introducing novel, undesired artefacts, and blurry images is not negligible.

In conclusion, in this study, we used GANs to generate GGN and validated these by four physicians and radiomics approaches, showing that GAN methods have great potential for augmentation of the original dataset.

References

1. Zhou XR (2020) Automatic segmentation of multiple organs on 3D CT images by using deep learning approaches. Adv Exp Med Biol 1213:135–147. https://doi.org/10.1007/978-3-030-33128-3_9

2. Mastouri R, Khlifa N, Neji H, Hantous-Zannad S (2020) Deep learningbased CAD schemes for the detection and classification of lung nodules from CT images: a survey. J Xray Sci Technol 28:591–617. https://doi.org/10.3233/XST-200660

3. Setio AAA, Traverso A, de Bel T et al (2017) Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medi Image Analysis 42:1–13. https://doi.org/10.1016/j.media.2017.06.015

4. Kaggle Data Science Bowl (2017). https://www.kaggle.com/c/data-scien cebowl-2017.

5. Bowles C, Chen L, Guerrero R, et al (2018) Gan augmentation: augmenting training data using generative adversarial networks. .arXiv preprint arXiv:1810.10863, 2018.

6. Migliore M, Fornito M, Palazzolo M et al (2018) Ground glass opacities management in the lung cancer screening era. Ann Transl Med 6(5):90. https://doi.org/10.21037/atm.2017.07.28.

7. Zhang H, Hu X, Ma D, Wang R, Xie X (2022) Insufficient data generative model for pipeline network leak detection using generative adversarial networks. IEEE Trans Cybern 52(7):7107–7120. https://doi.org/10.1109/

8. Bera S, Biswas PK (2021) Noise conscious training of non local neural network powered by self attentive spectral normalized markovian patch GAN for low dose CT denoising. IEEE Trans Med Imaging 40(12):36633673. https://doi.org/10.1109/tmi.2021.3094525

9. Do WJ, Seo S, Han Y, Ye JC, Choi SH, Park SH (2020) Reconstruction of multicontrast MR images through deep learning. Med Phys 47(3):983–997. https://doi.org/10.1002/mp.14006

10. Jiang Y, Chen H, Loew M, Ko H (2021) COVID-19 CT image synthesis with a conditional generative adversarial network. IEEE J Biomed Health Inform 25(2):441–452. https://doi.org/10.1109/jbhi.2020.3042523

11. Armato SG 3rd, McLennan G, Bidaut L et al (2011) The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 38(2):915–931. https://doi.org/10.1118/1.3528204

12. Ledig C, Theis L, Huszár F et al (2017) Photo-realistic single image superresolution using a generative adversarial network. Proc IEEE Conf Comput Vis Pattern Recognit 2017:4681–4690

13. He K, Zhang X, Ren J (2016) Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778

14. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen XJA inips (2016) Improved techniques for training gans. arXiv:1606.03498v1

15. de Farias EC, di Noia C, Han C, Sala E, Castelli M, Rundo L (2021) Impact of GAN-based lesion-focused medical image super-resolution on the robustness of radiomic features. Sci Rep 11(1):21361. https://doi.org/10. 1038/s41598-021-00898-z

16. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14(12):749–762. https://doi.org/10.1038/nrclinonc.2017.141

17. Tixier F, Jaouen V, Hognon C, Gallinato O, Colin T, Visvikis D (2021) Evaluation of conventional and deep learning based image harmonization methods in radiomics studies. Phys Med Biol 66 (24). https://doi.org/10. 1088/1361-6560/ac39e5

18. Chuquicusma MJ, Hussein S, Burt J, Bagci U (2018) How to fool radiologists with generative adversarial networks? A visual Turing test for lung cancer diagnosis. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE, pp 240–244. https://ieeexplore.ieee. org/document/8363564

19. Xu Z, Wang X, Shin H-C, Roth H, Yang D, Milletari F, Zhang L, Xu D (2019) Tunable CT lung nodule synthesis conditioned on background image and semantic features. Simulation and Synthesis in Medical Imaging. Springer International Publishing, Cham, pp 62–70

Chapter 4: A PET/CT Radiomics Model for Predicting Distant metastasis in Early-Stage Non–Small Cell Lung Cancer Patients Treated with Stereotactic Body Radiotherapy: A Multicentric Study

*Adapted from: **Zhen Zhang***, Lu Yu*, et al. A PET/CT Radiomics Model for Predicting Distant metastasis in Early-Stage Non–Small Cell Lung Cancer Patients Treated with Stereotactic Body Radiotherapy: A Multicentric Study. (Submitted, under review)*

** indicates equal contributions*

Chapter 5: Combining tumor radiomics features and whole-lung radiomics features to predict prognosis in locally advanced non-small cell lung cancer treated with curative radiotherapy

Adapted from: ***Zhen Zhang****, Meng Yan*, et al. Combining tumor radiomics features and whole-lung radiomics features to predict prognosis in locally advanced non-small cell lung cancer treated with curative radiotherapy (In preparation)*

** indicates equal contributions*

Chapter 6: Radiomics and dosiomics signature from whole lung predicts radiation pneumonitis: a model development study with prospective external validation and decision-curve analysis

Adapted from: ***Zhen Zhang****; Zhixiang Wang; Meng Yan; Jiaqi Yu; Andre Dekker; Lujun Zhao; Leonard Wee. Radiomics and Dosiomics Signature from Whole Lung Predicts Radiation Pneumonitis: A Model Development Study with Prospective External Validation and Decision-Curve Analysis. International Journal of Radiation Oncology, Biology, Physics 2022, S0360-3016(22)03189-3. https://doi.org/10.1016/j.ijrobp.2022.08.047*

Abstract

Purpose Radiation pneumonitis (RP) is one of the common side effects of radiotherapy in the thoracic region. Radiomics and dosiomics quantifies information implicit within medical images and radiotherapy dose distributions. In this study we demonstrated the prognostic potential of radiomics, dosiomics, and clinical features for RP prediction.

Materials and methods Radiomics, dosiomics, dose-volume histogram (DVH) metrics, and clinical parameters were obtained on 314 retrospectively-collected and 35 prospectively-enrolled patients diagnosed with lung cancer between 2013 to 2019. A radiomics risk score (R-score) and dosiomics risk score (D-score) and DVH-score were calculated based on logistic regression after feature selection. Six models were built using different combinations of R-score, D-score, DVH-score, and clinical parameters to evaluate their added prognostic power. Over-optimism was evaluated by bootstrap resampling from the training set, and the prospectively-collected cohort was used as the external test set. Model calibration and decision-curve characteristics of the best-performing models were evaluated. For ease of further evaluation, nomograms were constructed for selected models.

Results A model built by integrating all of R-score, D-score, and clinical parameters had the best discriminative ability with area under the curves (AUCs) of 0.793 (95%CI 0.735-0.851), 0.774 (95%CI 0.762-0.786), and 0.855 (95%CI 0.719-0.990) in the training set, bootstrapping set, and external test set, respectively. The calibration curve image showed good agreement between the predicted and actual values with a slope of 1.21 and an intercept of -0.04. The decision curve image showed positive net benefit for the final model based on the nomogram.

Conclusion Radiomics and dosiomics features have potential to assist with the prediction of RP, and the combination of radiomics, dosiomics, and clinical parameters led to the best prognostic model in the present study.

Keywords: Radiomics; Dosiomics; Lung cancer; Radiation Pneumonitis

Introduction

Radiotherapy (RT) plays a crucial role in the management of lung cancer (LC) [1], especially for locally advanced and unresectable cases [2, 3]. Advances in thoracic RT have led to steadily improving prognosis for LC patients, but RT-related side effects remain a treatment-limiting concern [4-6]. Radiation pneumonitis (RP) is a common adverse effect that degrades patients' quality of life and can be fatal in severe cases. To date, there is no highly effective cure for RP [7], thus prevention of RP remains one of the top clinical priorities during RT dose planning [8, 9]. Robust and reproducible predictive models that could estimate the risk of developing RP after lung RT would be of immense clinical value. Such estimates could be incorporated into treatment planning and informed shared decision-making consultations (such as a choice between starting prophylactic medication or active vigilance).

Studies to date suggest a number of clinical factors, such as smoking status, pre-existing lung disease [10], pre-existing cardiac disease [11], and chemotherapy [12], may affect an individual's pre-disposition to develop RP. Although these parameters may indicate towards susceptibility, RP remains a disease exhibiting strong inter-person variability [13]; this heterogeneity does not appear to be sufficiently well represented in conventional clinical factors. Single-nucleotide polymorphism (SNPs) [14] and plasma cytokines [15, 16] can also be indicative of heterogeneity, and several studies have revealed significant associations between SNPs and the occurrence of RP [17], which suggests the feasibility of genetic and molecular biomarkers. However, some biomarkers may be subject to vagaries of limited spatial sampling and are only available through invasive means.

Radiomics is the high-throughput extraction of quantitative handcrafted features from medical images. Image-based radiomics has the potential to characterize heterogeneity within the entire pre-RT lung parenchyma and, in the case where suitable repeated imaging could be available, to be able to quantify parenchymal changes during a course of RT in a non-invasive manner. It has been demonstrated that radiomics features are associated with genetic heterogeneity (radiogenomics) [18]. There have been several studies that demonstrate the potential of radiomics to predict RP [19-21], but building predictive models only from an image perspective may not be sufficient. Physicians routinely modify treatment strategies based on the patient's condition. For example, some patients with pre-existing lung disease diagnosed by imaging may be prescribed a relatively low dose thereby reducing the chance of developing RP and weakening the predictive power of radiomics. Therefore, there is a need to incorporate prescription dose information into predictive model.

In a different context, the occurrence of RP is strongly related to RT dose, and therefore a number of studies have used dose-volume histogram (DVH) metrics, such as mean lung dose (MLD) [22] and volume of the lung receiving 20 Gy (V20) [23], to predict RP. DVH parameters are not able to fully describe the immense spatial heterogeneity of dose distribution, which may be realized through intensity modulated radiation delivery (i.e., IMRT and/or VMAT) [23, 24]. Dosiomics, conceived as using radiomics tools to characterize spatial heterogeneity of RT dose (as opposed to image voxel intensities) provides a greater depth of information in contrast to traditional DVH measures [25, 26].

Previous works [19-21, 25, 26] attempt to predict RP solely on the basis of medical (tomographic) imaging alone, or on the basis of dose information, and those results show that it

is highly unlikely to be clinically sufficient by relying exclusively on either the imaging features (radiomics) or the dose-volume parameters. There is a lack of studies combining radiomics and dosiomics to predict RP in lung cancer, furthermore, studies using rigorous and rational steps of selecting handcrafted features are needed. A large sample-based, prospective study is also required to assess the objective predictive power of the models.

In this study, we extracted radiomics features in RT planning CT and dosiomics features in 3D dose grids from the RT treatment planning system (TPS) and performed objective and rigorous feature selection. We evaluated the performance of clinical parameters, radiomics features, dosiomics features, and DVH metrics, singly as well as in combination, to predict RP after RT to the chest area. We evaluated the prediction models in terms of discriminative performance and model calibration using a prospectively collected dataset. Moreover, decision-curve analysis was used to investigate the potential clinical relevance of such models if implemented in routine practice. A nomogram was provided to facilitate future independent validation of our work in other clinical settings.

Methods

1. Study design

This study was designed as a Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) type 3 study comprising model development and independent validation [27]. This study was registered on artificial intelligence in biomedical research platform (AIMe, ID: mn9jLf) [28]. The overarching study flow is illustrated in Figure 1.

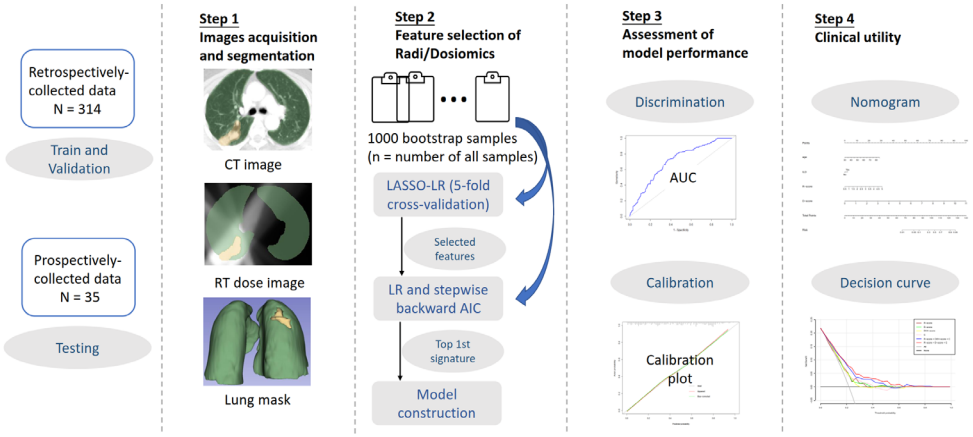


Figure 1. Analysis flowchart. Step 1, The radiomics and dosiomics features of the lung tissue region were extracted. Step 2, 1000 unique bootstrap samples were taken from all samples, features were selected by correlation, least absolute shrinkage (LASSO) embedded with logistic regression (LR) and Akaike information criterion (AIC) for modeling. Step 3, The model performance was evaluated using discrimination and calibration. Step 4, Clinical applications were evaluated using nomogram and decision curves.

2. Patients

A single-institutional model development cohort of 314 subjects was retrospectively extracted from institutional records after ethics board approval (IRB/bc2021135), comprising patients diagnosed with LC and treated with radical (chemo)-RT, with either IMRT or VMAT techniques, at Anonymized for Review Hospital between January 2013 and December 2018. For model validation, an additional 35 patients with the same criteria were prospectively enrolled with informed consent and same ethics approval (IRB/bc2021135), who were treated between October 2018 and March 2019 in the same institution. Detailed inclusion and exclusion criteria have been specified in the Supplementary Materials A.

3. Image acquisition and treatment planning

Intravenous contrast-enhanced planning CT scans were acquired on a single Brilliant (Philips Medical Systems; Best, The Netherlands) multislice scanner with a standardized protocol: 120 kVp, 100 mAs, 3 mm slice thickness, 512 x 512 image matrix, 50 cm fields of view, 0.977 mm pixel spacing and vendor's default convolution kernel. Experienced radiation oncologists delineated the LC gross tumor volume (GTV) and malignant lymph nodes in the Pinnacle TPS (Philips Radiation Oncology Systems; Fitchburg, Wisconsin, United States), with image fusion against complementary imaging studies whenever available (such as positron emission tomography).

The GTV was isotropically expanded by 5 mm, as well as subclinical microscopic malignant lesions to derive the clinical target volume (CTV). The planning target volume (PTV) was an additional 5 mm isotropic expansion around the CTV. Dosimetrist were instructed to cover at least 95% of the PTV with the prescribed RT dose. Delineations conformed to the guidelines set by the Radiotherapy and Oncology Group (RTOG). The relevant dose constraints were as follows: MLD < 20 Gy, V20 < 30%, and volume of the lung receiving 5 Gy (V5) < 60%. All patients were nominally prescribed 2 Gy per fraction once daily. Radiation oncologists determined the total prescribed dose based on each patient's overall physical condition and best achievable normal tissue constraints. The actual total RT dose delivered ranged between 50 to 70 Gy. The dose grid resolution is 4 mm, and the dose calculation algorithm is Collapsed Cone Convolution [29, 30]. The planning CT series with associated RT structure delineations and RT planned radiotherapy 3D dose grids were exported from Pinnacle in the standard DICOM format.

4. Lung segmentation and RP grading

We extracted radiomics features and dosiomics features from the region corresponding to total (left plus right) lung. To ensure consistency of lung segmentation, we quality assured the lung delineations for each subject using a deep-learning automatic lung contouring tool based on retraining of the published model. The original and automatically generated lung outlines were inspected and then manually edited by a single experienced radiation oncologist (author MY). Two other radiation oncologists (author JQY and ZZ) subsequently independently reviewed the lung organ segmentation, and any disputes were resolved by direct consultation among all three authors.

The primary outcome RP was defined, in accordance with the Common Terminology Criteria for Adverse Events (CTCAE) v5.0, as symptomatic radiation pneumonitis of CTCAE grade 2 or higher within 6 months after the end of RT [12, 16]. Monitoring of RP was based

on the combination of clinical examination, reported symptoms, outpatient medical records, laboratory tests, chest X-ray, and visual inspection of follow-up CTs, which were all performed at intervals of 1, 3, and 6 months after completion of RT, and then every 6 months thereafter.

5. Radiomics and dosiomics features extraction

A total of 103 handcrafted radiomics features were extracted from DICOM CT and RT Structures using the "O-RAW" package [31] (based on Pyradiomics v3.7 [32]). These features comprised 17 intensity histogram features, 13 morphological (shape) features, and 73 textural features. No digital image filters were applied during pre-processing. Most of the hand-crafted features conformed to the Image Biomarker Standardization Initiative (IBSI) [33]; specific divergences from the IBSI at the time of writing have been reported according to the PyRadiomics documentation. Radiomics extraction settings are the same as for a previous publication [31], and our PyRadiomics parameters setting file has been provided in the Supplementary Materials B. For dosiomics features, DICOM RT Dose files were first converted as NRRD images using 3D Slicer [34], and then the same feature extraction procedure in PyRadiomics was applied for the total lung region. Additionally, voxel-wise values in the "dose images" were scaled to represent the absolute physical dose in units of Gray (Gy). Isotropic spatial resampling (1 mm) was applied on the CT images and dose images prior to feature extraction as recommended by previous studies [35].

6. Feature selection

An overview of multi-step feature selection and model construction is given in Figure 1. The clinical parameters for modeling were evaluated by using univariate and multivariate analyses for twelve clinical parameters with predictive potential. Feature selection for the radiomics model and the dosiomics model were performed separately, and has been adapted from the feature pooling and signature pooling method used by Compter et al. [36]. In brief, the selection process was as follows:

(i) A thousand unique bootstrap samples (with replacement) were drawn from the whole training cohort. Within each bootstrap sample, we first minimized the number of strong pairwise normalized (Z-score, (original value-mean value)/standard deviation) feature correlations greater than 0.90 or less than -0.90. A least absolute shrinkage (LASSO) loop with 20-times repeated 5-fold cross-validation embedded with a logistic regression (LR) supervised classifier was used to select features. From each of the 1000 bootstraps, we ranked each individual feature according to how frequently it was retained by the LASSO-LR.

(ii) We arbitrarily selected some of the top most frequently-appearing individual features from the above table. From this small subset of selected features, we built a multivariable LR model on each of the same aforementioned bootstraps samples with stepwise backwards elimination using the Akaike information criterion (AIC) as metric. From each of these 1000 bootstraps, we tabulated how many times each combination of one or more features (i.e., potential signatures) was retained by the stepwise LR.

(iii) We arbitrarily selected the top most frequently-appearing signature arbitrarily selected to build the final multivariable LR model. The coefficients of the final model were fitted using the original non-bootstrapped development cohort.

7. Model construction

The clinical model was presented as a multivariable LR model. To this, we added an aggregated Radiomics Risk Score (R-score) and an aggregated Dosiomics Risk Score (D-score), separately. The R-score was defined as the linear predictor (LP) of the multivariable LR radiomics model, and likewise the D-score was defined as the LP of the multivariable dosiomics model. For combined models, we assessed the combinations of the clinical factors together with either, or both, of the R-score and D-score.

V20 and mean lung dose (MLD) were used to build DVH model, and details of feature selection and model construction are provided in the Supplementary Materials C. To address the issue of imbalanced data, we performed the Synthetic Minority Oversampling Technique (SMOTE) approach in the training set. We also examined the Pearson correlation between the R-score and clinical parameters, and between the D-score and dose-volume histogram metrics (dosimetrics).

8. Model validation – internal and external

We estimated the over-optimism in model development using the method recommended in the TRIPOD guidelines; for each of the 1000 abovementioned pre-defined bootstraps, we fitted the LR model coefficients on each bootstrap, and then computed its Area under the curve (AUC) of receiver operating characteristic curve (ROC) using the original non-bootstrapped development cohort. From these 1000 bootstraps, we computed the average AUC and its 95% confidence interval.

As external validation, we evaluated the aforementioned models using the prospectively-registered cohort of 35 subjects. Processing of these 35 subjects followed exactly the same procedure as for the model development cohort, and none of these subjects were used in any way during model construction.

The well-established calibration curve technique was used to assess model goodness of fit (i.e., the extent of concordance between the predicted and observed values) again using a bootstrap of 1000 repetitions. To facilitate clinical use and support fully independent validation of our model, a simple nomogram was generated for the R-score, D-score, and the selected clinical parameters. Lastly, we tried to discuss the potential clinical utility of our model using decision curve analysis (DCA) [37].

9. Statistical analyses

Baseline patient characteristics for continuous variables are presented as mean ± standard deviation. For univariate ranking of clinical predictors, Pearson X2 tests and exact Fisher tests were used for categorical variables and logistic regression for continuous variables. For significance of clinical factors, a two-sided hypothesis test at the $a = 0.05$ confidence level was assumed. Significant characteristics were subsequently combined in multivariable logistic regression.

All data had been collated and standardized using the Statistical Package for Social Science program (SPSS for Windows, version 27.0; SPSS Inc, Chicago, IL). Feature selection, model construction, model performance assessment and decision-curve analysis were all performed in R software (version 4.0.5).

Results

1. Patient characteristics and incidence of RP

The case mix of patients and treatments studied in this model are reported in Table 1. Univariate analysis showed statistically significant differences in interstitial lung disease (ILD), concurrent chemoradiotherapy (CCRT), and age between patients with and without RP. The overall incidence of CTCAE grade 2 or higher for RP was 21.5% (75 of 349), 21% (66 of 314) in the retrospective data set, and 25.7% (9 of 35) in the prospective validation set. Multivariable analysis indicated that ILD (OR 2.471; 95%CI 1.037-5.888, p = 0.041) and age (OR 1.051; 95%CI 1.012-1.085, p =0.008) were independent factors associated with RP. A forest plot for the coefficients in the multivariable LR model is shown in Figure 2.

Table 1 Patient Characteristics

Characteristics	All retro pts	Without RP2	With RP2	P*	Pros pts
	n (%)	Mean ± SD	Mean ± SD		n (%)
Age median	61 (30-85)	61 (30-85)	63 (44-79)	0.005	62 (34-75)
Gender				0.523	
Male	238 (75.8%)	186 (78.2%)	52 (21.8%)		23 (65.7%)
Female	76 (24.2%)	62 (81.6%)	14 (18.4%)		12 (34.3%)
Smoking				0.569	
Yes	244 (77.7%)	191 (78.3%)	53 (21.7%)		26 (74.3%)
No	70 (22.3%)	57 (81.4%)	13 (18.6%)		9 (25.7%)
KPS				0.725	
≤80	132 (42.0%)	103 (78.0%)	29 (22.0%)		13 (37.1%)
>80	182 (58.0%)	145 (79.7%)	37l (20.3%)		22 (62.9%)
Diabetes				0.609	
Yes	34 (10.8%)	28 (82.4%)	6 (17.6%)		2 (5.7%)
No	280 (89.2%)	220 (78.6%)	60 (21.4%)		33 (94.3%)
ILD				0.015	
Yes	25 (8.0%)	15 (60.0%)	10 (40.0%)		9 (25.7%)
No	289 (92.0%)	233 (80.6%)	56 (19.4%)		26 (74.3%)
Pathology				0.656	
LUSC	86 (27.4%)	65 (75.6%)	21 (24.4%)		8 (22.9%)
LUAD	73 (23.2%)	59 (80.8%)	14 (19.2%)		10 (28.6%)
SCLC	155 (49.4%)	124 (80.0%)	31 (20.0%)		17 (48.5%)
Inducchemo				0.739	
Yes	287 (91.4%)	226 (78.7%)	61 (21.3%)		31 (88.6%)
No	27 (8.6%)	22 (81.5%)	5 (18.5%)		4 (11.4%)
CCRT				0.047	

Yes	93 (29.6%)	168 (76.0%)	53 (24.0%)		8 (22.9%)
No	221 (70.4%)	80 (86.0%)	13 (14.0%)		27 (77.1%)
Conso chemo				0.116	
Yes	179 (57.0%)	147 (82.1%)	32 (17.9%)		19 (54.3%)
No	135 (43.0%)	101 (74.8%)	34 (25.2%)		16 (45.7%)
PGTV(Gy)	59.274±2.977	59.204±3.063	59.539±2.634	0.415	60.200±2.870
Smoking index	661.540±571.430	641.840±550.543	735.600±643.084	0.237	668.600±550.412

Abbreviations: Retro = retrospective; Pts = patients; Pros = prospective; LUSC = lung squamous cell carcinoma; LUAD = lung adenocarcinoma; SCLC = small cell lung cancer; IMRT = intensity-modulated radiotherapy; VMAT = volumetric modulated arc therapy; chemo = chemotherapy; KPS = Karnofsky performance score; Induc chemo = induction chemotherapy; CCRT = concurrent chemoradiotherapy; Conso chemo = consolidation chemotherapy; PGTV = planning gross tumor volume.

*The differences in characteristics were evaluated by logistic regression for continuous variables or Pearson X2 test and exact Fisher test for categorical variables

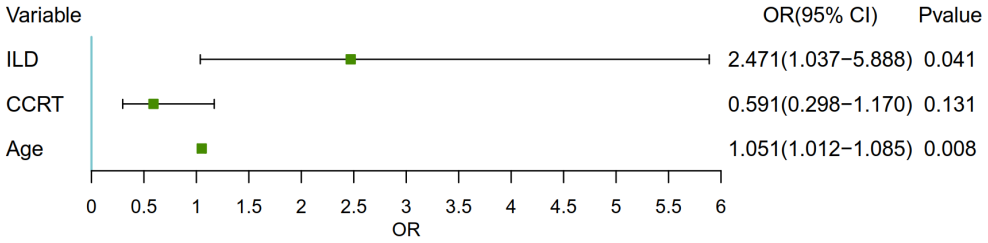


Figure 2. Multivariate analysis forest plot by logistic regression. Characteristics with statistically significant univariate analysis were subjected to multivariate analysis, with ILD and age as independent predictors of RP. *Abbreviations: OR = Odds ratio; ILD = Interstitial lung disease; CCRT = Concurrent chemoradiotherapy.*

2. Feature selection and risk scores

By inspecting the frequency ranking of individual features, we noted that a threshold frequency of around 600 yielded us 11 radiomics features and 12 dosiomics features. Subsequently, we derived a final radiomics signature comprising of 7 features for the R-score, and a final dosiomics model of 6 features for the R-score. Detailed tables and graphs from the feature selection process, along with the names and definitions of the selected features, are provided in the Supplementary Materials D.

The R-score and the D-score were calculated based on the coefficients weighted by LR. The formula of R-score and D-score are provided in the Supplementary Materials D. For ease of computing the R-score and D-score, a simple calculator has been provided and can be found here: only for Windows or MacOS operating systems, (https:// https://github.com/Radiologyzz/Calculator.git). Instructions for using the calculator are given in the Supplementary Materials E.

Examples of low and high R-score and D-score are given in Figure 3. In this example, ILD was evident in the patient with high R-score. The lung tissue of the patient with high D-score received higher dose of radiation than the patient with low D-score (the same prescription dose for both patients). The results showed no significant correlation (>0.8) by Spearman's analysis between R-score and clinical parameters, D-score and dosimetrics, respectively (Supplementary material F Figure 3). However, there were slight differences in the distribution of R-score for the population with and without ILD, and more noticeable differences in the distribution of D-score for the population with different MLD (Supplementary material F Figure 4).

3. Comparison of discrimination performance of different models

Prediction performance was quantified as AUC for six models and is summarized in Table 2. Other possible combinations of models are provided in the Supplementary material G. The model that yielded the highest AUC was the combination of R-score, D-score, and clinical parameters. The discrimination performances were 0.793 (95%CI 0.735-0.851) and 0.855 (95%CI 0.719-0.99), in the training and prospective validation sets, respectively. As the estimate of the degree of over-optimism (i.e., over-fitting) during model construction, our bootstrap-based validation yielded an AUC of 0.774 (95%CI 0.762-0.786).

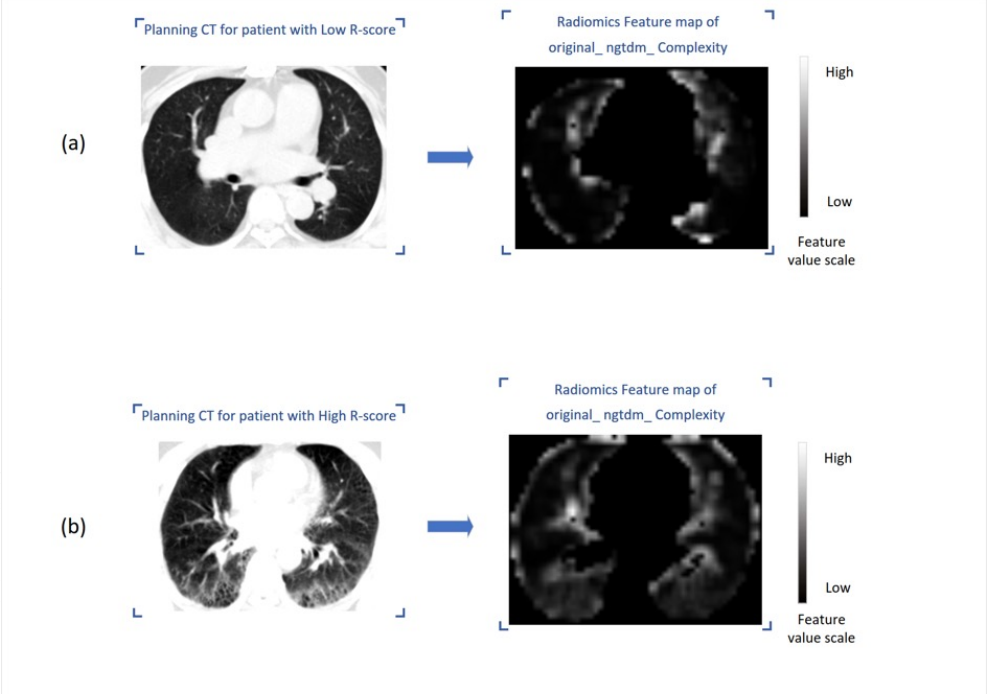
Table 2 Discrimination ability of different models according to area under the curve (AUC) with 95%CI provided between parentheses.

Model	Train (95%CI)	Validation by bootstrapping (95%CI)	Testing (95%CI)
R-score	0.676 (0.606-0.745)	0.619 (0.592-0.646)	0.671 (0.558-0.899)
D-score	0.728 (0.665-0.790)	0.687 (0.667-0.706)	0.684 (0.573-0.883)
DVH-score	0.637 (0.570-0.705)	0.628 (0.613-0.642)	0.661 (0.551-0.856)
Clinical parameters	0.664 (0.594-0.735)	0.654 (0.628-0.680)	0.709 (0.509-0.91)
R-score + DVH-score + C	0.728 (0.674-0.803)	0.719 (0.703-0.736)	0.782 (0.686-0.832)
R-score + D-score + C	0.793 (0.735-0.851)	0.774 (0.762-0.786)	0.855 (0.719-0.990)

Abbreviations: R = radiomics risk score; D = dosiomics risk score; DVH = dose-volume histogram; C = clinical parameters.

4. Model calibration and decision curve analysis

A nomogram based on clinical parameters, R-score, and D-score was constructed and is shown in Figure 4a. The calibration curve of nomogram validated by bootstrap resampling is displayed in Figure 4b, which illustrates good agreement between the predicted probabilities of RP versus the actual observed probabilities. The calibration curve of prospective validation set is provided in the Supplementary Material H with a slope of 1.21 and an intercept of - 0.04. DCA (Figure 4c) showed that the prediction model with the combination of R-score, D-score and clinical parameters has the best positive net benefits at threshold probabilities, implying that a proportion of patients could benefit from using the model to assist in clinical decision making.



nosis and reduces quality of life. Patients with RP are a highly heterogeneous group, hence this study evaluated non-invasive methods (radiomics and dosiomics) using only pre-treatment information to characterize individual differences. In this study, the dosiomics features were shown to have stronger predictive power than the conventional DVH parameters, and the combination of a radiomics signature, a dosiomics signature, and two clinical factors were found to be predictive of RP. The results demonstrated that all three types of data appear to carry complementary information relevant to the risk of developing RP. To facilitate further clinical evaluation, we provided a nomogram and discuss the potential clinical benefits of applying the RP predictive model.

Several studies to date have been conducted to predict RP by extracting handcrafted radiomics features from CT. Cunliffe et al. [38] explored the correlation between radiomics and RP and found that 12 radiomics features extracted from CT images of patients with esophageal cancer changed over time in association with the development of RP (AUC=0.78),

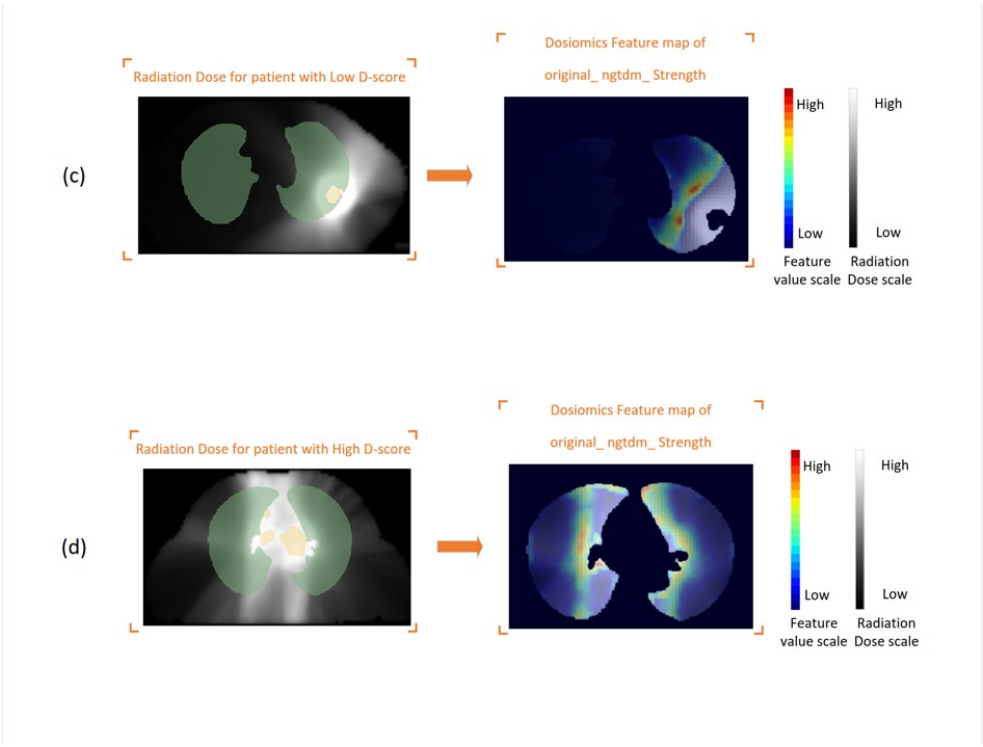
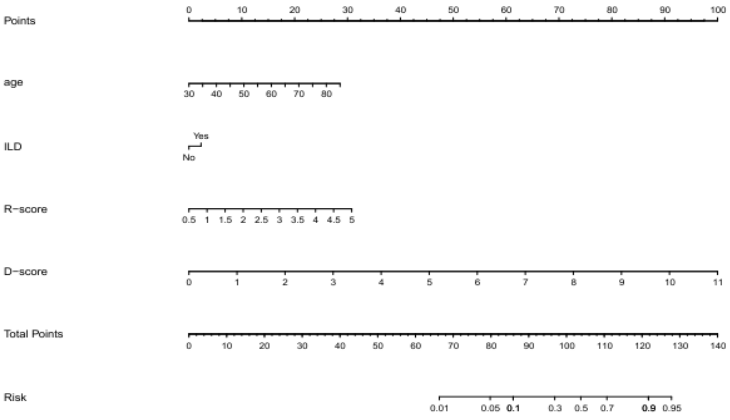


Figure 3. (a) The left image is the planning CT image of a patient with a low Radiomics risk score (R-score). The right image is the radiomics feature (original_ngtdm_Complexity) map of CT image at roughly the same level as shown on the left. Feature values are indicated from dark to light.

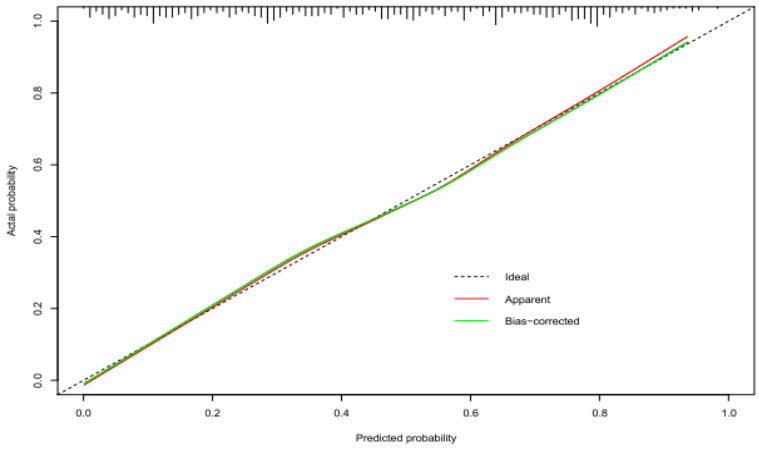
(b) The left image is the planning CT image of a patient with a high R-score. The right image is the radiomics feature (original_ngtdm_Complexity) map of CT image at roughly the same level as shown on the left.

(c) The left image is the radiation dose (RD) image of a patient with a low Dosiomics risk score (D-score). The right image is the dosiomics feature (original_ngtdm_Strength) map of RD image at roughly the same level as shown on the left. Feature values are represented by rainbow color bar, i.e., from blue to red. The irradiation dose is indicated from dark to light.

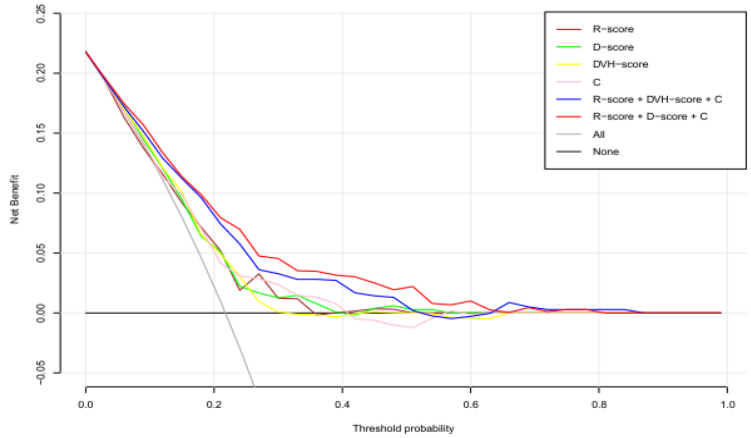
(d) The left image is the radiation dose (RD) image of a patient with a high D-score. The right image is the dosiomics feature (original_ngtdm_Strength) map of RD image at roughly the same level as shown on the left.


Points
0 10 20 30 40 50 60 70 80 90 100
age
30 40 50 60 70 80
ILD
Yes
No
R−score
0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
D−score
0 1 2 3 4 5 6 7 8 9 10 11
Total Points
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Risk
0.01 0.05 0.1 0.3 0.5 0.7 0.9 0.95

(a)


Actal probability
0.0 0.2 0.4 0.6 0.8 1.0
Predicted probability
Ideal
Apparent
Bias−corrected

(b)


Net Benefit
−0.05 0.00 0.05 0.10 0.15 0.20 0.25
0.0 0.2 0.4 0.6 0.8 1.0
Threshold probability
R−score
D−score
DVH−score
C
R−score + DVH−score + C
R−score + D−score + C
All
None

(c)

Figure 4. (a) Nomogram predicting the occurrence of symptom RP. Abbreviations: ILD: Interstitial lung disease; R-score = Radiomics risk score; D-score = Dosiomics risk score. (b) Calibration curve with a bootstrap resampling validation of prediction model combining radiomics risk score, dosiomics risk score, and clinical parameters. Dashed line indicated the ideal model in which predicted and actual probabilities were perfectly identical; Red line indicated actual performance with apparent accuracy; Green line indicated bootstrap corrected estimate of the calibration curve. (c) Decision curve analysis of prediction models. The color lines represent the DCA of different prediction models, the horizontal black line represents the hypothesis that no patients receive interventions, the oblique gray line represents the hypothesis that all patients receive the interventions. Abbreviations: R-score = Radiomics risk score; D-score = Dosiomics risk score; DVH-score = dose-volume histogram score; C. = clinical parameters.

Discussion

Identifying patients at higher risk of developing RP following thoracic irradiation remains an important and topical clinical question, as this adverse event directly affects patient prognosis and reduces quality of life. Patients with RP are a highly heterogeneous group, hence this study evaluated non-invasive methods (radiomics and dosiomics) using only pre-treatment information to characterize individual differences. In this study, the dosiomics features were shown to have stronger predictive power than the conventional DVH parameters, and the combination of a radiomics signature, a dosiomics signature, and two clinical factors were found to be predictive of RP. The results demonstrated that all three types of data appear to carry complementary information relevant to the risk of developing RP. To facilitate further clinical evaluation, we provided a nomogram and discuss the potential clinical benefits of applying the RP predictive model.

Several studies to date have been conducted to predict RP by extracting handcrafted radiomics features from CT. Cunliffe et al. [38] explored the correlation between radiomics and RP and found that 12 radiomics features extracted from CT images of patients with esophageal cancer changed over time in association with the development of RP (AUC=0.78), however, this study focuses on measurement and assessment rather than prediction. Krafft et al. [21] performed an in-depth study for lung cancer and concluded that the best predictive power (AUC=0.68) was achieved when combining radiomics, clinical and dosimetric parameters to build the model. Similar findings were obtained in a study of esophageal cancer by Du et al [20]. They developed a model combining radiomics, clinical and dosimetric parameters by studying 96 patients with esophageal cancer (AUC=0.91). Although these studies included small sample sizes, they inspired us that the combination of handcrafted radiomics features and dosimetric parameters can improve the predictive power of the model. For dosiomics, several studies have demonstrated its potential to predict radiotherapy-related endpoints, including prognosis [39-41] and treatment efficacy [42, 43], but there are very few studies using handcrafted dosiomics to predict side effects. A recent study published by Takanori et al. [25] used a combination of dosiomics and dose-volume indices to predict the occurrence of RP and concluded that dosiomics has the ability to predict RP. Liang et al. [26] conducted a study on dosiomics prediction of RP and confirmed that dosiomics predictive ability was superior to both dosimetric and NTCP predictors (AUC of 0.78 compared to 0.68 and 0.74), which gives us an idea that dosiomics relative to dosimetrics perhaps possessing more dimensional information.

Based on the results of this study (Table 2) we conclude that the predictive power and stability (with narrower 95%CI) of the model based on dosiomics features is stronger than the model based on dosimetrics. The correlation analysis between dosimetric and D-score showed that they are correlated, where D-score correlates with V30, V25, and V20 between 0.7 and 0.8 (Supplementary material F Figure 3b). Although both dosiomics and dosimetric are quantitative values obtained by calculating from 3D dose distributions, dosiomics obtains more detailed information from texture analysis of the dose distribution, while dosimetric obtains information based on dose-volume histograms. The shape features, which measure the dose delivery from another perspective, may also give a stronger predictive power to the dosiomics. Combining the results of this study and the published dosiomics studies to date, we suggest that neither can replace the other. Inspired by radiomics studies, we resampled the RD images to 1 mm. Different dose grids affect dosiomics feature values [29], however, the utility of resampling RD images, more specifically, whether resampling improves the reproducibility and stability of dosiomics features, requires more research. Placidi et al. conducted a multi-institutional basic study on dosiomics features, which concluded that dosiomics is a tool with predictive potential suitable for multi-institutional studies by analyzing the reproducibility, stability, and sensitivity of dosiomics features [29]. Our results also demonstrate that dosiomics have predictive potential and therefore it is worthwhile to investigate dosiomics more extensively and deeply.

To the best of our knowledge, no previous published studies have combined handcrafted radiomics, dosiomics, and clinical parameters of lung cancer in various ways and compared their ability to predict RP. In this work, we have compared models with radiomics alone, and with 3D spatial dose quantitative features (dosiomics) and we then go beyond current knowledge by proposing a combined model which shows that radiomics and dosiomics are complementary thus leading to improved model performance. We implemented a careful and objective feature selection approach, with robustness as the selection principle for each step of feature selection rather than best predictive ability, which to some extent avoids the occurrence of chance events. After this, the robust model validation approach was conducted and validated using bootstrap datasets and a prospective dataset, respectively, with over-optimism correction in both ways. Meanwhile, the number of variables in the model was controlled to avoid overfitting. The objective potential of radiomics/dosiomics for predicting RP was explored according to such a process.

We evaluated the performance of the model in three aspects, discrimination ability, calibration, and clinical application potential [44-46]. First, the differences between the training set, bootstrapping set, and test set are satisfactory in the results of discriminative validation, and the fluctuation range of 1000 repetitions is small. Based on this result, we think the model has stable prediction ability and low risk of overfitting. Second, the goodness of fit is another evaluation criterion for the prediction model. The final comprehensive model has excellent calibration, with no significant over- or under-estimation for different risk intervals. Third, a nomogram was built to assist clinical practice, and an offline calculator was provided to facilitate the calculation of R/D-score. The potential of the predictive model for clinical application was also evaluated using DCA. In Figure 4c, it can be seen that the nomogram-based prediction model has positive net benefits. In more detail, the net benefit of the prediction model is greater than the hypothesis that all patients receive RP prophylaxis or pro-active countermeasures (e.g., taking drugs to prevent RP or reducing the dose of radiotherapy) and that all patients do not receive such measures indiscriminately. It is worth

noting that the net benefit of the D-score-based model is higher than that of the DVH-score-based model, implying that the model with the D-score has more potential clinical benefit. In summary, the model we developed has potential clinical utility.

In univariate analysis of clinical parameters, whether receiving CCRT had an effect on the occurrence of RP, and the incidence of RP was lower in patients who received CCRT, which is not consistent with clinical experience and with findings in most studies [12, 13, 47]. This might be a bias due to subjective clinical decision making by physicians. Patients included in our study were evaluated by physicians for risk prior to receiving CCRT, and patients with poor health status and high incidence of radiation therapy side effects in the opinion of physicians would not be given CCRT. Some patients will receive potentially lower prescription doses in the radical dose range with stricter dose constraints of the lung to ensure they can complete a full cycle of radiotherapy without serious radiation therapy side effects. Similar views have been proposed by other researchers [8]. A negative correlation between age and CCRT can be seen in Supplementary material F Figure 3, which also illustrates the subjectivity in the setting of the CCRT protocol. Our findings suggest that ILD is a risk factor for the development of RP. Clinically, RT may lead to exacerbation of ILD and thus interfere with the diagnosis of RP [48]. Accordingly, in this study, the diagnosis of RP in patients with ILD was determined by collaboration with radiologists. And it should be noted that strictly to define, the ILD mentioned in this study is subclinical ILD, according to previous studies. [49, 50]. To investigate the effect of ILD on the model, we excluded patients with ILD in all datasets and performed the same independent validation methods as described previously. Based on the results (Supplementary Material G), we propose our hypothesis: 1. The radiomics model focuses not only on lung texture but also includes other information, as there is no significant difference between the model including or excluding patients with ILD. 2. The discrimination performance of the model built by dosiomics or DVH metrics is improved by excluding patients with ILD, as dose-based models are difficult to predict RP in patients with ILD. 3. ILD is a critical clinical predictor. In previous reports, patients with ILD have high risk of RP, and ILD has been considered a high risk factor for fatal RP [51, 52]. A number of studies have been conducted to analyze the relationship between age and RP [8, 53]. Several studies [54-56] and a meta-analysis [57] have shown that older patients have a higher risk of developing RP. However, some studies did not find an association between age and the risk of RP [58, 59]. In summary, patients who are elderly or/and have ILD should be given more attention and a more comprehensive risk assessment before receiving radiotherapy.

A current challenge in radiomics/dosiomics studies is interpretability, and we attempted to analyze the omics results from a clinical perspective. The analysis revealed no strong correlation between clinical parameters and the R-score (Supplementary material F Figure 3a). However, imaging radiomics contains a large amount of quantitative information and it may not be possible to interpret the full meaning of what it represents using a few clinical parameters. The feature maps of radiomics and dosiomics can provide the direct visualization of voxel-based feature values. As shown in Figure 3 (a) and (b), the radiomic feature "original_ngtdm_Complexity" can reflect the texture characteristics, and for ILD patients, higher voxel-based feature values were obtained compared to patients without pre-existing lung disease. The dosomic feature "original_ngtdm_Strength" (Figure 3(c) and (d)) shows a pattern of variation from high to low dose, which is some reflection of the radiotherapy planning pattern. Feature maps of other features are provided in the Supplementary material

F Figure 5. We compared the feature maps with the follow-up diagnostic CTs and found that the radiomics signature map did not match the areas of symptomatic RP. In contrast, there is a significant overlap between some dosiomics feature maps and the symptomatic RP regions (Supplementary Material F Figure 6). This is consistent with the clinical understanding that the regional localization of symptomatic RP is more closely related to the physical radiation dose distribution.

The Rad/Dosiomics features selected in this study include shape features, which give us a suggestion that the contouring of the lung tissue is important. Currently, manual segmentation is still the "gold standard", but it is time consuming. Therefore, we performed manual check to ensure the accuracy and quality of the automatic segmentation, following processing by the automatic segmentation software. We think this approach is suitable for future multi-institutional studies to assure accuracy while reducing physician workload. Since dosiomics is still relatively little studied, there are no standardized parameter settings yet. Although it has common points with imaging radiomics, some of the parameter settings are different and have a great impact on the results, so we provide the setting files in Supplementary material B, which also provides a reference for future investigators.

This present study has several limitations. First, although the sample size included in our study is relatively large for radiomics/dosiomics RP prediction study, the prospective validation sample size is too small. Our institution's prospective study is still ongoing and continues to expand the sample size. For the scope of this work, we did not yet optimize the plan based on the results of the omics model. We acknowledge that the prospective data set used in this study was derived from an observational prospective study and no interventions were implemented in those patients based on our abovementioned predictive models. By prospective inclusion, we were strictly only able to standardize the follow-up strategy, specifically, patients received regular follow-up examinations and RP grade was jointly diagnosed by the study investigators, which ensured the highest achievable accuracy and consistency of the endpoints, while giving more attention towards patients with likelihood of developing RP. At the present time, it is not yet clear which aspect of the treatment plan to change in order to intervene correctly in the planning dosimetry process, so this requires further work. A prospectively-enrolled clinical study would be important in the clinical implementation process, this is planned for future work, but is not the principal purpose of this paper. Second, the current gold standard for predictive model validation is still multi-institutional real-world external validation. Third, we built a binary prediction model because the sample size is limited and as the dataset expands, models that can predict different grades are needed. Fourth, pneumonitis associated with immune checkpoint inhibitor (ICI) therapy is an important adverse event. However, the relationship between ICI and RP or the relationship between ICI-associated pneumonitis and radiotherapy-associated pneumonitis remains unclear. Therefore, we excluded patients treated with ICI. Fifth, most current studies comparing machine learning and deep learning conclude that deep learning has stronger predictive power. This study is a pilot study. Therefore, deep learning which is currently a "black box" is not applied, and machine learning with observable processing is chosen. Finally, individualized treatment should incorporate more multidimensional omics information, including genomics and imaging multimodality data. To address several issues above, our institution is conducting a multi-institutional study.

Conclusions

This study was a TRIPOD type 3 prediction model development study, validated using bootstrap samples and a prospective validation set. The radiomics, dosiomics signature, and clinical parameters associated with RP were selected. By comparing the performance of the models built by combining different types of parameters, the best prediction model was found with the best performance of the three types of parameters combined. Furthermore, a comprehensive nomogram was built to assist in clinical decision making and individualized treatment. In the future, a multi-institutional study is needed.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: a cancer journal for clinicians. 2021;71:209-49. doi:10.3322/caac.21660.

2. Yang W-C, Hsu F-M, Yang P-C. Precision radiotherapy for non-small cell lung cancer. J Biomed Sci. 2020;27:82. doi:10.1186/s12929-020-00676-5.

3. Vinod SK, Hau E. Radiotherapy treatment for lung cancer: Current status and future directions. Respirology. 2020;25 Suppl 2:61-71. doi:10.1111/resp.13870.

4. Luo H-S, Huang H-C, Lin L-X. Effect of modern high-dose versus standard-dose radiation in definitive concurrent chemo-radiotherapy on outcome of esophageal squamous cell cancer: a meta-analysis. Radiation Oncology. 2019;14:178. doi:10.1186/s13014-019-1386-x.

5. Ladbury CJ, Rusthoven CG, Camidge DR, Kavanagh BD, Nath SK. Impact of Radiation Dose to the Host Immune System on Tumor Control and Survival for Stage III Non-Small Cell Lung Cancer Treated with Definitive Radiation Therapy. International Journal of Radiation Oncology*Biology*Physics. 2019;105:346-55. doi:10.1016/j.ijrobp.2019.05.064.

6. Kong F-M, Ten Haken RK, Schipper MJ, Sullivan MA, Chen M, Lopez C, et al. High-dose radiation improved local tumor control and overall survival in patients with inoperable/unresectable non-small-cell lung cancer: long-term results of a radiation dose escalation study. International Journal of Radiation Oncology, Biology, Physics. 2005;63:324-33. doi:10.1016/j.ijrobp.2005.02.010.

7. Niu S, Zhang Y. Applications and therapeutic mechanisms of action of mesenchymal stem cells in radiation-induced lung injury. Stem Cell Res Ther. 2021;12:212. doi:10.1186/s13287-021-02279-9.

8. Ullah T, Patel H, Pena GM, Shah R, Fein AM. A contemporary review of radiation pneumonitis. Curr Opin Pulm Med. 2020;26:321-5. doi:10.1097/MCP.0000000000000682.

9. Käsmann L, Dietrich A, Staab-Weijnitz CA, Manapov F, Behr J, Rimner A, et al. Radiation-induced lung toxicity - cellular and molecular mechanisms of pathogenesis, management, and literature review. Radiation Oncology (London, England). 2020;15:214. doi:10.1186/s13014-020-01654-9.

10. Thomas R, Chen Y-H, Hatabu H, Mak RH, Nishino M. Radiographic patterns of symptomatic radiation pneumonitis in lung cancer patients: Imaging predictors for clinical severity and outcome. Lung Cancer. 2020;145:132-9. doi:10.1016/j.lungcan.2020.03.023.

11. Nalbantov G, Kietselaer B, Vandecasteele K, Oberije C, Berbee M, Troost E, et al. Cardiac comorbidity is an independent risk factor for radiation-induced lung toxicity in lung cancer patients. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2013;109:100-6. doi:10.1016/j.radonc.2013.08.035.

12. Arroyo-Hernández M, Maldonado F, Lozano-Ruiz F, Muñoz-Montaño W, Nuñez-

Baez M, Arrieta O. Radiation-induced lung injury: current evidence. BMC Pulm Med. 2021;21:9. doi:10.1186/s12890-020-01376-4.

13. Kong F-MS, Wang S. Nondosimetric risk factors for radiation-induced lung toxicity. Seminars in Radiation Oncology. 2015;25:100-9. doi:10.1016/j.semradonc.2014.12.003.

14. Huang Q, Xie F, Ouyang X. Predictive SNPs for radiation-induced damage in lung cancer patients with radiotherapy: a potential strategy to individualize treatment. Int J Biol Markers. 2015;30:e1-11. doi:10.5301/jbm.5000108.

15. Niu X, Li H, Chen Z, Liu Y, Kan M, Zhou D, et al. A study of ethnic differences in TGFβ1 gene polymorphisms and effects on the risk of radiation pneumonitis in non-small-cell lung cancer. Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer. 2012;7:1668-75. doi:10.1097/JTO.0b013e318267cf5b.

16. Yu H, Wu H, Wang W, Jolly S, Jin J-Y, Hu C, et al. Machine Learning to Build and Validate a Model for Radiation Pneumonitis Prediction in Patients with Non-Small Cell Lung Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2019;25:4343-50. doi:10.1158/1078-0432.CCR-18-1084.

17. Mak RH, Alexander BM, Asomaning K, Heist RS, Liu C-y, Su L, et al. A single-nucleotide polymorphism in the methylene tetrahydrofolate reductase (MTHFR) gene is associated with risk of radiation pneumonitis in lung cancer patients treated with thoracic radiation therapy. Cancer. 2012;118:3654-65. doi:10.1002/cncr.26667.

18. Bodalal Z, Trebeschi S, Nguyen-Kim TDL, Schats W, Beets-Tan R. Radiogenomics: bridging imaging and genomics. Abdominal Radiology (New York). 2019;44:1960-84. doi:10.1007/s00261-019-02028-w.

19. Wang L, Gao Z, Li C, Sun L, Li J, Yu J, et al. Computed tomography-based delta-radiomics analysis for discriminating radiation pneumonitis in patients with esophageal cancer after radiation therapy. International Journal of Radiation Oncology, Biology, Physics. 2021. doi:10.1016/j.ijrobp.2021.04.047.

20. Du F, Tang N, Cui Y, Wang W, Zhang Y, Li Z, et al. A Novel Nomogram Model Based on Cone-Beam CT Radiomics Analysis Technology for Predicting Radiation Pneumonitis in Esophageal Cancer Patients Undergoing Radiotherapy. Front Oncol. 2020;10:596013. doi:10.3389/fonc.2020.596013.

21. Krafft SP, Rao A, Stingo F, Briere TM, Court LE, Liao Z, et al. The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis. Med Phys. 2018;45:5317-24. doi:10.1002/mp.13150.

22. Liu Y, Wang W, Shiue K, Yao H, Cerra-Franco A, Shapiro RH, et al. Risk factors for symptomatic radiation pneumonitis after stereotactic body radiation therapy (SBRT) in patients with non-small cell lung cancer. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;156:231-8. doi:10.1016/j.radonc.2020.10.015.

23. Saha A, Beasley M, Hatton N, Dickinson P, Franks K, Clarke K, et al. Clinical and

dosimetric predictors of radiation pneumonitis in early-stage lung cancer treated with Stereotactic Ablative radiotherapy (SABR) - An analysis of UK's largest cohort of lung SABR patients. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;156:153-9. doi:10.1016/j.radonc.2020.12.015.

24. Bourbonne V, Da-Ano R, Jaouen V, Lucia F, Dissaux G, Bert J, et al. Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;155:144-50. doi:10.1016/j.radonc.2020.10.040.

25. Adachi T, Nakamura M, Shintani T, Mitsuyoshi T, Kakino R, Ogata T, et al. Multi-institutional dose-segmented dosiomic analysis for predicting radiation pneumonitis after lung stereotactic body radiation therapy. Med Phys. 2021;48:1781-91. doi:10.1002/mp.14769.

26. Liang B, Yan H, Tian Y, Chen X, Yan L, Zhang T, et al. Dosiomics: Extracting 3D Spatial Features From Dose Distribution to Predict Incidence of Radiation Pneumonitis. Front Oncol. 2019;9:269. doi:10.3389/fonc.2019.00269.

27. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015;162:W1-W73. doi:10.7326/M14-0698.

28. Matschinske J, Alcaraz N, Benis A, Golebiewski M, Grimm DG, Heumos L, et al. The AIMe registry for artificial intelligence in biomedical research. Nat Methods. 2021;18:1128-31. doi:10.1038/s41592-021-01241-0.

29. Placidi L, Gioscio E, Garibaldi C, Rancati T, Fanizzi A, Maestri D, et al. A Multicentre Evaluation of Dosiomics Features Reproducibility, Stability and Sensitivity. Cancers (Basel). 2021;13:3835. doi:10.3390/cancers13153835.

30. Adachi T, Nakamura M, Kakino R, Hirashima H, Iramina H, Tsuruta Y, et al. Dosiomic feature comparison between dose-calculation algorithms used for lung stereotactic body radiation therapy. Radiol Phys Technol. 2022. doi:10.1007/s12194-022-00651-9.

31. Shi Z, Traverso A, van Soest J, Dekker A, Wee L. Technical Note: Ontology-guided radiomics analysis workflow (O-RAW). Med Phys. 2019;46:5677-84. doi:10.1002/mp.13844.

32. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77:e104-e7. doi:10.1158/0008-5472.CAN-17-0339.

33. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295:328-38. doi:10.1148/radiol.2020191145.

34. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Mag-

netic Resonance Imaging. 2012;30:1323-41. doi:10.1016/j.mri.2012.05.001.

35. Larue RTHM, van Timmeren JE, de Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncologica. 2017;56:1544-53. doi:10.1080/0284186X.2017.1351624.

36. Compter I, Verduin M, Shi Z, Woodruff HC, Smeenk RJ, Rozema T, et al. Deciphering the glioblastoma phenotype by computed tomography radiomics. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;160:132-9. doi:10.1016/j.radonc.2021.05.002.

37. Shi Z, Zhang Z, Liu Z, Zhao L, Ye Z, Dekker A, et al. Methodological quality of machine learning-based quantitative imaging analysis studies in esophageal cancer: a systematic review of clinical outcome prediction after concurrent chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2021. doi:10.1007/s00259-021-05658-9.

38. Cunliffe A, Armato SG, Castillo R, Pham N, Guerrero T, Al-Hallaq HA. Lung Texture in Serial Thoracic Computed Tomography Scans: Correlation of Radiomics-based Features With Radiation Therapy Dose and Radiation Pneumonitis Development. International Journal of Radiation Oncology*Biology*Physics. 2015;91:1048-56. doi:10.1016/j.ijrobp.2014.11.030.

39. Rossi L, Bijman R, Schillemans W, Aluwini S, Cavedon C, Witte M, et al. Texture analysis of 3D dose distributions for predictive modelling of toxicity rates in radiotherapy. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2018;129:548-53. doi:10.1016/j.radonc.2018.07.027.

40. Gabryś HS, Buettner F, Sterzing F, Hauswald H, Bangert M. Design and Selection of Machine Learning Methods Using Radiomics and Dosiomics for Normal Tissue Complication Probability Modeling of Xerostomia. Front Oncol. 2018;8:35. doi:10.3389/fonc.2018.00035.

41. Zhen X, Chen J, Zhong Z, Hrycushko B, Zhou L, Jiang S, et al. Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. Phys Med Biol. 2017;62:8246-63. doi:10.1088/1361-6560/aa8d09.

42. Buizza G, Paganelli C, D'Ippolito E, Fontana G, Molinelli S, Preda L, et al. Radiomics and Dosiomics for Predicting Local Control after Carbon-Ion Radiotherapy in Skull-Base Chordoma. Cancers (Basel). 2021;13. doi:10.3390/cancers13020339.

43. Wu A, Li Y, Qi M, Lu X, Jia Q, Guo F, et al. Dosiomics improves prediction of locoregional recurrence for intensity modulated radiotherapy treated head and neck cancer cases. Oral Oncol. 2020;104:104625. doi:10.1016/j.oraloncology.2020.104625.

44. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128-38. doi:10.1097/EDE.0b013e3181c30fb2.

45. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimi-

nation and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. JAMA. 2017;318:1377. doi:10.1001/jama.2017.12126.

46. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2016;34:2534-40. doi:10.1200/JCO.2015.65.5654.

47. Bledsoe TJ, Nath SK, Decker RH. Radiation Pneumonitis. Clinics in Chest Medicine. 2017;38:201-8. doi:10.1016/j.ccm.2016.12.004.

48. Kocak Z, Evans ES, Zhou S-M, Miller KL, Folz RJ, Shafman TD, et al. Challenges in defining radiation pneumonitis in patients with lung cancer. International Journal of Radiation Oncology, Biology, Physics. 2005;62:635-8. doi:10.1016/j.ijrobp.2004.12.023.

49. Yamaguchi S, Ohguri T, Matsuki Y, Yahara K, Oki H, Imada H, et al. Radiotherapy for thoracic tumors: association between subclinical interstitial lung disease and fatal radiation pneumonitis. Int J Clin Oncol. 2015;20:45-52. doi:10.1007/s10147-014-0679-1.

50. Doyle TJ, Hunninghake GM, Rosas IO. Subclinical interstitial lung disease: why you should care. Am J Respir Crit Care Med. 2012;185:1147-53. doi:10.1164/rccm.201108-1420PP.

51. Doi H, Nakamatsu K, Nishimura Y. Stereotactic body radiotherapy in patients with chronic obstructive pulmonary disease and interstitial pneumonia: a review. Int J Clin Oncol. 2019;24:899-909. doi:10.1007/s10147-019-01432-y.

52. Okumura M, Hojo H, Nakamura M, Hiyama T, Nakamura N, Zenda S, et al. Radiation pneumonitis after palliative radiotherapy in cancer patients with interstitial lung disease. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;161:47-54. doi:10.1016/j.radonc.2021.05.026.

53. Giuranno L, Ient J, De Ruysscher D, Vooijs MA. Radiation-Induced Lung Injury (RILI). Front Oncol. 2019;9:877. doi:10.3389/fonc.2019.00877.

54. Leprieur EG, Fernandez D, Chatellier G, Klotz S, Giraud P, Durdux C. Acute radiation pneumonitis after conformational radiotherapy for nonsmall cell lung cancer: clinical, dosimetric, and associated-treatment risk factors. J Cancer Res Ther. 2013;9:447-51. doi:10.4103/0973-1482.119339.

55. Dang J, Li G, Zang S, Zhang S, Yao L. Risk and predictors for early radiation pneumonitis in patients with stage III non-small cell lung cancer treated with concurrent or sequential chemoradiotherapy. Radiation Oncology (London, England). 2014;9:172. doi:10.1186/1748-717X-9-172.

56. Tsujino K, Hashimoto T, Shimada T, Yoden E, Fujii O, Ota Y, et al. Combined analysis of V20, VS5, pulmonary fibrosis score on baseline computed tomography, and patient age improves prediction of severe radiation pneumonitis after concurrent chemoradiotherapy for locally advanced non-small-cell lung cancer. Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer. 2014;9:983-90. doi:10.1097/JTO.0000000000000187.

57. Vogelius IR, Bentzen SM. A literature-based meta-analysis of clinical risk factors for development of radiation induced pneumonitis. Acta Oncologica (Stockholm, Sweden). 2012;51:975-83. doi:10.3109/0284186X.2012.718093.

58. Wen J, Liu H, Wang Q, Liu Z, Li Y, Xiong H, et al. Genetic variants of the LIN28B gene predict severe radiation pneumonitis in patients with non-small cell lung cancer treated with definitive radiation therapy. European Journal of Cancer (Oxford, England: 1990). 2014;50:1706-16. doi:10.1016/j.ejca.2014.03.008.

59. Bradley JD, Hope A, El Naqa I, Apte A, Lindsay PE, Bosch W, et al. A nomogram to predict radiation pneumonitis, derived from a combined analysis of RTOG 9311 and institutional data. International Journal of Radiation Oncology, Biology, Physics. 2007;69:985-92. doi:10.1016/j.ijrobp.2007.04.077.

Supplementary Materials

Supplementary material A

Inclusion and exclusion criteria for retrospective and prospective data

The first dataset was collected retrospectively as a training set and validation set. A total of 314 patients treated between January 2013 and December 2018 with definitive RT at Anonymized for Review hospital were considered for the retrospective dataset. The inclusion criteria were as follows: (1) Patients identified with histologically confirmed NSCLC or SCLC. (2) Diagnosed with Stage I-III NSCLC and limited-stage SCLC (American Joint Committee on Cancer, 8th edition, 2017) before RT, and patients underwent radical RT. (3) No thoracic RT or thoracic surgery prior to RT. (4) CT examinations were performed at 1, 3, and 6 months (± 15 days) after treatment at Anonymized for Review Hospital. Patients were excluded, if treatment break of more than 5 days during RT, if patients received surgical treatment within 6 months after radiotherapy, if patients received adjuvant/concurrent immunotherapy, if there was also a second primary tumor, and if the patients had a lung infection within 6 months after radiotherapy, so it was difficult to identify whether it was RP.

The second dataset was collected prospectively at the same institution as a test set. A total of 56 patients were enrolled in the study from October 2018 to March 2019. Finally, 35 patients were included in the analysis. 21 patients were excluded because did not meet the eligible criteria, fourteen of which did not follow up CT as planned, six of which did not complete radiotherapy, and one patient died two months after radiation therapy. The inclusion and exclusion criteria were the same as the retrospective dataset and these patients were followed-up every month after had received radiotherapy. The follow-up items included blood routine examination, C-reactive protein, tumor markers associated with lung cancer, chest X-rays, and patients received CT examination at 1, 3, and 6 months (± 7 days) after radiotherapy.

Patient Characteristics for prospective data

Supplementary Table 1. Patient Characteristics for prospective data

Characteristics	Pros pts	Without RP2	With RP2	P*
	n (%)	Mean ± SD	Mean ± SD	
Age median	62 (34-75)	61.5 (34-75)	62 (59-68)	0.363
Gender				1.000
Male	23 (65.7%)	17 (73.9%)	6 (26.1%)	
Female	12 (34.3%)	9 (75.0%)	3 (25.0%)	
Smoking				1.000
Yes	26 (74.3%)	19 (73.1%)	7 (26.9%)	
No	9 (25.7%)	7 (77.8%)	2 (22.2%)	
KPS				1.000
≤80	13 (37.1%)	10 (76.9%)	3 (23.1%)	
>80	22 (62.9%)	16 (72.7%)	6 (27.3%)	

Diabetes				1.000
Yes	2 (5.7%)	2 (100.0%)	0 (0%)	
No	33 (94.3%)	24 (72.7%)	9 (27.3%)	
ILD				0.192
Yes	9 (25.7%)	5 (55.6%)	4 (44.4%)	
No	26 (74.3%)	21 (80.8%)	5 (19.2%)	
Pathology				0.776
LUSC	8 (22.9%)	5 (62.5%)	3 (37.5%)	
LUAD	10 (28.6%)	8 (80.0%)	2 (20.0%)	
SCLC	17 (48.5%)	13 (76.5%)	4 (23.5%)	
Induc chemo				0.553
Yes	31 (88.6%)	22 (71.0%)	9 (29.0%)	
No	4 (11.4%)	4 (100.0%)	0 (0%)	
CCRT				0.081
Yes	8 (22.9%)	8 (100.0%)	0 (0%)	
No	27 (77.1%)	18 (66.7%)	9 (33.3%)	
Conso chemo				0.245
Yes	19 (54.3%)	16 (84.2%)	3 (15.8%)	
No	16 (45.7%)	10 (62.5%)	6 (37.5%)	
PGTV (Gy)	60.200±2.870	60.423±2.862	59.556±2.963	0.436
Smoking index	668.600±550.412	646.154±566.555	733.333±527.376	0.679

Abbreviations: Retro = retrospective; Pts = patients; Pros = prospective; LUSC = lung squamous cell carcinoma; LUAD = lung adenocarcinoma; SCLC = small cell lung cancer; IMRT = intensity-modulated radiotherapy; VMAT = volumetric modulated arc therapy; chemo = chemotherapy; KPS = Karnofsky performance score; Induc chemo = induction chemotherapy; CCRT = concurrent chemoradiotherapy; Conso chemo = consolidation chemotherapy; PGTV = planning gross tumor volume.

*The differences in characteristics were evaluated by logistic regression for continuous variables or Pearson X2 test and exact Fisher test for categorical variables

Supplementary material B

Radiomics and dosiomics features extraction parameter settings file

Radiomics

imageType:

Original:

binWidth: 25

```
featureClass:
  shape: # Remove VoxelVolume, correlated to Volume
    - Elongation
    - Flatness
    - LeastAxisLength
    - MajorAxisLength
    - Maximum2DDiameterColumn
    - Maximum2DDiameterRow
    - Maximum2DDiameterSlice
    - Maximum3DDiameter
    - MeshVolume
    - MinorAxisLength
    - Sphericity
    - SurfaceArea
    - SurfaceVolumeRatio
  firstorder: # Remove Total Energy, correlated to Energy  (due to resampling enabled)
    - 10Percentile
    - 90Percentile
    - Energy
    - Entropy
    - InterquartileRange
    - Kurtosis
    - Maximum
    - Mean
    - MeanAbsoluteDeviation
    - Median
    - Minimum
    - Range
```

- RobustMeanAbsoluteDeviation
- RootMeanSquared
- Skewness
- Uniformity
- Variance

glcm: # Disable SumAverage by specifying all other GLCM features available

- ‘Autocorrelation’
- ‘JointAverage’
- ‘ClusterProminence’
- ‘ClusterShade’
- ‘ClusterTendency’
- ‘Contrast’
- ‘Correlation’
- ‘DifferenceAverage’
- ‘DifferenceEntropy’
- ‘DifferenceVariance’
- ‘JointEnergy’
- ‘JointEntropy’
- ‘Imc1’
- ‘Imc2’
- ‘Idm’
- ‘Idmn’
- ‘Id’
- ‘Idn’
- ‘InverseVariance’
- ‘MaximumProbability’
- ‘SumEntropy’
- ‘SumSquares’

```
  glrlm:
  glszm:
  gldm:
  ngtdm:
setting:
  interpolator: 'sitkBSpline'
  resampledPixelSpacing: [2, 2, 2]
  padDistance: 10  # Extra padding for large sigma valued LoG filtered images
  resegmentRange: [-3, 3]
  resegmentMode: sigma
  voxelArrayShift: 1000
```

Dosiomics

```
imageType:
  Original:
    binWidth: 0.5
featureClass:
  shape: # Remove VoxelVolume, correlated to Volume
    - Elongation
    - Flatness
    - LeastAxisLength
    - MajorAxisLength
    - Maximum2DDiameterColumn
    - Maximum2DDiameterRow
    - Maximum2DDiameterSlice
    - Maximum3DDiameter
    - MeshVolume
    - MinorAxisLength
    - Sphericity
```

- SurfaceArea
- SurfaceVolumeRatio

firstorder: # Remove Total Energy, correlated to Energy (due to resampling enabled)

- 10Percentile
- 90Percentile
- Energy
- Entropy
- InterquartileRange
- Kurtosis
- Maximum
- Mean
- MeanAbsoluteDeviation
- Median
- Minimum
- Range
- RobustMeanAbsoluteDeviation
- RootMeanSquared
- Skewness
- Uniformity
- Variance

glcm: # Disable SumAverage by specifying all other GLCM features available

- 'Autocorrelation'
- 'JointAverage'
- 'ClusterProminence'
- 'ClusterShade'
- 'ClusterTendency'
- 'Contrast'
- 'Correlation'

```
    - 'DifferenceAverage'
    - 'DifferenceEntropy'
    - 'DifferenceVariance'
    - 'JointEnergy'
    - 'JointEntropy'
    - 'Imc1'
    - 'Imc2'
    - 'Idm'
    - 'Idmn'
    - 'Id'
    - 'Idn'
    - 'InverseVariance'
    - 'MaximumProbability'
    - 'SumEntropy'
    - 'SumSquares'
  glrlm:
  glszm:
  gldm:
  ngtdm:
setting:
interpolator: 'sitkBSpline'
  resampledPixelSpacing: [2, 2, 2]
  padDistance: 10  # Extra padding for large sigma valued LoG filtered images
  voxelArrayShift: 0
```

Supplementary material C

Dose-volume histogram (DVH) metrics selection and model construction

Due to the colinearity of DVH metrics, it does is not suitable to perform the same feature selection approaches as radiomics/dosiomics. Instead, the predictive model is built using

the already acknowledged metrics V20 and MLD. The DVH-score was defined as the linear predictor of the multivariable LR radiomics model.

The validation method was performed in exactly the same way as for the radiomics/dosiomics model: (1) For each of the 1000 bootstraps, we fitted the logistic regression model coefficients on each bootstrap, and then computed its Area under the curve (AUC) of receiver operating characteristic curve (ROC) using the original non-bootstrapped development cohort. From these 1000 bootstraps, we computed the average AUC and its 95% confidence interval. (2) As external validation, we evaluated the DVH model using the prospectively-registered cohort of 35 subjects. Processing of these 35 subjects followed exactly the same procedure as for the model development cohort, and none of these subjects were used in any way during model construction.

Since V5 is an important predictor in the IMRT/VMAT era, we also built another DVH model by combining V5 and MLD. However, based on this dataset, the predictive power of the "V5+MLD" model is worse than that of the "V20+MLD" model, so we used the DVH model of V20 and MLD as the comparative model in this study.

Supplementary material D

Feature selection results and graphs

The top twenty features that were screened are displayed in **Supplementary Table 1.** The features are sorted according to the number of frequencies selected and shown in the **Supplementary Figure 1**. The cut-off points were decided based on the frequency breakpoints shown in the graphs. The cut-off points for both radiomics and dosiomics features are around 600.

The three most frequently selected signatures are shown in **Supplementary Table 2**, with the highest selected frequencies of 45 and 105 for the radiomics and dosiomics signatures, respectively.

Definitions of the selected features are provided in **Supplementary Table 3.**

Supplementary Table 1a. The top twenty radiomics features that were selected

No.	Radiomics features	Frequency
1	original_shape_Elongation	1000
2	original_shape_Flatness	922
3	original_shape_MinorAxisLength	871
4	original_shape_MeshVolume	746
5	original_firstorder_90Percentile	700
6	original_glcm_JointEntropy	696
7	original_ngtdm_Complexity	694
8	original_firstorder_Median	684
9	original_shape_Maximum2DDiameterSlice	677
10	original_glszm_LargeAreaEmphasis	670

11	original_shape_Maximum2DDiameterRow	663
12	original_gldm_DependenceNonUniformityNormalized	603
13	original_gldm_DependenceEntropy	587
14	original_glcm_DifferenceEntropy	563
15	original_ngtdm_Contrast	562
16	original_glszm_SmallAreaLowGrayLevelEmphasis	542
17	original_gldm_SmallDependenceLowGrayLevelEmphasis	537
18	original_shape_LeastAxisLength	525
19	original_shape_SurfaceVolumeRatio	525
20	original_ngtdm_Strength	521

Supplementary Table 1b. The top twenty dosiomics features that were selected

No.	dosiomics features	Frequency
1	original_shape_Elongation	1000
2	original_glszm_LargeAreaEmphasis	864
3	original_shape_Flatness	736
4	original_ngtdm_Strength	715
5	original_shape_SurfaceArea	704
6	original_shape_MeshVolume	693
7	original_shape_Maximum2DDiameterRow	643
8	original_glszm_GrayLevelVariance	642
9	original_shape_MinorAxisLength	641
10	original_ngtdm_Coarseness	622
11	original_ngtdm_Contrast	605
12	original_glszm_SmallAreaLowGrayLevelEmphasis	594
13	original_gldm_LargeDependenceEmphasis	555
14	original_shape_LeastAxisLength	554
15	original_glcm_DifferenceEntropy	545
16	original_glrlm_ShortRunLowGrayLevelEmphasis	544
17	original_glszm_ZoneEntropy	544
18	original_glrlm_RunLengthNonUniformity	537
19	original_glszm_SmallAreaHighGrayLevelEmphasis	526
20	original_gldm_DependenceEntropy	523

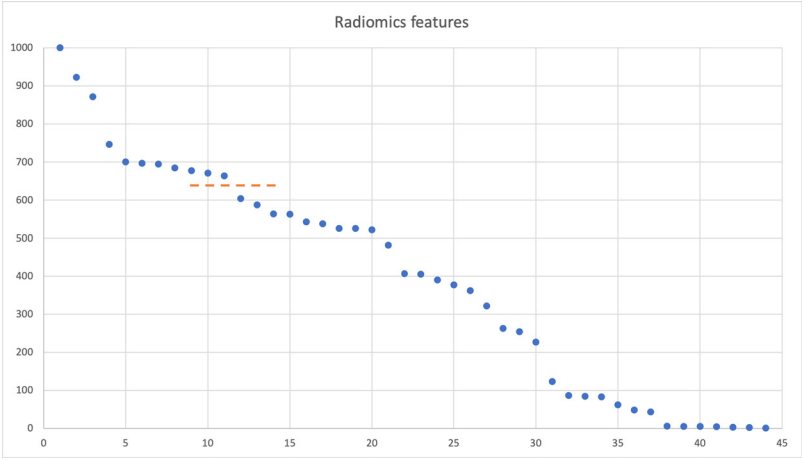
Supplementary Table 2a. The top three frequently selected radiomics signatures

No	Signature	Freq
1	original_glcm_JointEntropy + original_ngtdm_Complexity + original_shape_Elongation + original_shape_Flatness + original_shape_Maximum2DDiameterSlice + original_shape_MeshVolume + original_shape_MinorAxisLength	45

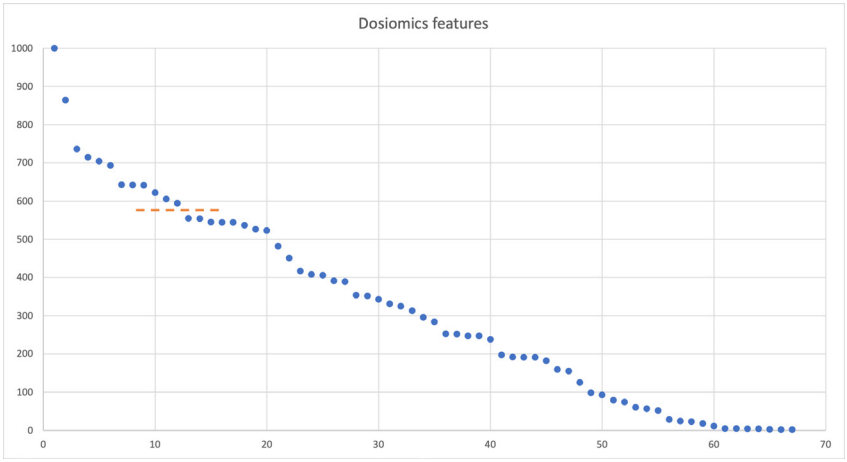
2	original_firstorder_90Percentile + original_firstorder_Median + original_glcm_JointEntropy + original_ngtdm_Complexity + original_shape_Elongation + original_shape_Flatness + original_shape_MeshVolume + original_shape_MinorAxisLength	40
3	original_firstorder_90Percentile + original_firstorder_Median + original_glszm_LargeAreaEmphasis + original_ngtdm_Complexity + original_shape_Elongation + original_shape_Flatness + original_shape_Maximum2DDiameterRow + original_shape_MeshVolume + original_shape_MinorAxisLength	36

Supplementary Table 2b. The top three frequently selected dosiomics signatures

No	Signature	Freq
1	original_glszm_GrayLevelVariance + original_glszm_LargeAreaEmphasis + original_ngtdm_Contrast + original_ngtdm_Strength + original_shape_MeshVolume + original_shape_SurfaceArea	105
2	original_glszm_GrayLevelVariance + original_glszm_LargeAreaEmphasis + original_ngtdm_Contrast + original_ngtdm_Strength + original_shape_MeshVolume	70
3	original_glszm_GrayLevelVariance + original_glszm_LargeAreaEmphasis + original_ngtdm_Contrast + original_ngtdm_Strength + original_shape_MeshVolume + original_shape_MinorAxisLength + original_shape_SurfaceArea	43



(a)



(b)

Supplementary Figure 1. (a) The radiomics features are sorted according to the number of frequencies selected. (b) The dosiomics features are sorted according to the number of frequencies selected.

Supplementary Table 3. Definitions of the selected features.

Feature	Definition
original_glcm_JointEntropy	Joint entropy is a measure of the randomness/variability in neighborhood intensity values.
original_ngtdm_Complexity	An image is considered complex when there are many primitive components in the image, i.e. the image is non-uniform and there are many rapid changes in gray level intensity.
original_shape_Elongation	Elongation shows the relationship between the two largest principal components in the ROI shape. For computational reasons, this feature is defined as the inverse of true elongation.
original_shape_Flatness	Flatness shows the relationship between the largest and smallest principal components in the ROI shape. For computational reasons, this feature is defined as the inverse of true flatness.
original_shape_Maximum2DDiameterSlice	Maximum 2D diameter (Slice) is defined as the largest pairwise Euclidean distance between tumor surface mesh vertices in the row-column (generally the axial) plane.
original_shape_MeshVolume	The volume of the ROI V is calculated from the triangle mesh of the ROI.
original_shape_MinorAxisLength	This feature yield the second-largest axis length of the ROI-enclosing ellipsoid and is calculated using the largest principal component λminor.

original_glszm_GrayLevelVariance	GLV measures the variance in gray level intensities for the zones.
original_glszm_LargeAreaEmphasis	LAE is a measure of the distribution of large area size zones, with a greater value indicative of more larger size zones and more coarse textures.
original_ngtdm_Contrast	Contrast is a measure of the spatial intensity change, but is also dependent on the overall gray level dynamic range. Contrast is high when both the dynamic range and the spatial change rate are high, i.e. an image with a large range of gray levels, with large changes between voxels and their neighbourhood.
original_ngtdm_Strength	Strength is a measure of the primitives in an image. Its value is high when the primitives are easily defined and visible, i.e. an image with slow change in intensity but more large coarse differences in gray level intensities.
original_shape_SurfaceArea	To calculate the surface area, first the surface area of each triangle in the mesh is calculated (1). The total surface area is then obtained by taking the sum of all calculated sub-areas (2).

Radiomics (R)-score and Dosiomics (D)-score

We added a constant offset in order to return strictly positive scores.

The R-score was calculated as follows: -1.383 + 1.067*original_glcm_JointEntropy - 0.370*original_ngtdm_Complexity + 1.605*original_shape_Elongation - 0.635*original_shape_Flatness + 0.398*original_shape_Maximum2DDiameterSlice + 1.557* original_shape_MeshVolume - 2.148*original_shape_MinorAxisLength + 4.

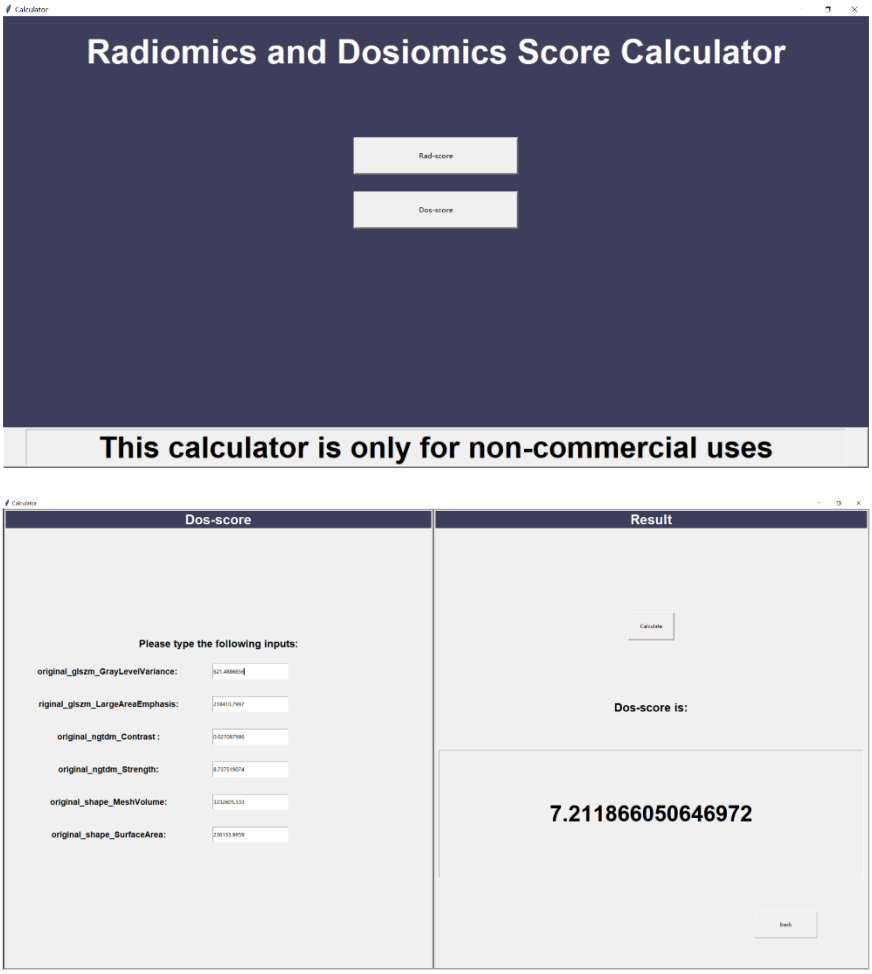
The D-score: -1.522 - 0.616*original_glszm_GrayLevelVariance - 0.868*original_glszm_LargeAreaEmphasis + 0.878*original_ngtdm_Contrast + 0.922*original_ngtdm_Strength + 1.457*original_shape_MeshVolume - 0.625*original_shape_SurfaceArea + 9.

Supplementary material E

Instructions for R-score and D-score calculator

Selecting either Radiomics risk score (R-score) or Dosiomics risk score (D-score), then enter the feature values into the corresponding input boxes and click the “Calculate” button to get the scores.

**This calculator can only be used for research purposes, not for commercial use*.

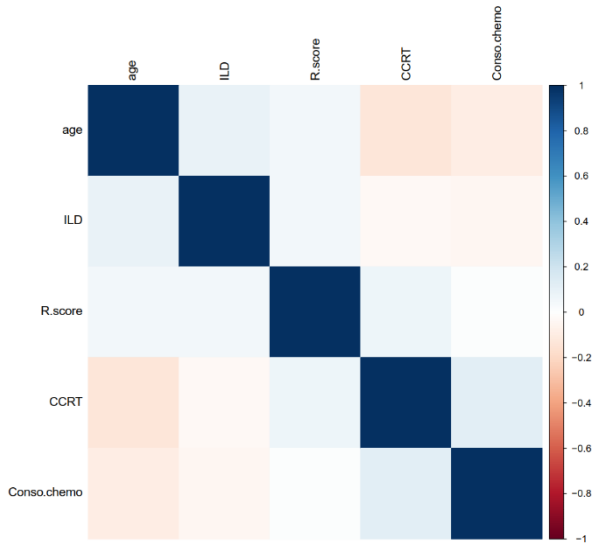


Supplementary Figure 2. The operator interface of the calculator.

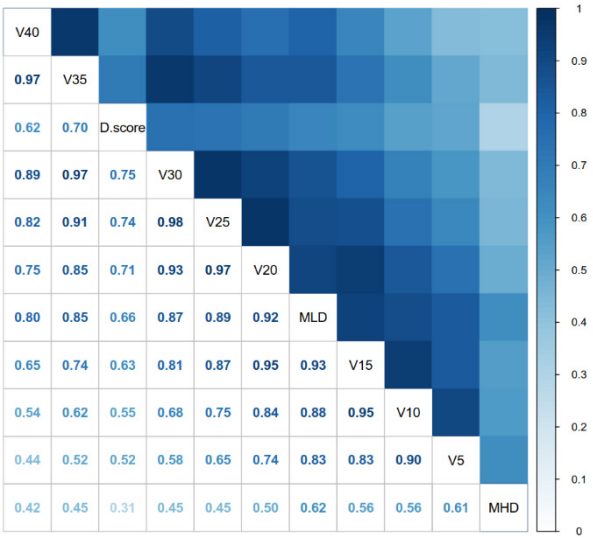
Supplementary material F

Correlations between different parameters

The correlation between the different parameters was calculated (Spearman correlation, R version 4.0.5). The results showed no significant correlation (>0.8) between radiomics risk score (R-score) and clinical parameters, dosiomics risk score (D-score) and dosimetrics.

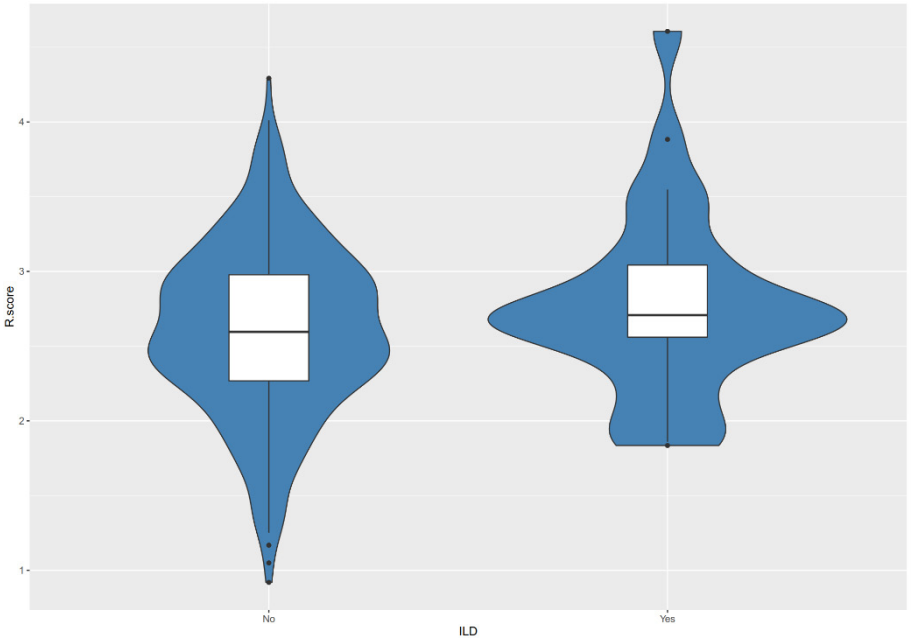


(a)

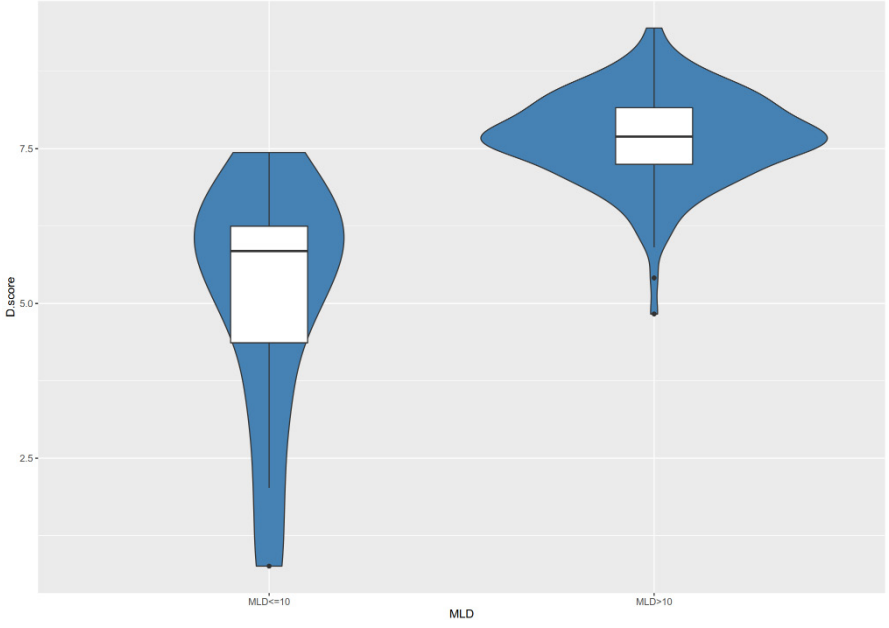


(b)

Supplementary Figure 3. (a) Correlations between R-score and clinical parameters. (b) Correlation between D-score and dosimetrics. *Abbreviations*: CCRT = concurrent chemo-radiotherapy; Conso chemo = consolidation chemotherapy; R-score = radiomics risk score; D-score = dosiomics risk score; MLD = mean lung dose; MHD = mean heart dose.

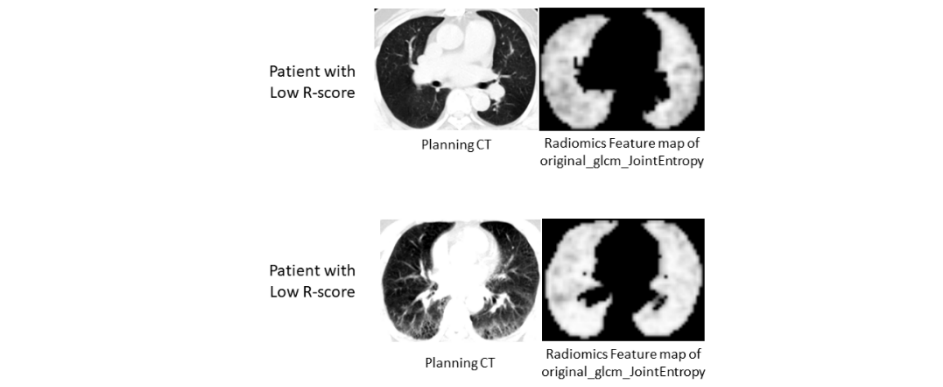


(a)

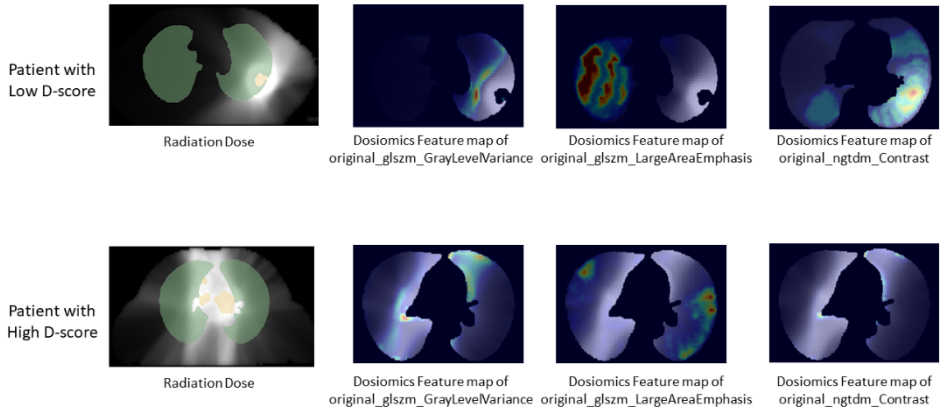


(b)

Supplementary Figure 4. (a) Distribution of radiomics risk score (R-score) in patients with and without interstitial lung disease (ILD). (b) Distribution of dosiomics risk score (D-score) among patients with mean lung dose (MLD) greater than 10Gy and less than or equal to 10Gy.



(a)



(b)

Supplementary Figure 5. (a) Radiomics feature map of feature "original_glcm_JointEntropy" for patient with low radiomics risk score (R-score) and patient with high R-score. (b) Dosiomics feature map of feature "original_glszm_GrayLevelVariance", "original_glszm_LargeAreaEmphasis" and "original_ngtdm_Contrast" for patient with low dosiomics risk score (D-score) and patient with high D-score.

Supplementary material G

Discrimination ability of different combination of Radiomics score, Dosiomics score and clinical parameters

Model	Train (95%CI)	Validation by bootstrapping (95%CI)	Testing (95%CI)
R-score + D-score	0.735 (0.673-0.796)	0.729 (0.720-0.736)	0.739 (0.553-0.926)
R-score + C	0.717 (0.652-0.782)	0.701 (0.683-0.719)	0.771 (0.585-0.962)
D-score + C	0.770 (0.710-0.830)	0.755 (0.744-0.765)	0.756 (0.559-0.954)

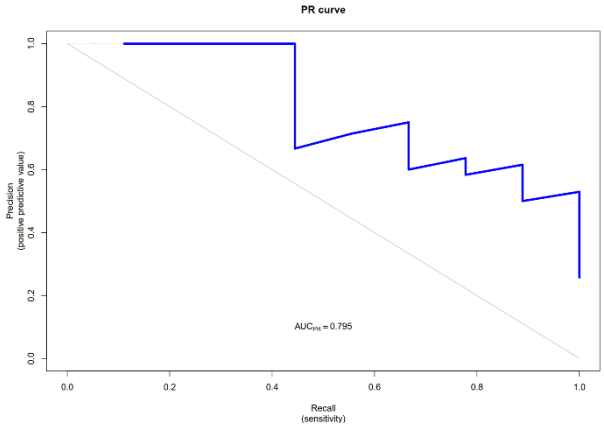
Abbreviations: R = radiomics risk score; D = dosiomics risk score; C = clinical parameters.

Discrimination ability of different models without patients with interstitial lung disease (ILD)

Model	Testing (95%CI)	Testing without patient with ILD (95%CI)
R-score	0.671 (0.558-0.899)	0.714 (0.348-1.000)
D-score	0.684 (0.573-0.883)	0.800 (0.613-0.987)
DVH-score	0.661 (0.551-0.856)	0.752 (0.505-1.000)
Clinical parameters	0.709 (0.509-0.91)	0.629 (0.392-0.865)
R-score + D-score + C	0.855 (0.719-0.990)	0.914 (0.785-1.000)

Abbreviations: ILD = interstitial lung disease.

Precision Recall (RP) curve of the model combing R-score, D-score and Clinical parameters on the test set

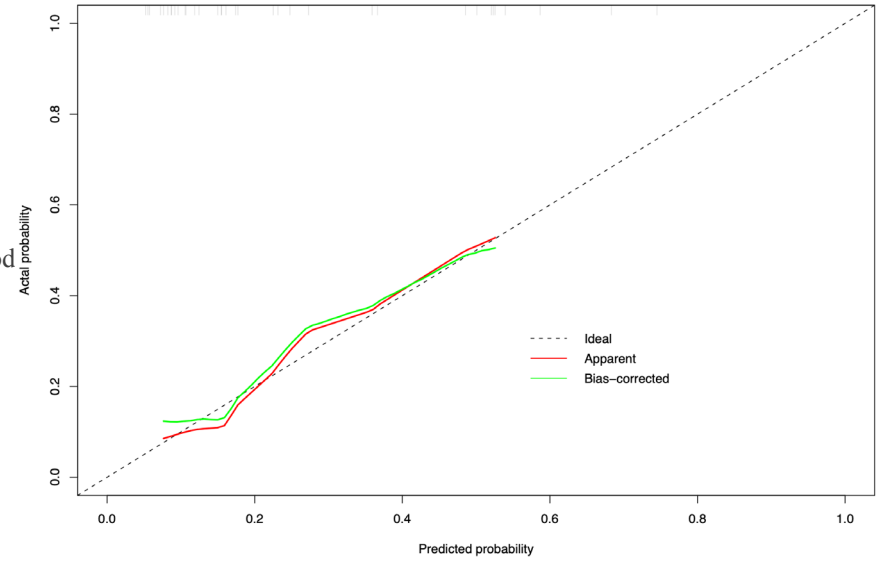


Supplementary Figure 6. Precision Recall (RP)-curve

Supplementary material H

Calibration curve of prospective validation set with a bootstrap resampling meth-

od



Supplementary Figure 7. Calibration curve of prospective validation set with a bootstrap resampling method. Dashed line indicated the ideal model in which predicted and actual probabilities were perfectly identical; Red line indicated actual performance with apparent accuracy; Green line indicated bootstrap corrected estimate of the calibration curve.

Chapter 7: Computed tomography and radiation dose images-based deep-learning model for predicting radiation pneumonitis in lung cancer patients after radiation therapy: A pilot study with external validation

Adapted from: ***Zhen Zhang****; Zhixiang Wang*; Tianchen Luo; Meng Yan; Andre, Dekker; Dirk De Ruysscher; Alberto Traverso; Leonard Wee; Lujun Zhao. Computed tomography and radiation dose images-based deep-learning model for predicting radiation pneumonitis in lung cancer patients after radiation therapy: A pilot study with external validation. Radiotherapy and Oncology, https://doi.org/10.1016/j.radonc.2023.109581.*

** indicates equal contributions*

Abstract

Purpose: To develop a deep learning model that combines CT and radiation dose (RD) images to predict the occurrence of radiation pneumonitis (RP) in lung cancer patients who received radical (chemo)radiotherapy.

Methods: CT, RD images and clinical parameters were obtained from 314 retrospectively-collected patients (training set) and 35 prospectively-collected patients (test-set-1) who were diagnosed with lung cancer and received radical radiotherapy in the dose range of 50 Gy and 70 Gy. Another 194 (60 Gy group, test-set-2) and 158 (74 Gy group, test-set-3) patients from the clinical trial RTOG 0617 were used for external validation. A ResNet architecture was used to develop a prediction model that combines CT and RD features. Thereafter, the CT and RD weights were adjusted by using 40 patients from test-set-2 or 3 to accommodate cohorts with different clinical settings or dose delivery patterns. Visual interpretation was implemented using a gradient-weighted class activation map (grad-CAM) to observe the area of model attention during the prediction process. To improve the usability, ready-to-use online software was developed.

Results: The discriminative ability of a baseline trained model had an AUC of 0.83 for test-set-1, 0.55 for test-set-2, and 0.63 for test-set-3. After adjusting CT and RD weights of the model using a subset of the RTOG-0617 subjects, the discriminatory power of test-set-2 and 3 improved to AUC 0.65 and AUC 0.70, respectively. Grad-CAM showed the regions of interest to the model that contribute to the prediction of RP.

Conclusion: A novel deep learning approach combining CT and RD images can effectively and accurately predict the occurrence of RP, and this model can be adjusted easily to fit new cohorts.

Keywords: Radiotherapy; Radiation pneumonitis; Deep learning; Artificial intelligence; actuarial outcome models

Introduction

Radiation pneumonitis (RP) is a relatively common radiotherapy(RT)-related side effect [1, 2]; estimates of RP vary from 5%-58% [3] but it is challenging to forecast accurately on the individual patient level. The risk of RP constrains the tumoricidal dose that can be prescribed and, in serious instances, may directly threaten the life of the patient. Prediction models of a patient's RP risk are hence an active topic in current research work [4, 5].

Dose-volume histogram (DVH) metrics, such as mean lung dose [6], V5 and V20 [7], are presently in broad clinical use as surrogates for RP risk. Normal tissue control probability (NTCP) can be computed from a DVH of total lungs [8]. These aforementioned DVH indicators do not explicitly account for the spatially heterogeneous distribution of dose in lungs, nor do they account for the functional state of lung parenchymal tissue prior to commencement of RT. Hand-crafted features that describe spatial dose non-uniformity (i.e. "dosiomics") have been recently investigated [9], as were characterization of non-tumour lung tissue via image-based analysis (i.e. "radiomics" and texture) [10-12]. To date, few RP studies have been performed that combine both dosiomics from a clinical treatment plan and radiomics from its corresponding planning CT [13, 14]. These studies have treated the two types of data as disjoint feature domains.

A promising direction for predicting RP is a deeper exploration of inter-related effects of dose and morphology. First, it is supposed that information about the underlying radio-sensitivity of lung tissue might be encoded into CT-based imaging features. Second, that variations in applying RT planning national guidelines leads to divergent spatial dose distributions that are not fully captured in traditional indices such as V20. For example, in China, the constraint V20 not exceed 25% - 30% [15, 16], however National Comprehensive Cancer Network (NCCN) guidelines recommend 35% - 40%. Within a set of DVH constraints, there exists an unlimited number of feasible RT plans that would meet those constraints but result in non-comparable spatial dose distributions in normal lung. Third, it is not entirely clear how to explicitly define hand-crafted measures that combine both CT and dose information into a common feature domain. Last, it remains an open debate about the relative merits of hand-crafted features versus deep-learning features in regard to a given clinical question.

The objective of this study was to develop and evaluate a deep-learning (DL) model to predict RP on the basis of CT intensities and Radiotherapy Dose (RD) distributions, using a joint feature representation for CT attenuation (radiomics) and dose distribution (dosiomics), rather than making an ensemble of separated models. A design criterion was that any such DL-based predictions need to be "adjustable" in a relatively simple way to adapt to alternative prescribed dose and RT planning protocols.

This work describes the implementation a well-known 3D ResNet DL architecture as a generator of "deep features" in the joint CT-RD representation. A fully-connected (FC) network is appended to the end of the ResNet to estimate class probabilities of RP based on deep features. We assumed a linearly-weighted mixture of CT and RD, with tunable weights, as the input. In the event of different prescribed doses or dose planning procedures at different institutions, we assumed that a baseline model has to be subsequently adjusted only for a different mixing ratio of CT and RD, as well as to retrain the FC component to use the new deep features resulting from the alternative mixing. However, the ResNet part will be kept

frozen after training an initial baseline model.

Methods

1. Study design

The overall flow in this study has been illustrated in Figure 1. This study utilizes private data from a single institution to train a baseline model. Subsequent model adjustments and model performance evaluations used a prospectively collected cohort from the same single institution, plus the RTOG-0617 randomized trial dataset [17-20] split into two sets according to the prescribed dose (60Gy in control arm and 72Gy in the experiment arm). Grad-CAM heatmaps were overlaid on the input CT and RD to support clinical interpretation. Model discrimination was reported as receiver-operator "area under the curve" (AUC) and model calibration was assessed as goodness-of-fit for binary classification. The details of each part of the study are as follows.

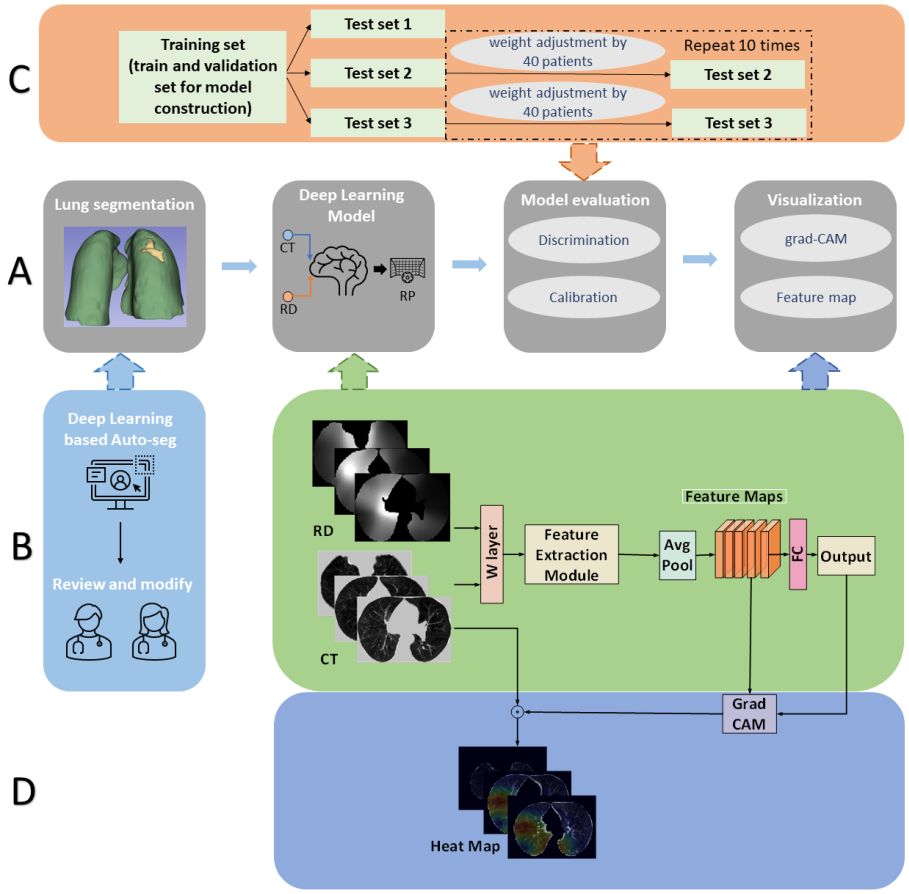


Figure 1. A, The pipeline of this study: lung segmentation, model construction, model evaluation and visualization. B, Lung mask contouring using deep learning based automatic tool

and reviewed and modified by two physicians. And model architecture. C, First, test-set-1, 2, and 3 were used to validate the base model built from training set. Second, forty patients from test-sets-2 and 3 were used to adjust the weight and fully connected layers, respectively, and validated by the test-sets-2 and 3 (without forty patients). This step was repeated ten times. D, Visualization of the model was achieved by the guided gradient weighted class activation mapping.

2. Study population

All patients included this study had confirmed diagnosis of lung cancer and were treated with radically-intended radiotherapy (IMRT or VMAT), either with or without concurrent chemotherapy. The primary endpoint was symptomatic RP grade 2 or higher according to Common Terminology Criteria for Adverse Events (CTCAE) v4.0. In the private institutional datasets, the presence (or absence) of RP was assessed by experienced radiation oncologists based on follow-up CT, blood test and symptoms. In the RTOG dataset, the status of RP was documented in individual case reports. In this work, we considered only RP events which occurred anytime from the last fraction of radiotherapy up to 6 months after the last fraction of radiotherapy, as specifically RT-treatment induced RP.

The Training set consisted of 314 routine care patients retrospectively extracted from archives at one medical university cancer hospital. These were primarily intended for treatment with 60 Gy, but a range of delivered doses between 50 Gy to 70 Gy was prescribed at the treating physicians discretion. Test-set-1 comprised of 35 prospectively registered patients from the same institution, also predominantly 60 Gy total intended dose, with variations of delivered dose at treating physician's discretion. Training set and Test-set-1 were obtained with approval from an internal review board (ref. IRBbc2021135). The discretionary deviations in delivered dose were based on each patient's overall physical condition and best achievable normal tissue constraints. Specific details of Training set and Test-set-1 are provided in Supplementary Materials 1A.

Access for secondary re-use of data from the prospectively randomized controlled trial RTOG-0617 was obtained through the trial sponsor. From the control arm (60 Gy prescribed dose), 194 subjects were defined as Test-set-2, and from the intervention arm (74 Gy prescribed dose), 158 subjects were allocated as Test-set-3. Specific details for filtering the RTOG-0617 subjects are provided in Supplementary Materials 1B.

3. Data preparation

Planning CT and RD were originally extracted in DICOM format for all subjects. The voxel-wise values in the RD images were scaled to represent absolute physical dose in units of Gy. We used a deep-learning automatic lung contouring tool based on previous work [22] to automatically segment whole lungs. Experienced radiation oncologists (ZZ and MY) inspected and (where needed) manually corrected the auto-generated lung masks to ensure accuracy and segmentation consistency. Data preparation and preprocessing steps are described in Supplementary Materials 2A.

4. Development of deep learning RP models

A 3D ResNet architecture was implemented as the main backbone of the RP model (see technical schematic in Supplementary Materials Figure S1). In brief, the pre-processed CT

and RD arrays of the same dimensions were passed to the ResNet via linear mixing (W layer) immediately followed by a 7x7 convolution layer. In the W layer, we defined the composite input source as , where . A and B were thus the mixing ratio of CT and RD, such that A was always fixed at unity. Values of A and B were tuned as part of the model training process and were determined by back-propagation of the error.

The ResNet was used as an image-based "deep feature" generator; its weights were determined by training an initial baseline model and thereafter the entire ResNet weights were frozen. The RP classification model consisted of average pooling and a fully-connected (FC) layer at the end, which uses the deep feature maps generated by the ResNet in order to compute a class probability of RP at a sigmoid function layer. A purely binary classification (RP or non-RP) was computed by applying a threshold of 0.5. The core of the ResNet comprised eight repeating residual blocks containing convolution (conv), batch normalization (BN) and Rectified Linear Unit (ReLU) activation. We used an Adam optimizer with a learning rate of 0.0001 and Binary Cross-Entropy as the loss function. The training strategy, loss function definition and model tuning hyperparameters are shown in Supplementary Materials 2D.

After training using exclusively the Training set, the baseline model was evaluated in each of the three hitherto unseen cohorts i.e., Test-set-1 (medical university cancer hospital, 60Gy prescribed), Test-set-2 (RTOG-0617 control arm, 60Gy prescribed) and Test-set-3 (RTOG-0617 experiment arm, 74Gy prescribed).

To examine the feasibility of "adjusting" the model for the same nominal prescribed dose but different planning protocol, we attempted two related experiments. First, we randomly chose 40 subjects from Test-set-2 without replacement and then proceeded to re-train only the CT-RD mixing ratio (i.e., the W layer) and the FC classifier – the ResNet was kept frozen as abovementioned.

As cross-validation, we evaluated the adjusted model using the remainder of Test-set-2 subjects (hereafter, Test-set-2* = the initial 194 subjects minus the 40 selected for adjustment = 154). To check for random vagaries of selecting 40 patients, we repeated the entire experiment 10 times, each time choosing different subsets of 40 patients. Secondly, to see if there was added value of using more patients, we adjusted the baseline model using all 194 subjects prescribed to 60Gy in the RTOG control arm. However, it is no longer possible to check for over-optimism using repeated cross-validation, so 1000 times bootstrapping with replacement from Test-set-2 (hereafter Test-set-2#) was used to estimate a range of validation results.

To examine the feasibility of "adjusting" the model for simultaneously different prescribed dose and different planning protocol, we re-did the two related experiments above only utilizing Test-set-3.

To help visualize imaging and dose features that influence RP/non-RP prediction, and thus assist with clinical interpretation of the model attention area, activation heatmaps were generated by back-projecting Grad-CAM values as overlay on the planning CT and dose images (see detail in Supplementary Materials 2C).

5. Comparator RP models as alternatives to deep learning of mixed CT and RD models

A model employing only CT images and a model employing only RD images were constructed with the same process as the combined model described above for comparison.

We compared the aforementioned models against simple logistic regression based either on (i) dose-volume histograms (DVH) only, or (ii) clinical parameters only, or (iii) a combination of DVH and clinical parameters. Due to the high degree of correlation that is well-known in DVH metrics, we only considered V20 and mean lung dose (MLD) in the DVH-based model. For the clinical model, the patient age and presence of interstitial lung abnormalities were selected according to Supplementary Material Table S1. Further detailed information on the construction of the DVH model and clinical parameters model are provided in Supplementary Materials 5B.

6. Statistical analysis

The discrimination performance of the model was quantified using area under the receiver-operator curve (AUC), accuracy, sensitivity, and specificity of RP prediction. For all performance metrics reported, we estimated 95% confidence intervals by 1000 times bootstrapping. Goodness-of-fit was tested by calculating the model calibration error [23, 24].

Patients' baseline characteristics for continuous variables are presented as mean ± standard deviation. For univariate analysis of clinical parameters, Pearson chi-squared tests and exact Fisher tests were used for categorical variables and logistic regression for continuous variables. For significance of clinical factors, a two-sided hypothesis test at the a = 0.05 confidence level was assumed. Clinical and DVH data were analyzed in the Statistical Package for Social Science program (SPSS for Windows, version 27.0; SPSS Inc, Chicago, IL). All deep learning models were constructed and test set performance assessed using Python (version 3.8.5) and R software (version 4.0.5), respectively.

7. Code and data availability

Code packages and libraries for constructing our deep learning models are given in Supplementary Materials 2. The source code is made open access at https://gitlab.com/w654053334/rp_prediction.

The RTOG trial dataset may be obtained by contacting the sponsors for secondary re-use of data. Training set and Test-set-1 are private institutional collections, which may be made available to other researchers upon reasonable request and subject to data sharing agreements – please contact the corresponding author. To assist readers with using our RP model, we have prepared an open access online version with user interface (see Supplementary Materials 3).

Results

The characteristics of patients are shown in Table 1. Statistically significant heterogeneity between groups was observed across the majority of clinical factors, except for age and smoking. In Table 2, the clinical factors were grouped by RP versus non-RP. In univariate analysis, age, planning tumor volume (PTV), volume of the lung receiving 5 Gy (V5_lung) and 20 Gy (V20_lung), and mean lung dose (MLD) were each statistically significantly higher in patients with RP versus non-RP. Additional detailed clinical characteristics in the four datasets are given in Supplementary materials 1C (Table S1-3).

Table 1. Patient characteristics in Training set, Test-set-1, 2, and 3.

Characteristics	Training set (n=314) Mean ± SD	Test set 1 (n=35) Mean ± SD	Test set 2 (n=194) Mean ± SD	Test set 3 (n=158) Mean ± SD	P-value
Age	61 (30-85)	62 (34-75)	64 (37-82)	63 (41-82)	0.243
Gender					<0.001
Male	238 (75.8%)	23 (65.7%)	115 (59.3%)	89 (56.3%)	
Female	76 (24.2%)	12 (34.3%)	79 (40.7%)	69 (43.7%)	
Smoking					<0.001
Yes	71 (22.6%)	9 (25.7%)	14 (7.2%)	11 (7.0%)	
No	241 (76.8%)	26 (74.3%)	167 (86.1%)	144 (91.1%)	
Unknow	2 (0.6%)	0	13 (6.7%)	3 (1.9%)	
Histology					<0.001
LUSC	84 (26.8%)	8 (22.9%)	75 (38.7%)	70 (44.3%)	
LUAD	75 (23.9%)	10 (28.6%)	86 (44.3%)	63 (39.9%)	
LCU	——	——	4 (2.1%)	1 (0.6%)	
NOS	——	——	29 (14.9%)	24 (15.2%)	
SCLC	155 (49.4%)	17 (48.6%)	——	——	
Rt_technique					<0.001
3D-CRT	——	——	115 (59.3%)	81 (51.3%)	
IMRT	87 (27.7%)	5 (14.3%)	79 (40.7%)	77 (48.7%)	
VMAT	227 (72.3%)	30 (85.7%)	——	——	
Conso chemo					<0.001
Yes	179 (57.0%)	19 (54.3%)	173 (89.2%)	136 (86.1%)	
No	135 (43.0%)	16 (45.7%)	21 (10.8%)	22 (13.9%)	
PTV (cc)	446.82±188.51	417.72±179.70	507.93±273.31	482.66±261.40	0.014
V5_lung (%)	48.80±10.15	48.82±10.83	57.68±15.29	57.11±14.65	<0.001
V20_lung (%)	24.43±5.24	24.06±4.90	29.06±7.47	31.22±7.96	<0.001
MLD (Gy)	13.37±2.62	13.06±2.61	16.66±4.15	19.16±4.55	<0.001

Abbreviations: Pts = patients; LUSC = lung squamous cell carcinoma; LUAD = lung adenocarcinoma; LCU= Large cell undifferentiated; NOS= Non-small cell lung cancer; SCLC = small cell lung cancer; Rt_technique = radiotherapy technique used to treat patient; 3D-CRT=3dimensional comformal radiation therapy; IMRT = intensity-modulated radiotherapy; VMAT = volumetric modulated arc therapy; chemo = chemotherapy; Conso chemo = consolidation chemotherapy; PTV = planning tumor volume; V5_lung= Lung V5 (%);V20_lung= Lung V20 (%); MLD = Mean lung dose (Gy).

Table 2. Patient characteristics group according to outcome of RP or without RP.

Characteristics	Without RP (n=565) Mean ± SD	With RP (n=136) Mean ± SD	P-value
Age median, range (years)	62 (30-86)	65 (38-80)	0.004
Gender			0.170
Male	368 (65.1%)	97 (71.3%)	
Female	197 (34.9%)	39 (28.7%)	
Smoking			0.229
Yes	464 (82.1%)	114 (83.8%)	
No	84 (14.9%)	21 (15.4%)	
Unknow	17 (3.0%)	1 (0.7%)	
Histology			0.926
LUSC	189 (33.5%)	48 (35.3%)	
LUAD	191 (33.8%)	43 (31.6%)	
LCU	5 (0.9%)	0	
NOS	43 (7.6%)	10 (7.4%)	
SCLC	137 (24.2%)	35 (25.7%)	
Histology			0.778
LUSC	189 (33.5%)	48 (35.3%)	
NSC-NSCLC	239 (42.3%)	53 (39.0%)	
SCLC	137 (24.2%)	35 (25.7%)	
Rt_technique			0.620
3D-CRT	162 (28.7%)	34 (25.0%)	
IMRT	200 (35.4%)	48 (35.3%)	
VMAT	203 (35.9%)	54 (39.7%)	
Conso chemo			0.046
Yes	418 (74.0%)	89 (65.4%)	
No	147 (26.0%)	47 (34.6%)	
PTV (cc)	459.54±226.35	515.30±253.67	0.013
V5_lung (%)	52.50±13.66	55.72±12.70	0.013
V20_lung (%)	26.91±7.22	28.52±6.92	0.020
MLD (Gy)	15.42±4.37	16.21±3.99	0.056

Abbreviations: Pts = patients; LUSC = lung squamous cell carcinoma; LUAD = lung adenocarcinoma; LCU= Large cell undifferentiated; NOS= Non-small cell lung cancer; SCLC = small cell lung cancer; Rt_technique = radiotherapy technique used to treat patient; 3D-CRT=3dimensional comformal radiation therapy; IMRT = intensity-modulated radiotherapy; VMAT = volumetric modulated arc therapy; chemo = chemotherapy; Conso

chemo = consolidation chemotherapy; PTV = planning tumor volume; V5_lung= Lung V5 (%);V20_lung= Lung V20 (%); MLD = Mean lung dose (Gy).

The predictive performance of models for RP is summarized in Table 3. The baseline model performed well on Test-set-1 (AUC 0.83) compared to Test-set-2 (AUC 0.55) and Test-set-3 (AUC 0.63). However, after adjustment, model discrimination was improved in Test-set-2* (AUC 0.65) and Test-set-3* (AUC 0.70), respectively. The discrimination metrics of using only a subset of forty patients to adjust model were close to using the entire dataset, with small differences in AUC of 0.03 and 0.01, respectively. The accuracy, sensitivity and specificity largely followed the same pattern of findings as for AUC.

Table 3. Performance of baseline model and adjustments (using 60Gy and 74Gy arms of RTOG-0617 trial).

Model	Adjustment	Evaluation	AUC (95%CI)	Accuracy (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)
Baseline model	No adjustment	Test-set-1 (35)	0.83 (0.82-0.91)	0.82 (0.75-0.81)	0.70 (0.67-0.77)	0.88 (0.83-0.89)
		Test-set-2 (194)	0.55 (0.47-0.69)	0.70 (0.57-0.82)	0.41 (0.39-0.52)	0.69 (0.61-0.84)
		Test-set-3 (158)	0.63 (0.53-0.72)	0.66 (0.60-0.73)	0.60 (0.46-0.74)	0.68 (0.61-0.75)
Adjusted for RTOG 60 Gy arm	40 randomly selected from Test-set-2	Test-set-2* (154)	0.65 (0.54-0.77)	0.76 (0.63-0.91)	0.58 (0.48-0.83)	0.70 (0.66-0.98)
Adjusted for RTOG 60 Gy arm	All subjects from Test-set-2	Test-set-2# (194 bootstrap samples)	0.68 (0.58-0.83)	0.78 (0.80-0.89)	0.77 (0.62-0.97)	0.65 (0.60-0.74)
Adjusted for RTOG 74 Gy arm	40 randomly selected from Test-set-3	Test-set-3* (118)	0.70 (0.63-0.76)	0.71 (0.63-0.83)	0.62 (0.56-0.86)	0.73 (0.67-0.95)
Adjusted for RTOG 74 Gy arm	All subjects from Test-set-3	Test-set-3# (158 bootstrap samples)	0.71 (0.62-0.81)	0.78 (0.73-0.84)	0.68 (0.54-0.83)	0.77 (0.71-0.82)

Abbreviations: AUC = area under receiver operating characteristic curve; 95% CI = 95% confidence interval; * the asterisk indicates that the coefficients of CT and RD for this model are adjusted with 40 patients for each set; # The pound symbol indicates that the coefficients of CT and RD for this model are adjusted using the entire data set. The number in parentheses are the sample size for the evaluations.

The mixing ratio, i.e. A and B weights, for CT and RD from each model were summarized in Supplementary Materials 4 Table S4. Across the baseline model and its subsequent adjustments, RD was overall more important than CT from an RP prediction perspective. Among nominally 60Gy subjects, the post-training weight of RD relative to CT was reasonably stable around 1.5 (range 1.42-1.67). Among nominally 74Gy subjects, the relative weight of RD to CT was suppressed to about 1.2 (range 1.20-1.25).

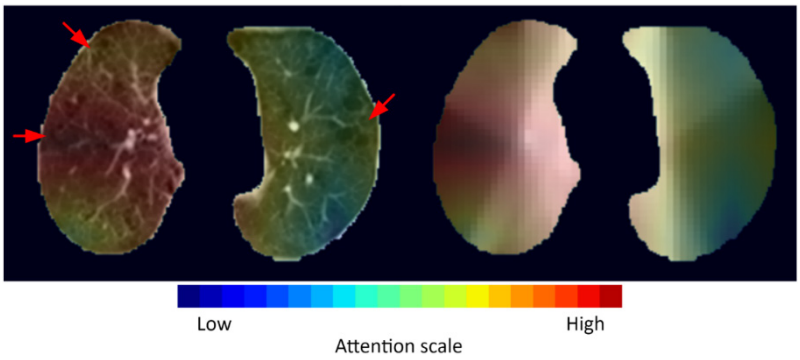
For comparison with the baseline model that included CT and RD, an alternative baseline model was constructed by either CT or RD alone and then tested on Test-set-1. The original baseline model (with both CT and RD, AUC 0.83) performed better than either CT-only or RD-only alternatives (AUC 0.63 for CT and 0.69 for RD, additionally accuracy, sensitivity and specificity were reported in Supplementary Materials 5A Table S5).

The discrimination of the DVH-based logistic model was poorer than that of the RD-only deep learning model (AUC 0.66 vs. 0.69) when evaluated in Test-set-1, and both were markedly poorer than the baseline model results. Discrimination of the logistic regression model based on clinical parameters (AUC 0.71 in Test-set-1) was poorer than the baseline model, but was slightly better than either of the RD-only deep learning and the DVH-only logistic regression.

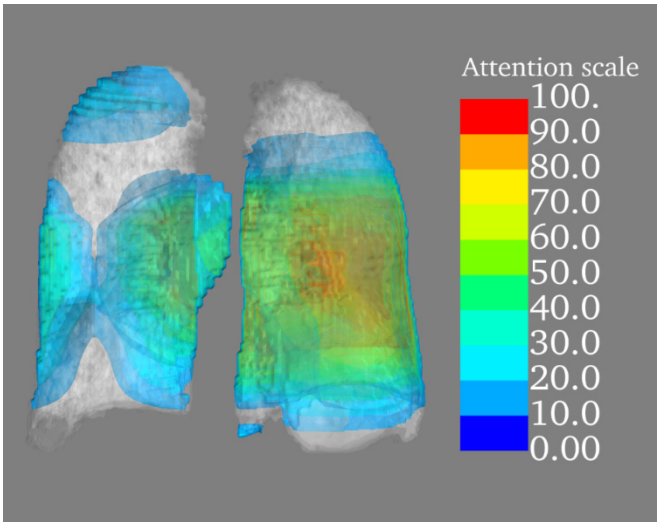
The calibration error of the baseline model was 0.07 in Test-set-1, 0.22 in Test-set-2, and 0.18 in Test-set-3, indicating that there was no major calibration issue. However, after the model adjustment, the average expected calibration error was reduced to 0.14 for Test-set-2, and 0.13 for Test-set-3.

Some representative examples of 3D (Supplementary Materials Video) and 2D heatmaps (Figure 2 and Supplementary Figure S11-13) generated by Grad-CAM may help to illustrate the global view for the whole lung and detailed view of each slice, respectively. In patients with pre-existing lung disease (the area indicated by the pointer in Figure 2 and Supplementary Figure S11), such as interstitial lung abnormalities or emphysema, model attention appears more widely dispersed overall in the lungs. In contrast, for patients without pre-existing lung disease, relatively narrow distribution of model attention has been observed that follows the distribution of dose in the RD (Supplementary Figure S13). This clearly shows that, as far as the prediction of RP goes, a good model needs to be trained that can make use of (CT) features associated with pre-existing lung disease as well as (RD) features related to prominent dose distribution in the normal lungs.

Representative feature maps extracted from each residual block of the RP and non-RP cases are shown in Supplementary Materials 2E (Figure S2-9). As the level deepens in the model, the extracted features become more complex and abstract. While these features maps are very important since the FC layer uses these ResNet-generated feature maps to estimate the probability of RP, it nonetheless remains challenging to interpret the feature maps and thus visually associate them with clinically meaningful features. Thus, in this respect, the grad-CAM heatmaps overlaid onto the CT and RD might be potentially more useful by way of clinical interpretation.



A



B

Figure 2. Illustration of attention (heat) map of a 60-year-old male with non-small cell lung cancer. A, Two-dimensional attention map. The left image is an overlay of the CT image and the attention map, with blue to red representing increasing levels of importance (attention scale). The area of interstitial lung abnormalities indicated by pointers. The image on the right is an overlay of the radiation dose (RD) image and the attention map. From dark to light represents low to high dose, and from blue to red represents increasing importance. B, Three-dimensional attention map. The different colors represent different levels of importance (attention scale).

Discussion

In this study, we used pre-treatment radiotherapy planning CT and planned radiation dose distribution to build a ResNet-based deep learning model to predict RP. The baseline model is trained using a joint representation of features from CT and RD, which we implemented using a linear mixing method of the intensity/dose magnitudes. We then showed that such

a baseline model can be subsequently adjusted by only re-training the mixing ratio (i.e., the W layer) and the FC classifier for RP, at the start and at the end of the ResNet, respectively, without changing any other weights in the ResNet feature extractor itself.

The combination of CT and RD predicted RP reasonably well in Test-set-1, which was expected since the test set most closely resembled the Training set in terms of prescribed dose, RT planning procedure and race cohort. Model performance and model calibration on the RTOG-0617 datasets, i.e., Test-set-2 and Test-set-3 were overall improved after adjusting the baseline model with either some or all of the each dataset.

However, the adjusted model did not perform as well on either of the RTOG-0617 subsets as it performed on Test-set-1. We hypothesize this is because RTOG-0617 data was contributed unevenly across 185 institutions [17], which may leave a large amount of heterogeneity among patients as well as residual differences between scanners, physicians delineations and RT planners that the trial protocol could not reconcile, as one can see in Table 1. It was interesting that the baseline model initially performed better in Test-set-3 (74Gy) with higher AUC and sensitivity compared to Test-set-2 (60Gy), which should have been closer to the prescription setting of the training dataset. However, we cannot rule out random chance since the baseline model initially performed sub-optimally for both Test-set-2 and Test-set-3. This may also suggest that treatment delivery modality may not be the critical factor for the model, at least relative to lung tissue and dose hotspots, and other sources of clinical heterogeneity may be more important. We are unable to resolve this question at present, and resolution of such questions needs more detailed study.

Grad-CAM heatmaps overlaid onto CT and RD suggested synergistic information for the prediction of RP, that is, the influential features point towards pre-existing lung injury in CT and regions of high dose in normal lung. Moreover, we proposed a computationally simplified way to adjust the model to fit different clinical settings. We suggest this a feasible method to adapt to different dose groups and planning protocols. However, it must be noted that even this limited adjustment-based retraining is still more computationally intensive than retraining a conventional machine learning model from scratch; as such, it is presently computationally unfeasible to perform more than a dozen repetitions of cross-validation or bootstraps during training.

This study included a retrospective single-institutional dataset as training set, and three other cohorts to evaluate the performance of our model. All test sets were prospectively collected to ensure the best available accuracy of registering the primary outcome of RP. In clinical practice, an RP event needs to be diagnosed by following up patients' symptoms and examinations. To distinguish RP from other types of pneumonia, follow-up CT examinations, routine blood tests, and C-reactive protein may be used. The endpoint of this study is grade 2 or higher RP, because patients with grade 2 RP require medical intervention and their activities of daily living are affected.

In this study, As mentioned, the relative importance of RD relative to CT was about 1.5 in most cases, except for the 74Gy Test-set-3 where it appeared suppressed to about 1.2. A possible reason for this is that the standard dose (60Gy) can induce RP in patients with intrinsic lung susceptibility to RP, but increasing prescribed dose to 74Gy seems not to be additionally effective at inducing RP. Although the method proposed in this study is potentially an efficient way to update the baseline model for a new clinical setting, it is still possi-

ble to obtain a biased dataset with randomly sampling 40 patients [27], therefore if a larger dataset may be used for adjustment, we expect the model will be more robust.

Some interesting points were found based on the attention maps (Figure 2 and Supplementary Materials 6 Figure S11-13), where we tried to understand the "diagnostic" logic of the model. The results of this study are consistent with previous insights [28-30] and based on our data set, interstitial lung abnormality is an influential factor for the occurrence of RP. In the future, as the sample size expands, the model based only on patients with interstitial lung abnormalities can be developed and compared with the model developed in this study by Grad-CAM approach. Another feature is that the attentional areas tend to be located more in the central area of the lungs than in the peripheral areas. We speculate that there are two reasons for this phenomenon. First, RD is denser in the central part because of the irradiation of metastatic lymph nodes [31, 32]. Secondly, the dose received by the heart may be another factor in the development of RP [33, 34]. Krafft et al. found that cardiac DVH metrics improved the predictive power of radiomics models for RP prediction [12]. In our previous study, cardiac comorbidity was also found to be an independent predictor of RP [35].

Based on these observations, we speculate that the predictive logic of the model may be as follows: for patients with pre-existing lung disease, which was determined in collaboration with radiologists, the model pays attention to lung tissue with disease and analyzes these areas in conjunction with RD distribution. For patients with overall good (no lung disease) status, the model preferentially pays attention to regions of high dose and predicts RP mainly using the RD features. For most patients, the central part of the lung and the regions adjacent to the heart are more important than the peripheral lung. We also compared an RD-only deep learning model with the DVH-based model, which is another commonly used model in clinical practice. From the results, the predictive power of the DVH-based model is not better than that of the RD-based deep learning model.

The result of this study has a few real-world clinical implications. In this study, we did not iteratively tune the decision threshold of the model. In practice, we may select the thresholds that prioritize either higher sensitivity or higher specificity, but we could not do both. Patients with very low probability of RP could receive standard or adequate doses if adjusted models with high specificity were used for these hospitals, which might improve their prognosis [1, 2]. For patients with a very high probability of RP, physicians can give these patients more frequent examinations or preventive medications to lower the grade of RP or prevent it from occurring [36]. Alternatively, this clinical tool may be of assistance during the doctor-patient consultation about risks and expectations of treatment.

There were several limitations in this study. The deep learning model with complex neural networks needs a large dataset to avoid overfitting. We included 701 patients in this study and although, to best of our knowledge, this is the largest dataset on the topic of artificial intelligence model to predict RP, model development will benefit further from even larger datasets including heterogeneity of CT scanners, dose planning systems, etc. different institutions with improvements expected both in terms of performance and in terms of generalizability across backgrounds, scanners, treatment strategies and patients. Second, this model did not include combinations of clinical parameters including cytokines. Our previous studies and others have demonstrated that it has predictive value for RP [37, 38]. The combination of cytokines could improve the performance of the model [39], however, the present

aim of our study was to focus on a non-invasive approach to modeling and therefore cytokines were not included. A potential benefit is that this model can be directly embedded into RT planning systems, as it only needs CT and RD information and can export its predictions directly to other systems for clinical decision support. In addition, patients included in this study did not receive concurrent chemotherapy with the same regimen, and we think that the predictive power of the model could be improved if clinical factors were harmonized. On the reverse side, it is difficult to maintain the same treatment regimen everywhere in the world, and the generalizability of the model would be affected if only patients receiving the same chemotherapy regimen were included.

Finally, we did not include patients who received immunotherapy, which is already a standard therapy for local advanced lung cancer patients now. And the incidence of pneumonitis is higher with the addition of durvalumab after concurrent chemo-radiotherapy [40, 41]. There are still challenges to be addressed before including patients receiving immunotherapy in the analysis, such as differential diagnosis of immune checkpoint inhibitor therapy-related pneumonitis and RP and datasets containing large sample sizes of patients receiving immunotherapy. The model we developed in this study can served as a base (pre-trained) model for future studies that include patients receiving immunotherapy [42].

In summary, we successfully developed a deep learning model to predict RP, and this model can be adjusted easily to fit new cohorts. We tried to uncover the model prediction logic by a visualization approach. In addition, a ready-to-use online software was developed to assist clinical practice. Despite several limitations, we believe that deep learning algorithm possesses great potential to sever as a clinical assistant tool.

References

1. Luo H-S, Huang H-C, Lin L-X. Effect of modern high-dose versus standard-dose radiation in definitive concurrent chemo-radiotherapy on outcome of esophageal squamous cell cancer: a meta-analysis. Radiation Oncology. 2019;14:178. doi:10.1186/s13014-019-1386-x.

2. Ladbury CJ, Rusthoven CG, Camidge DR, Kavanagh BD, Nath SK. Impact of Radiation Dose to the Host Immune System on Tumor Control and Survival for Stage III Non-Small Cell Lung Cancer Treated with Definitive Radiation Therapy. International Journal of Radiation Oncology*Biology*Physics. 2019;105:346-55. doi:10.1016/j.ijrobp.2019.05.064.

3. Arroyo-Hernández M, Maldonado F, Lozano-Ruiz F, Muñoz-Montaño W, Nuñez-Baez M, Arrieta O. Radiation-induced lung injury: current evidence. BMC Pulm Med. 2021;21:9. doi:10.1186/s12890-020-01376-4.

4. Ullah T, Patel H, Pena GM, Shah R, Fein AM. A contemporary review of radiation pneumonitis. Curr Opin Pulm Med. 2020;26:321-5. doi:10.1097/MCP.0000000000000682.

5. Käsmann L, Dietrich A, Staab-Weijnitz CA, Manapov F, Behr J, Rimner A, et al. Radiation-induced lung toxicity - cellular and molecular mechanisms of pathogenesis, management, and literature review. Radiation Oncology (London, England). 2020;15:214. doi:10.1186/s13014-020-01654-9.

6. Liu Y, Wang W, Shiue K, Yao H, Cerra-Franco A, Shapiro RH, et al. Risk factors for symptomatic radiation pneumonitis after stereotactic body radiation therapy (SBRT) in patients with non-small cell lung cancer. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;156:231-8. doi:10.1016/j.radonc.2020.10.015.

7. Saha A, Beasley M, Hatton N, Dickinson P, Franks K, Clarke K, et al. Clinical and dosimetric predictors of radiation pneumonitis in early-stage lung cancer treated with Stereotactic Ablative radiotherapy (SABR) - An analysis of UK's largest cohort of lung SABR patients. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;156:153-9. doi:10.1016/j.radonc.2020.12.015.

8. Prasanna PG, Rawojc K, Guha C, Buchsbaum JC, Miszczyk JU, Coleman CN. Normal Tissue Injury Induced by Photon and Proton Therapies: Gaps and Opportunities. International Journal of Radiation Oncology, Biology, Physics. 2021;110:1325-40. doi:10.1016/j.ijrobp.2021.02.043.

9. Bourbonne V, Da-Ano R, Jaouen V, Lucia F, Dissaux G, Bert J, et al. Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;155:144-50. doi:10.1016/j.radonc.2020.10.040.

10. Wang L, Gao Z, Li C, Sun L, Li J, Yu J, et al. Computed tomography-based delta-radiomics analysis for discriminating radiation pneumonitis in patients with esophageal cancer after radiation therapy. International Journal of Radiation Oncology, Biology, Physics. 2021. doi:10.1016/j.ijrobp.2021.04.047.

11. Du F, Tang N, Cui Y, Wang W, Zhang Y, Li Z, et al. A Novel Nomogram Model Based on Cone-Beam CT Radiomics Analysis Technology for Predicting Radiation Pneumonitis in Esophageal Cancer Patients Undergoing Radiotherapy. Front Oncol. 2020;10:596013. doi:10.3389/fonc.2020.596013.

12. Krafft SP, Rao A, Stingo F, Briere TM, Court LE, Liao Z, et al. The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis. Med Phys. 2018;45:5317-24. doi:10.1002/mp.13150.

13. Puttanawarut C, Sirirutbunkajorn N, Tawong N, Jiarpinitnun C, Khachonkham S, Pattaranutaporn P, et al. Radiomic and Dosiomic Features for the Prediction of Radiation Pneumonitis Across Esophageal Cancer and Lung Cancer. Front Oncol. 2022;12.

14. Zhang Z, Wang Z, Yan M, Yu J, Dekker A, Zhao L, et al. Radiomics and dosiomics signature from whole lung predicts radiation pneumonitis: a model development study with prospective external validation and decision-curve analysis. International Journal of Radiation Oncology, Biology, Physics. 2022:S0360-3016(22)03189-3. doi:10.1016/j.ijrobp.2022.08.047.

15. Wei J, Zhang Z, Yu J, Jia H, Tian J, Meng C, et al. Meta-analysis of the incidence of radiation pneumonitis between European, American and Asian populations. Chinese Journal of Radiation Oncology. 2021;30:556-62. doi:10.3760/cma.j.cn113030-20201114-00554.

16. Liu Z, Liu W, Ji K, Wang P, Wang X, Zhao L. Simultaneous integrated dose reduction intensity-modulated radiotherapy applied to an elective nodal area of limited-stage small-cell lung cancer. Experimental and Therapeutic Medicine. 2015;10:2083-7. doi:10.3892/etm.2015.2835.

17. Bradley JD, Paulus R, Komaki R, Masters G, Blumenschein G, Schild S, et al. Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study. Lancet Oncol. 2015;16:187-99. doi:10.1016/S1470-2045(14)71207-0.

18. Bradley JD, Hu C, Komaki RR, Masters GA, Blumenschein GR, Schild SE, et al. Long-Term Results of NRG Oncology RTOG 0617: Standard- Versus High-Dose Chemoradiotherapy With or Without Cetuximab for Unresectable Stage III Non-Small-Cell Lung Cancer. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2020;38:706-14. doi:10.1200/JCO.19.01162.

19. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of Digital Imaging. 2013;26:1045-57. doi:10.1007/s10278-013-9622-7.

20. Bradley J, Forster K. Data from NSCLC-Cetuximab. The Cancer Imaging Archive. 2018. doi:10.7937/TCIA.2018.jze75u7v.

21. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magnetic Resonance Imaging. 2012;30:1323-41. doi:10.1016/j.mri.2012.05.001.

22. Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. European Radiology Experimental. 2020;4:50. doi:10.1186/s41747-020-00173-2.

23. Shi Z, Zhang Z, Liu Z, Zhao L, Ye Z, Dekker A, et al. Methodological quality of machine learning-based quantitative imaging analysis studies in esophageal cancer: a systematic review of clinical outcome prediction after concurrent chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2021. doi:10.1007/s00259-021-05658-9.

24. Naeini MP, Cooper GF, Hauskrecht M. Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proc Conf AAAI Artif Intell. 2015;2015:2901-7.

25. Vial A, Stirling D, Field M, Ros M, Ritz C, Carolan M, et al. The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. Translational Cancer Research. 2018;7. doi:10.21037/21823.

26. Georgiou T, Liu Y, Chen W, Lew M. A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision. Int J Multimed Info Retr. 2020;9:135-70. doi:10.1007/s13735-019-00183-w.

27. Shahinfar S, Meek P, Falzon G. "How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. Ecological Informatics. 2020;57:101085. doi:10.1016/j.ecoinf.2020.101085.

28. Kocak Z, Evans ES, Zhou S-M, Miller KL, Folz RJ, Shafman TD, et al. Challenges in defining radiation pneumonitis in patients with lung cancer. International Journal of Radiation Oncology, Biology, Physics. 2005;62:635-8. doi:10.1016/j.ijrobp.2004.12.023.

29. Doi H, Nakamatsu K, Nishimura Y. Stereotactic body radiotherapy in patients with chronic obstructive pulmonary disease and interstitial pneumonia: a review. Int J Clin Oncol. 2019;24:899-909. doi:10.1007/s10147-019-01432-y.

30. Okumura M, Hojo H, Nakamura M, Hiyama T, Nakamura N, Zenda S, et al. Radiation pneumonitis after palliative radiotherapy in cancer patients with interstitial lung disease. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;161:47-54. doi:10.1016/j.radonc.2021.05.026.

31. Jiang X, Li T, Liu Y, Zhou L, Xu Y, Zhou X, et al. Planning analysis for locally advanced lung cancer: dosimetric and efficiency comparisons between intensity-modulated radiotherapy (IMRT), single-arc/partial-arc volumetric modulated arc therapy (SA/PA-VMAT). Radiation Oncology. 2011;6:140. doi:10.1186/1748-717X-6-140.

32. Chang JY. Intensity-Modulated Radiotherapy, Not 3 Dimensional Conformal, Is the Preferred Technique for Treating Locally Advanced Lung Cancer. Seminars in Radiation Oncology. 2015;25:110-6. doi:10.1016/j.semradonc.2014.11.002.

33. Shepherd AF, Iocolano M, Leeman J, Imber BS, Wild AT, Offin M, et al. Clinical and Dosimetric Predictors of Radiation Pneumonitis in Patients With Non-Small Cell

Lung Cancer Undergoing Postoperative Radiation Therapy. Practical Radiation Oncology. 2021;11:e52-e62. doi:10.1016/j.prro.2020.09.014.

34. Keffer S, Guy CL, Weiss E. Fatal Radiation Pneumonitis: Literature Review and Case Series. Advances in Radiation Oncology. 2020;5:238-49. doi:10.1016/j.adro.2019.08.010.

35. Nalbantov G, Kietselaer B, Vandecasteele K, Oberije C, Berbee M, Troost E, et al. Cardiac comorbidity is an independent risk factor for radiation-induced lung toxicity in lung cancer patients. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2013;109:100-6. doi:10.1016/j.radonc.2013.08.035.

36. Konkol M, Śniatała P, Milecki P. Radiation-induced lung injury — what do we know in the era of modern radiotherapy? Reports of Practical Oncology and Radiotherapy. 2022;0. doi:10.5603/RPOR.a2022.0046.

37. Niu X, Li H, Chen Z, Liu Y, Kan M, Zhou D, et al. A study of ethnic differences in TGFβ1 gene polymorphisms and effects on the risk of radiation pneumonitis in non-small-cell lung cancer. Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer. 2012;7:1668-75. doi:10.1097/JTO.0b013e318267cf5b.

38. Zhao L, Wang L, Ji W, Wang X, Zhu X, Hayman JA, et al. Elevation of plasma TGF-beta1 during radiation therapy predicts radiation-induced lung toxicity in patients with non-small-cell lung cancer: a combined analysis from Beijing and Michigan. International Journal of Radiation Oncology, Biology, Physics. 2009;74:1385-90. doi:10.1016/j.ijrobp.2008.10.065.

39. Wang L, Liang S, Li C, Sun X, Pang L, Meng X, et al. A Novel Nomogram and Risk Classification System Predicting Radiation Pneumonitis in Patients With Esophageal Cancer Receiving Radiation Therapy. Int J Radiat Oncol Biol Phys. 2019;105:1074-85. doi:10.1016/j.ijrobp.2019.08.024.

40. Bi J, Qian J, Yang D, Sun L, Lin S, Li Y, et al. Dosimetric Risk Factors for Acute Radiation Pneumonitis in Patients With Prior Receipt of Immune Checkpoint Inhibitors. Front Immunol. 2021;12:828858. doi:10.3389/fimmu.2021.828858.

41. Vansteenkiste J, Naidoo J, Faivre-Finn C, Özgüroğlu M, Villegas A, Daniel D, et al. MA05.02 PACIFIC Subgroup Analysis: Pneumonitis in Stage III, Unresectable NSCLC Patients Treated with Durvalumab vs. Placebo After CRT. Journal of Thoracic Oncology. 2018;13:S370-S1. doi:10.1016/j.jtho.2018.08.350.

42. Greenspan H, Ginneken Bv, Summers RM. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. IEEE Trans Med Imaging. 2016;35:1153-9. doi:10.1109/TMI.2016.2553401.

Acknowledgments

This manuscript was prepared using data from datasets (RTOG-0617; NCT00533949-D1, D2, D3) from the NCTN/NCORP Data Archive of the National Cancer Institute's (NCI's) National Clinical Trials Network (NCTN). Data were originally collected from a clinical trial (identifier NCT00533949; "A Randomized Phase III Comparison of Standard-Dose (60 Gy) Versus High-Dose (74 Gy) Conformal Radiotherapy with Concurrent and Consolidation Carboplatin/Paclitaxel +/- Cetuximab (IND #103444) in Patients With Stage IIIA/IIIB Non-Small Cell Lung Cancer"). All analyses and conclusions in this manuscript are the sole responsibility of the authors and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, the NCORP or the NCI.

Supplementary Materials

1. Dataset

A. Details of training set and test set 1

Training set

The training dataset was collected retrospectively from hospital archives. A total of 314 patients treated between January 2013 and December 2018 with definitive RT at Tianjin medical university cancer hospital were retrieved with IRB permission. The inclusion criteria were: (1) Patients identified with histologically confirmed NSCLC or SCLC. (2) Diagnosed with Stage I-III NSCLC and limited-stage SCLC (American Joint Committee on Cancer, 8th edition, 2017) before RT, and patients underwent radical RT. (3) No thoracic RT or thoracic surgery prior to RT. (4) CT examinations were performed at 1, 3, and 6 months (± 15 days) after treatment. Patients were excluded, if treatment break of more than 5 days occurred during RT, or if patients received surgical treatment within 6 months after radiotherapy, of if there was also a second primary tumor, or if the patients had a significant lung infection within 6 months after radiotherapy leading to concern about a potentially non-RT related origin RP.

Test set 1

This dataset was registered prospectively at the same institution as the above training set. A total of 56 patients were enrolled from October 2018 to March 2019. Finally, 35 patients were included in the analysis. 21 patients were excluded because did not meet the eligible criteria, fourteen of which did not follow up CT as planned, six of which did not complete radiotherapy, and one patient died two months after radiation therapy. The inclusion and exclusion criteria were the same as for the training dataset. These registered patients were followed-up every month after concluding radiotherapy. The follow-up items included blood routine examination, C-reactive protein, tumor markers associated with lung cancer, and chest X-rays. Furthermore, patients received CT examination at 1, 3, and 6 months (± 7 days) after end of radiotherapy.

Image acquisition and treatment planning

Intravenous contrast-enhanced planning CT scans were acquired on a single Brilliant (Philips Medical Systems; Best, The Netherlands) multislice scanner with a standardized protocol: 120 kVp, 100 mAs, 3 mm slice thickness, 512 x 512 image matrix, 50 cm fields of view, 0.977 mm pixel spacing and vendor's default convolution kernel. Experienced radiation oncologists delineated the lung cancer gross tumor volume (GTV) and malignant lymph nodes in the Pinnacle treatment planning system (Philips Radiation Oncology Systems; Fitchburg, Wisconsin, United States), with image fusion against complementary imaging studies whenever available (such as positron emission tomography).

The GTV was isotropically expanded by 5 mm, as well as subclinical microscopic malignant lesions to derive the clinical target volume (CTV). The planning target volume (PTV) was an additional 5 mm isotropic expansion around the CTV. Dosimetrist were instructed to cover at least 95% of the PTV with the prescribed RT dose. Delineations conformed to the guidelines set by the Radiotherapy and Oncology Group (RTOG). The relevant dose con-

straints were as follows: MLD <20 Gy, V20 <30%, and volume of the lung receiving 5 Gy (V5) <60%. All patients were nominally prescribed 2Gy per fraction once daily. Radiation oncologists determined the total prescribed dose based on each patient's overall physical condition and best achievable normal tissue constraints. The actual total RT dose delivered ranged between 50 to 70 Gy. The planning CT series with associated RT structure delineations and RT planned radiotherapy 3D dose grids were exported from Pinnacle in the standard DICOM format.

B. Details for re-use of RTOG-0617 dataset (test set 2 and 3)

Patients randomized to the control arm (60 Gy dose) were designated as Test Set 2, and patients that were allocated to the intervention arm (74 Gy dose) were designated as Test set 3. Specific details of the RTOG-0617 trial may be obtained from the trial protocol and published articles.

Inclusion criteria

1. Patients received full course of radiotherapy. 2. The thickness of CT images ranges from 1.25mm to 3mm. 3. Field of view is 500mm diameter and each axial image dimension should be 512 x 512, such that the final reconstructed per pixel spatial resolution falls between 0.9mm and 1.3mm. 4. Either IV contrast or non-IV contrast CT images.

Exclusion criteria

1. Patients diagnosed with (infectious) pneumonia rather than radiation pneumonitis. 2. No corresponding CT images were available. 3. Abnormal CT images with same pixel values (-1024) for the whole lung. 4. With multiple plan and dose files, cannot determine which one was applied. 5. CT images that include only part of the lung.

C. Characteristics of included patients (Table S1-3)

Supplementary table S1. Patient characteristics in training set, test set 1.

Characteristics	Training set n (%)	Without RP Mean ± SD	With RP Mean ± SD	P*	Test set 1 n (%)
Age median	61 (30-85)	61 (30-85)	63 (44-79)	0.005*	62 (34-75)
Gender				0.523	
Male	238(75.8%)	186(78.2%)	52(21.8%)		23 (65.7%)
Female	76(24.2%)	62(81.6%)	14(18.4%)		12 (34.3%)
Smoking				0.569	
Yes	244(77.7%)	191(78.3%)	53(21.7%)		26 (74.3%)
No	70(22.3%)	57(81.4%)	13(18.6%)		9 (25.7%)
KPS				0.725	
≤80	132(42.0%)	103(78.0%)	29(22.0%)		13 (37.1%)
>80	182(58.0%)	145(79.7%)	37l(20.3%)		22 (62.9%)
Diabetes				0.609	

Yes	34(10.8%)	28(82.4%)	6(17.6%)		2 (5.7%)
No	280(89.2%)	220(78.6%)	60(21.4%)		33 (94.3%)
ILA				0.015*	
Yes	25(8.0%)	15(60.0%)	10(40.0%)		9 (25.7%)
No	289(92.0%)	233(80.6%)	56(19.4%)		26 (74.3%)
Pathology				0.656	
LUSC	86(27.4%)	65(75.6%)	21(24.4%)		8 (22.9%)
LUAD	73(23.2%)	59(80.8%)	14(19.2%)		10 (28.6%)
SCLC	155(49.4%)	124(80.0%)	31(20.0%)		17 (48.5%)
Induc chemo				0.739	
Yes	287(91.4%)	226(78.7%)	61(21.3%)		31 (88.6%)
No	27(8.6%)	22(81.5%)	5(18.5%)		4 (11.4%)
CCRT				0.047	
Yes	93(29.6%)	168(76.0%)	53(24.0%)		8 (22.9%)
No	221(70.4%)	80(86.0%)	13(14.0%)		27 (77.1%)
Conso chemo				0.116	
Yes	179(57.0%)	147(82.1%)	32(17.9%)		19 (54.3%)
No	135(43.0%)	101(74.8%)	34(25.2%)		16 (45.7%)
PGTV(Gy)	59.274±2.977	59.204±3.063	59.539±2.634	0.415	60.200±2.870
Smoking index	661.540±571.430	641.840±550.543	735.600±643.084	0.237	668.600±550.412

Abbreviations: LUSC = lung squamous cell carcinoma; LUAD = lung adenocarcinoma; SCLC = small cell lung cancer; IMRT = intensity-modulated radiotherapy; VMAT = volumetric modulated arc therapy; chemo = chemotherapy; KPS = Karnofsky performance score; ILA = interstitial lung abnormalities; Induc chemo = induction chemotherapy; CCRT = concurrent chemoradiotherapy; Conso chemo = consolidation chemotherapy; PGTV = planning gross tumor volume.

* Statistically significant

Supplementary table S2. Patient characteristics in test set 2.

Characteristics	**All pts**	**Without RP**	**With RP**	**P***
	n (%)	**Mean ± SD**	**Mean ± SD**	
Age median	64(37-82)	64(37-82)	65.5(38-80)	0.239
Arm				0.792
No cetuximab	105(54.1%)	87(82.9%)	18(17.1%)	
Cetuximab	89(45.9%)	75(84.3%)	14(15.7%)	

Gender				0.990
Male	115(59.3%)	96(83.5%)	19(16.5%)	
Female	79(40.7%)	66(83.5%)	13(16.5%)	
Race				0.316
American Indian/Alaskan Native	1(0.5%)	1(100.0%)	0(0.0%)	
Asian	6(3.1%)	6(100.0%)	0(0.0%)	
Black or African American	16(8.2%)	16(100.0%)	0(0.0%)	
Native Hawaiian/Other Pacific Islander	1(0.5%)	1(100.0%)	0(0.0%)	
White	169(87.1%)	137(81.1%)	32(18.9%)	
Unknown	1(0.5%)	1(100.0%)	0(0.0%)	
Ethnicity				0.112
Hispanic or Latino	5(2.6%)	3(60.0%)	2(40.0%)	
Not Hispanic or Latino	181(93.3%)	151(83.4%)	30(16.6%)	
Unknown	8(4.1%)	8(100.0%)	0(0.0%)	
Zubrod				0.906
Normal activity	117(60.3%)	98(83.8%)	19(16.2%)	
Symptoms,but nearly fully ambulatory	77(39.7%)	64(83.1%)	13(16.9%)	
Histology				0.880
Squamous cell carcinoma	78(38.7%)	64(85.3%)	11(14.7%)	
Adenocarcinoma	86(44.3%)	70(81.4%)	16(18.6%)	
Large cell undifferentiated	4(2.1%)	4(100.0%)	0(0.0%)	
Non-small cell lung cancer NOS	29(14.9%)	24(82.8%)	5(17.2%)	
Histology grouped				0.586
Non-squamous histology	119(61.3%)	98(82.4%)	21(17.6%)	
Squamous histology	75(38.7%)	64(85.3%)	11(14.7%)	
AJCC Stage				0.073
IIIA, or N2	135(69.6%)	117(86.7%)	18(13.3%)	
IIIB, or N3	59(30.4%)	45(76.3%)	14(23.7%)	
RT technique				0.079
3D-CRT	106(54.6%)	84(79.2%)	22(20.8%)	
IMRT	88(45.4%)	78(88.6%)	10(11.4%)	
EGFR H-Score				0.020*
No H-Score	103(53.1%)	92(89.3%)	11(10.7%)	

H-Score able to be determined	91(46.9%)	70(76.9%)	21(23.1%)	
Smoking history				0.351
Non-smoker	14(7.2%)	10(71.4%)	4(28.6%)	
Former light smoker	16(8.2%)	14(87.5%)	2(12.5%)	
Former heavy smoker	70(36.1%)	57(81.4%)	13(18.6%)	
Current smoker	81(41.8%)	68(84.0%)	13(16.0%)	
Unknown	13(6.7%)	13(100.0%)	0(0.0%)	
PTV (cc)	507.934±273.307	501.690±264.952	539.548±315.003	0.474
V5_lung	57.676±15.288	57.458±15.122	58.781±16.312	0.654
V20_lung	29.056±7.472	28.890±7.440	29.896±7.696	0.486
Dmean_lung	16.664±4.146	16.551±4.163	17.238±4.075	0.394
Received_conc_cetuximab				0.792
No	105(54.1%)	87(82.9%)	18(17.1%)	
Yes	89(45.9%)	75(84.3%)	14(15.7%)	
Received_cons_chemo				0.982
No	21(10.8%)	17(81.0%)	4(19.0%)	
Yes	173(89.2%)	145(83.8%)	28(16.2%)	
Received_cons_cetuximab				0.853
No	112(57.7%)	94(83.9%)	18(16.1%)	
Yes	82(42.3%)	68(82.9%)	14(17.1%)	
Survival_status				0.172
Alive	88(45.4%)	77(87.5%)	11(12.5%)	
Dead	106(54.6%)	85(80.2%)	21(19.8%)	
Survival_months	23.604(2.562-61.465)	23.982(2.562-61.465)	19.333(4.468-57.293)	0.380
Local_failure_months	18.693(1.413-60.447)	19.136(1.413-60.447)	17.559(4.468-56.669)	0.931
Distant_failure_months	15.621(2.562-61.465)	16.196(2.562-61.465)	10.972(3.679-47.963)	0.149
Progression_free_survival_months	11.531(1.413-60.447)	11.662(1.413-60.447)	10.627(3.679-47.963)	0.492

Abbreviations: Pts = patients; RP = radiation pneumonitis; NOS = Non-small-cell lung cancer not otherwise specified; AJCC = American Joint Committee on Cancer; Rt technique = radiotherapy technique used to treat patient; 3D-CRT=3dimensional comformal radiation therapy; IMRT = intensity-modulated radiotherapy; EGFR H-Score = epidermal growth factor receptor immunohistochemistry scores; PTV = planning tumor volume; V5_lung= Lung V5 (%);V20_lung= Lung V20 (%); Dmean_lung = Mean lung dose (Gy); conc_cetuximab = concurrent cetuximab; cons_chemo = consolidation chemotherapy; cons_cetuximab = consolidation cetuximab. * Statistically significant

Supplementary table S3. Patient characteristics in test set 3.

Characteristics	All pts	Without RP	With RP	P*
	n (%)	Mean ± SD	Mean ± SD	
Age median	63(41-82)	62(41-82)	67 (46-76)	0.244
Arm				0.284
No cetuximab	85(53.8%)	72(84.7%)	13(15.3%)	
Cetuximab	73(46.2%)	57(78.1%)	16(21.9%)	
Gender				0.129
Male	89(56.3%)	69(77.5%)	20(22.5%)	
Female	69(43.7%)	60(87.0%)	9(13.0%)	
Race				1.000
American Indian/Alaskan Native	1(0.6%)	1(100.0%)	0(0.0%)	
Asian	4(2.5%)	4(100.0%)	0(0.0%)	
Black or African American	17(10.8%)	14(82.4%)	3(17.6%)	
White	136(86.1%)	110(80.9%)	26(19.1%)	
Ethnicity				0.710
Hispanic or Latino	4(2.5%)	3(75.0%)	1(25.0%)	
Not Hispanic or Latino	152(96.2%)	124(81.6%)	28(18.4%)	
Unknown	2(1.3%)	2(100.0%)	0(0.0%)	
Zubrod				0.962
Normal activity	92(58.2%)	75(81.5%)	17(18.5%)	
Symptoms,but nearly fully ambulatory	66(41.8%)	54(81.8%)	12(18.2%)	
Histology				0.938
Squamous cell carcinoma	70(44.3%)	57(81.4%)	13(18.6%)	
Adenocarcinoma	63(39.9%)	52(82.5%)	11(17.5%)	
Large cell undifferentiated	1(0.6%)	1(100.0%)	0(0.0%)	
Non-small cell lung cancer NOS	24(15.2%)	19(79.2%)	5(20.8%)	
Histology grouped				0.950
Non-squamous histology	88(55.7%)	72(81.8%)	16(18.2%)	
Squamous histology	70(44.3%)	57(81.4%)	13(18.6%)	
AJCC Stage				0.696
IIIA, or N2	103(65.2%)	85(82.5%)	18(17.5%)	
IIIB, or N3	55(34.8%)	44(80.0%)	11(20.0%)	
RT technique				0.528

3D-CRT	90(57.0%)	75(83.3%)	15(16.7%)	
IMRT	68(43.0%)	54(79.4%)	14(20.6%)	
EGFR H-Score				0.804
No H-Score	85(53.8%)	70(82.4%)	15(17.6%)	
H-Score able to be determined	73(46.2%)	59(80.8%)	14(19.2%)	
Smoking history				0.133
Non-smoker	11(7.0%)	10(90.9%)	1(9.1%)	
Former light smoker	14(8.9%)	12(85.7%)	2(14.3%)	
Former heavy smoker	52(32.9%)	37(71.2%)	15(28.8%)	
Current smoker	78(49.4%)	68(87.2%)	10(12.8%)	
Unknown	3(1.9%)	2(66.7%)	1(33.3%)	
PTV (cc)	482.660±261.390	472.956±243.796	525.828±330.203	0.328
V5_lung	57.109±14.654	56.205±14.385	61.133±15.411	0.104
V20_lung	31.223±7.960	30.853±7.473	32.869±9.828	0.218
Dmean_lung	19.159±4.553	18.967±4.480	20.017±4.849	0.262
Received_conc_cetuximab				0.284
No	85(53.8%)	72(84.7%)	13(15.3%)	
Yes	73(46.2%)	57(78.1%)	16(21.9%)	
Received_cons_chemo				0.749
No	22(13.9%)	19(86.4%)	3(13.6%)	
Yes	136(86.1%)	110(80.9%)	26(19.1%)	
Received_cons_cetuximab				0.089
No	93(58.9%)	80(86.0%)	13(14.0%)	
Yes	65(41.1%)	49(75.4%)	16(24.6%)	
Survival_status				0.499
Alive	52(32.9%)	44(84.6%)	8(15.4%)	
Dead	106(67.1%)	85(80.2%)	21(19.8%)	
Survival_months	20.253(0.493-59.560)	20.828(0.493-59.560)	18.528(3.022-47.799)	0.229
Local_failure_months	13.732(0.493-59.560)	13.798(0.493-59.560)	12.122(3.022-47.799)	0.557
Distant_failure_months	11.827(0.493-59.560)	11.662(0.493-59.560)	13.535(3.022-46.058)	0.935
Progression_free_survival_months	9.823(0.493-59.560)	9.888(0.493-59.560)	9.527(3.022-46.058)	0.475

Abbreviations: Pts = patients; RP = radiation pneumonitis; NOS = Non-small-cell lung cancer not otherwise specified; AJCC = American Joint Committee on Cancer; Rt technique = radiotherapy technique used to treat patient; 3D-CRT=3dimensional comformal radiation

therapy; IMRT = intensity-modulated radiotherapy; EGFR H-Score = epidermal growth factor receptor immunohistochemistry scores; PTV = planning tumor volume; V5_lung= Lung V5 (%);V20_lung= Lung V20 (%); Dmean_lung = Mean lung dose (Gy); conc_cetuximab = concurrent cetuximab; cons_chemo = consolidation chemotherapy; cons_cetuximab = consolidation cetuximab.

* Statistically significant

2. Pre-processing and model construction

A. Pre-processing

1) all images were resampled to a 1mm x 1mm x 1mm resolution; 2) images were cropped according to total lung mask that was contoured using a deep learning-based automatic tool and reviewed and modified by two physicians (ZZ and MY) to ensure the correctness of the delineations, and 3) normalization of CT images was based on the adjusted lung window. Since the lung window used in daily clinical practice is not perfectly suited for deep learning approaches, we modified it to -500 (center) and 1200 (range) for the image normalization process. RD images are normalized according to a range of 0-80. 4) To reduce the hardware requirements, images were compressed to 84 x 84 x 84. 5) 20% of the training set was randomly chosen as the validation set to evaluate the model performance during training. 6) Data augmentation methods were used to increase the number and diversity of training data.

Data augmentation:

1. Random affine transform.

Random affine transformation of the image keeping center invariant with the probability of 0.8.

2.Random rotate

Random rotate of the images with 0-90 degrees with the probability of 0.8.

2. Random flip

Random flip of the image with the channels of x, y and z with the probability of 0.8.

3. Random zoom transform.

Random zoom of the images from the 80% to 100% of input size with the probability of 0.5.

4. Random spatial crop

Random crop the spatial crop of the images from the size of 70 to 84 in each channel with the random center.

The data preprocessing method was applied by SimpleITK (v2.0.2). The data augmentation was applied based on MONAI package (v0.8.0).

B. Model architecture (Figure S1)

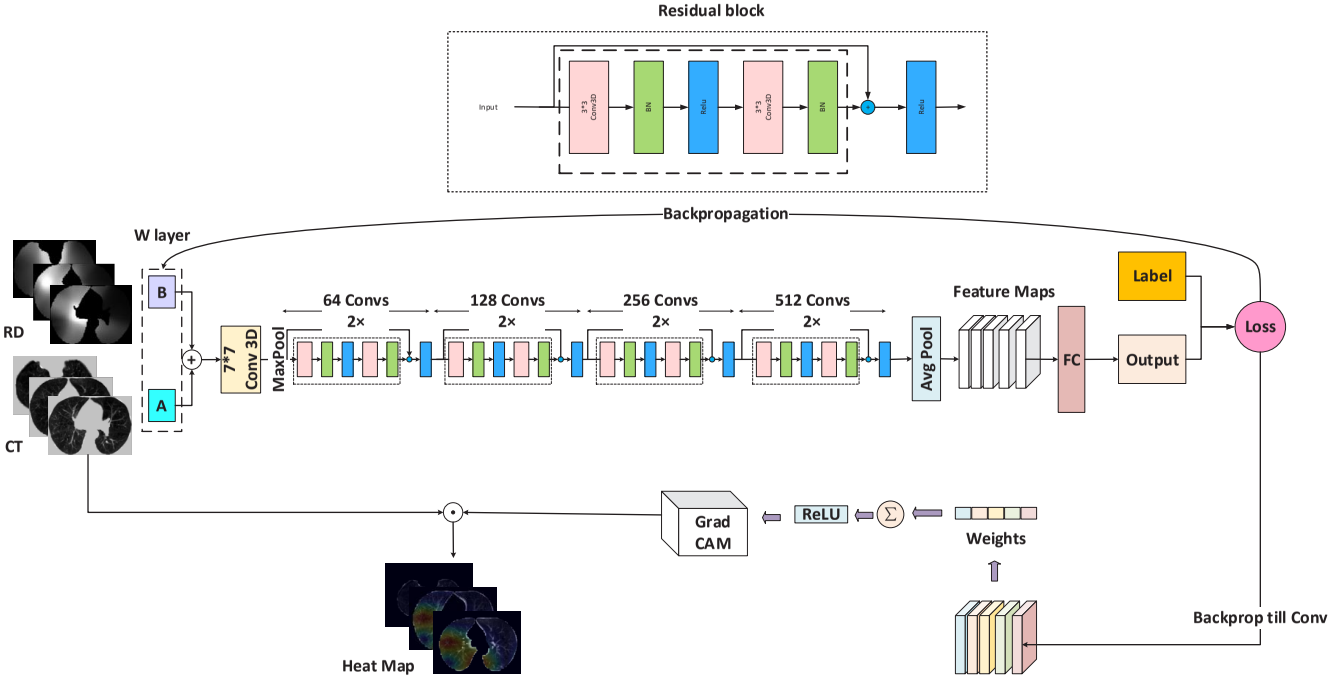


Figure S1. The model architecture

ResNet

The ResNet(Residual Neural Network)model is one of the widely used classification model was designed by Kaiming He [1]. Traditional convolutional networks have more or less information loss, gradients disappear or explode in the model The ResNet solves this problem by directly passing the input information to the output, the integrity of the information is protected. The entire network only needs to learn the part of the difference between the input and output, which simplifies the learning goals and difficulty. The biggest innovation of ResNet is that there are many bypasses to directly connect the input to the output layers, this structure is also called shortcut or skip connections. The residual block was designed by this innovation shown in Figure 1B. There are different ResNet models from ResNet10 to ResNet152 that are named according to the number of residual blocks.

Architecture

The architecture of proposed network is shown in Figure 1B, which is composed of weight layer, 3D convolution layers and one fully connection (FC) layer.

The weight layer is composed of two weights, which can be obtained for CT and RD from different datasets with diverse treatment patterns. The function of weights is shown below.

$$Weight_i = \frac{W_i}{\sum_{i=0}^{n} Wi}$$

$Weight_i$ is the weight for each input channel. W_i is the value in weight layers. $\frac{W_i}{\sum_{i=0}^{n} Wi}$ calculates the rate of specific input channel's weight in all input channels.

The ResNet was selected as the backbone of the model with the weight layer as the top of the model to give the weights of combine the CT and radiation dose images.

In RP prediction processing, the paired CT and radiation dose images was sent into the model as the input. Then, it was combined by the weight layer and the convolution layers will extract 512 high-level features from the input. Finally, the fully connection layer with Sigmoid activation function will predict the probability of the patient will have RP after treatment according to extracted features.

The model was built on PyTorch (v1.7.1). All code was written on python language (v3.8.5) The experiments were performed on a workstation with one NVIDIA Quadro T2000 (4GB) GPU.

C. Grad-CAM

To obtain the deeper understanding how the model makes the decision to classification whether the patient will have RP. The gradient-weighted class activation mapping (Grad CAM) was applied to visualize the interpretation for the proposed model.

Class activation map (CAM) shows the most significant position of the model through the visual thermal map, so it can be used to explain how the model make the decision [2]. The CAM method exploits the abundant spatial and semantic information in the convolution layers, replace the fully connection layer with global average layer, and replace the feature

map with the mean value of all pixels of the feature map as the activation map.

The Grad CAM is class-specific, meaning it can produce a separate visualization for every class present in the image, which is the upgrade method of CAM, uses the gradients of the target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept [3].

The workflow of Grad CAM shown in Figure 1B and Figure S1.

The function of Grad CAM is shown below:

$$a_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

$$L_{Grad-CAM}^c = ReLU(\sum_k a_k^c A^k)$$

A_{ij}^k presents the feature map from the last convolution layer in the model, k is the number of channels, i,j are the position of pixels. y^c is the prediction of the specific c-th class. $\frac{\partial y^c}{\partial A_{ij}^k}$ means the gradient of the c-th class to the feature map. By averaging this gradient per channel, we can get a k-dimensional vector which is the weights for feature map channels.

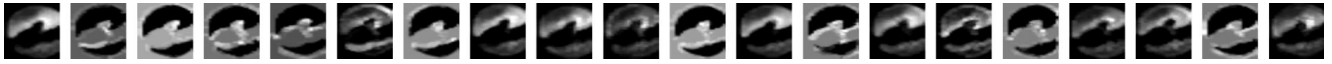
D. Deep learning network training strategy

The model was trained by the Adam optimizer with a learning rate of 0.0001. The Binary Cross-Entropy loss function was selected to train the model.

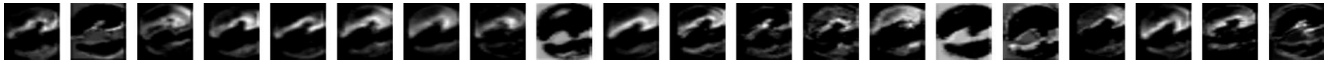
To avoid overfitting, the following method were used: 1. The Batch Normalization (BN) was used on the features after the average pooling. 2.The Dropout layer was used on the features after BN with the rate of 0.5. 3. The L2 regulation penalty were added on the weights of fully connection layers. 4. The early stopping method was used in the training process.

E. Feature map (Figure S2-9)

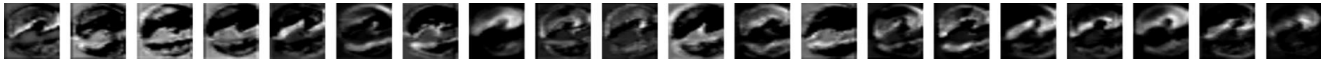
Supplementary Figure S2. Feature map extract from residual block 1



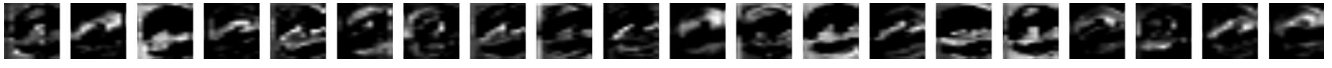
Supplementary Figure S3. Feature map extract from residual block 2



Supplementary Figure S4. Feature map extract from residual block 3

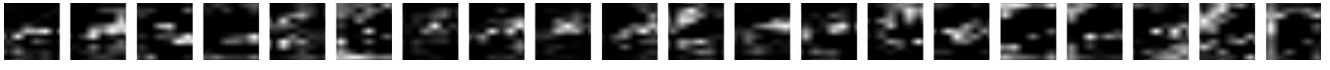


Supplementary Figure S5. Feature map extract from residual block 4

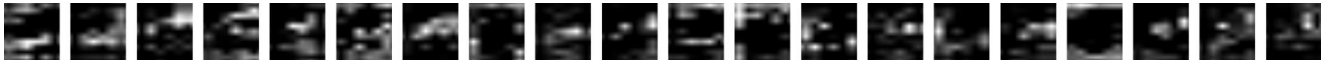


Supplementary Figure S6. Feature map extract from residual block 5

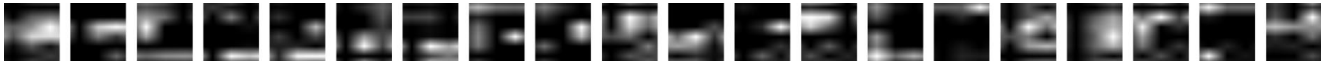
Supplementary Figure S7. Feature map extract from residual block 6



Supplementary Figure S8. Feature map extract from residual block 7



Supplementary Figure S9. Feature map extract from residual block 8



3. Instructions of the online tool for radiation pneumonitis prediction (Figure S10)

**** This tool can only be used for research purposes, not for commercial use.***

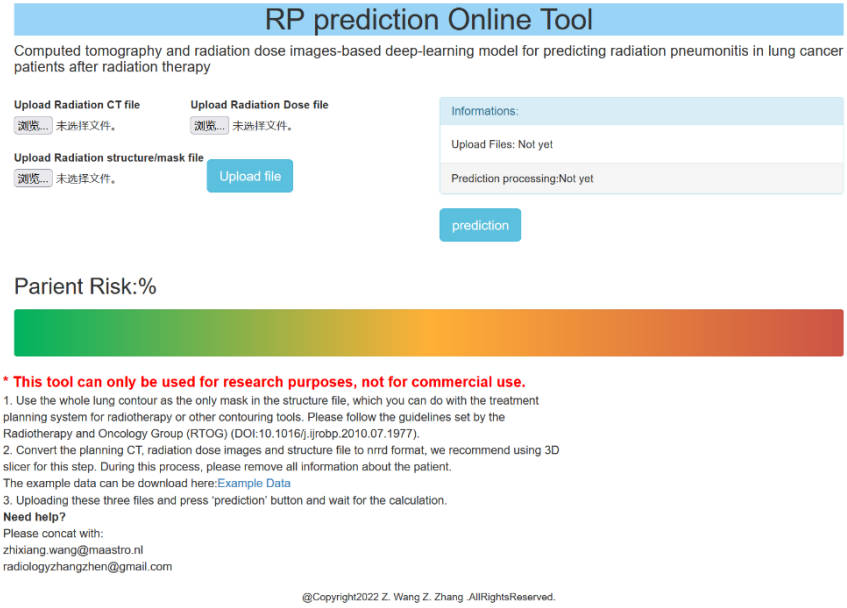
1. Use the whole lung contour as the only mask in the structure file, which you can do with the treatment planning system for radiotherapy or other contouring tools. Please follow the guidelines set by the Radiotherapy and Oncology Group (RTOG) (DOI:10.1016/j.ijrobp.2010.07.1977) [4].

2. Convert the planning CT, radiation dose images and structure file to nrrd format and rename them to 'CT', 'RD' and 'RS', we recommend using 3D slicer for this step. During this process, please remove all information about the patient.

3. Uploading these three files and press 'prediction' button and wait for the calculation.

Online tool: https://flask-web-zx.herokuapp.com/

Supplementary Figure S10. Radiation pneumonitis prediction online tool



4. Weights of CT and radiation dose (RD) for the models (Table S4)

Supplementary Table S4. The weights for CT and RD from each model.

Dataset	Weight of CT	Weight of RD (mean, range)
Training set	1	1.42
Test set 2*	1	1.53 (1.50-1.58)
Test set 2#	1	1.67

Test set 3*	1	1.20 (1.02-1.41)
Test set 3#	1	1.25

* The asterisk indicates that the coefficients of CT and RD for this model are adjusted with 40 patients for each set; # The pound symbol indicates that the coefficients of CT and RD for this model are adjusted using the entire data set.

5. CT, RD, DVH and clinical based model

A. Deep learning model constructed by CT or RD

Supplementary Table S5. Performance of deep learning model constructed by CT or RD alone.

Model	AUC (95%CI)	Accuracy (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)
CT	0.630 (0.571-0.674)	0.582 (0.535-0.628)	0.565 (0.503-0.617)	0.612 (0.552-0.689)
RD	0.686 (0.637-0.735)	0.646 (0.602-0.690)	0.659 (0.599-0.719)	0.632 (0.568-0.695)

B. Dose-volume histogram (DVH) metrics and clinical parameters selection and model construction

Due to the collinearity of DVH metrics, it does is not suitable to perform the complex feature selection approaches. Instead, the predictive model is built using the already acknowledged metrics V20 and mean lung dose.

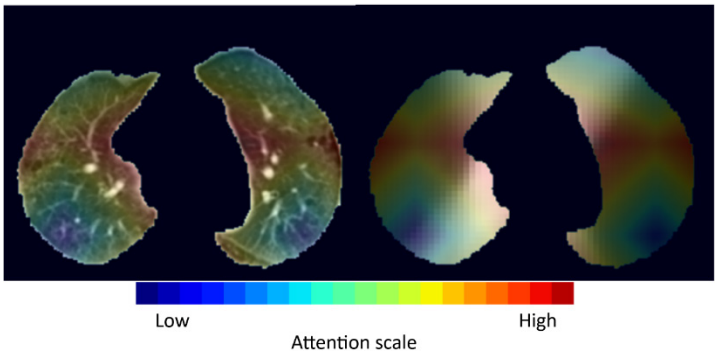
The logistic regression method was used to build this model based on the training set (314 subjects). As external validation, we evaluated the DVH model using the prospectively-registered cohort of 35 subjects (test-set-1). Processing of these 35 subjects followed exactly the same procedure as for the model development cohort, and none of these subjects were used in any way during model construction.

The clinical model was presented as a multivariable logistic regression model. The predictors (age and interstitial lung abnormalities) were selected according to Table S1.

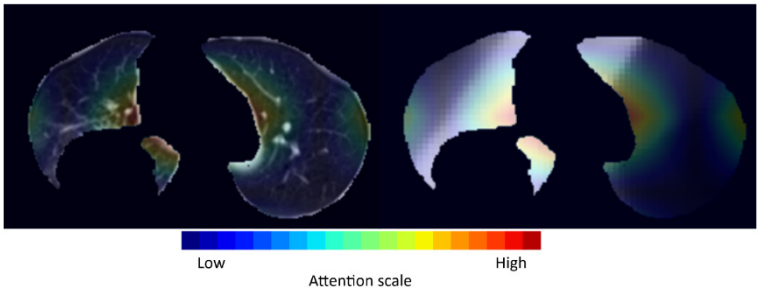
The training and validation strategies are the same as those described in the previous paragraph for the DVH model.

6. Attention maps (Figure S11-13)

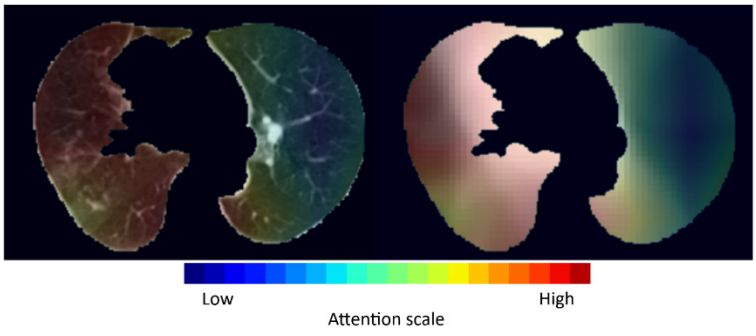
Supplementary Figure S11 Attention map of a patient with interstitial lung abnormality.



Supplementary Figure S12 Attention map shows attention areas around the heart region



Supplementary Figure S13 Attention map shows attention areas roughly follow radiation dose distribution



Reference

1. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv:151203385 [cs]. 2015.

2. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 2921-9.

3. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Int J Comput Vis. 2020;128:336-59. doi:10.1007/s11263-019-01228-7.

4. Kong F-M, Ritter T, Quint DJ, Senan S, Gaspar LE, Komaki RU, et al. CONSIDERATION OF DOSE LIMITS FOR ORGANS AT RISK OF THORACIC RADIOTHERAPY: ATLAS FOR LUNG, PROXIMAL BRONCHIAL TREE, ESOPHAGUS, SPINAL CORD, RIBS, AND BRACHIAL PLEXUS. International journal of radiation oncology, biology, physics. 2011;81:1442-57. doi:10.1016/j.ijrobp.2010.07.1977.

Chapter 8: Discussion

8.1 Executive summary

In this thesis, I have explored the potential utilization of radiomics and image-based biomarkers in radiotherapy. Three aspects were investigated: 1. methodological refinement and exploration of radiomics studies (Chapter 2 and 3). 2. The role of radiomics in radiotherapy prognosis (Chapter 4 and 5). 3. The value of image-based biomarker for radiotherapy side effect prediction (Chapter 6 and 7). Radiomics, as a field at the intersection of medicine, computer science, and engineering, has many pitfalls that are likely to be overlooked in the steps of implementation. I think that a practical checklist covering methodological and clinical utility is a solution to this problem. Therefore, in Chapter 2 I integrated engineering and clinical perspectives to propose assessment criteria for evaluating the quality of machine learning-based quantitative imaging analysis studies. The checklist covers several aspects such as data preparation, data processing, and clinical potential assessment. The checklist will allow investigators to self-check the repeatability, reproducibility, and clinical potential utilities of image-based biomarker studies in their future experiments. It is worth stating that it is not our initial intention to require or expect future studies to conform to every item in the checklist, which would be very difficult to achieve, limited by the objective research environment. Because researchers with different disciplinary backgrounds pursue varying research priorities, investigators can be selective in the use of certain items in the checklist depending on the goals pursued. In developing the checklist, I found that most of the published studies included relatively small sample sizes, thus leading to the study in Chapter 3. In this study, I demonstrated that deep learning has the potential to generate synthetic samples that expand the training set. Its application to lung cancer is currently under investigation. While the results are encouraging, we should be aware that there is still a long way to go from demonstrating its potential to being used in real-world research. The fact that some of the generated samples will still be recognized by doctors as artificial also shows that our algorithm is not perfect.

After specifying the qualitative evaluation criteria that should be followed for the implementation of radiomics and image-based biomarker studies, I explored the application of radiomics to practical clinical problems. Prospective data or/and multicenter data were used in the studies included in this thesis to validate the models to ensure that model validation results were as objective as possible. Improving patient prognosis is one of the main clinical concerns and the most important goal in refining treatment modalities. In Chapter 4, I predicted distant metastases in early-stage lung cancer patients who received SBRT. This is of great relevance in a clinical context where clinical decisions after SBRT are diverse for this group of patients, where the decision of whether to give the patient systemic therapy or not is a difficult one. Therefore, based on this prediction model, patients with a high risk of distant metastases can be treated aggressively as well as followed up more intensely [1]. Also, tumors with a high risk of metastasis tend to be more aggressive, so this model might also guide the choice of drugs. It should be noted that the patients in this study were all from China, a developing country with unevenly developed medical resources. Some of the patients were treated in developed cities and then returned to places of residence where medical resources are relatively scarce, so close imaging follow-up was difficult for them. Selective and aggressive follow-up of patients, i.e., those at high risk, is more practical and easier to accomplish. In addition, this study examined the application of multimodality radiomics in prognosis. Both CT images reflecting tissue anatomical information and PET functional images containing metabolic information were investigated. Based on this study

and the results of Chapters 6 and 7, I believe that performing integration of images based on different imaging rationales, and thus association with multidimensional data such as clinical parameters, can improve the predictive power and generalizability of prediction models. Patients with early-stage lung cancer have a high likelihood of achieving long-term survival [2], so I chose distant metastasis as an the endpoint for clinical decision making. In contrast, for patients with locally advanced disease, their overall survival is shorter and the primary indicator for assessing many treatment strategies and therapeutic techniques [3]. In Chapter 5, I explore the prediction of overall survival in locally advanced lung cancer. Region of interest (ROI) selection is an important step in radiomics research, but most of the studies published so far focus on malignant tissues, which is admittedly one of the determinants of overall survival, but the underlying status of normal tissues, especially organs such as the heart and lungs, is also strongly associated with survival [4]. In the study, I used radiomics features of both tumor tissue and lung tissue to build prediction models, and in a subgroup analysis, the results showed that overall survival could be predicted based on tumor tissue or lung tissue alone. This demonstrated that both tissues contain survival-related information and combining these two aspects might lead to models with better predictive power. This is reasonable because the underlying status of lung tissue reflects not only the patient's pre-existing disease, but also the tolerance to treatment for some cases.

Overall, in Chapter 4 and 5, I demonstrated that radiomics can make prognostic predictions. From a technical point of view, the use of simple statistical models, such as logistic regression or Cox proportional hazards regression, can be effective in predicting prognosis. From a clinical perspective, the results of the radiomics prediction models are consistent with clinical understanding, as discussed in Chapters 4 and 5, respectively. Having an accurate assessment of prognosis is essential for making clinical decisions, and based on the risk assessment values output by the model, physicians can individualize patient treatment to improve patient prognosis. On the flip side, predictive models can also help physicians stratify risk populations, allowing physicians to achieve greater clinical management effectiveness with their limited time. The model also has a contribution to patients and from an economic point of view, which is described in the "Research Impact" section.

Radiotherapy is bittersweet and can be accompanied by side effects while treating the disease. In Chapters 6 and 7 I explored one of the common side effects of radiotherapy, radiation pneumonitis. Radiation pneumonitis is a non-infectious pneumonia due to radiation that is difficult to treat [5], making it important to predict its occurrence [6]. Clinically, patients with high chance of developing radiation pneumonitis will be given relatively low radiation doses or prophylactic medications. From a practical standpoint, it is not possible to give prophylactic medications to all patients or to closely monitor every patient, which is uneconomical and time constrained. Therefore, based on the output of the prediction model, we can target and closely monitor the high-risk group, which also increases the likelihood of early and timely detection of radiation pneumonitis. It broadens the treatment window and has the potential to reduce the rate of severe disease and mortality in radiation pneumonitis. In these two studies, I used dose images that are not much explored yet. The application of dose images in radiomics is known as dosiomics. As mentioned before, radiation pneumonitis is caused by radiation and therefore focusing only on the basal state of the lung tissue is not sufficient and should be included in the evaluation together with information on radiation dose. Dose-volume histogram (DVH) parameters have been widely used in previous studies and in clinical practice, but they are only an approximation of the radiation dose dis-

tribution. Dosiomics (Chapter 6) allows one to explore the details of the dose distribution, like radiomics to obtain texture information from medical images. Unlike dosiomics with handcrafted features, features extracted by deep learning models are not pre-defined, and in Chapter 7 I demonstrated that deep learning-based dose features also have predictive power.

For image-based biomarker studies, the generalizability of the model is a challenge. Applying established models to different cohorts is difficult for complex reasons, including heterogeneity of human races and healthcare systems, different equipment for data collection, and differences in treatment protocols. In Chapter 7, I used a deep learning approach to address the issue of generalizability to clusters with heterogeneity to some extent. Although in Chapter 6 I demonstrated that the inclusion of clinical parameters improved the predictive power of the model, how to include clinical parameters in deep learning is currently a question without a perfect answer, and specifically, in which module of deep learning and in what form, remains to be investigated.

In conclusion, in Chapters 6 and 7, I demonstrated the feasibility of using CT images and dose images to predict radiation pneumonitis. From a technical point of view, the use of a suitable model construction allows the synergistic effect of the two different sources of images. From a clinical point of view, the radiation pneumonitis prediction model can be used as a part of precision radiotherapy. It allows physicians to give individualized treatment regimens, more appropriate follow-up to patients, and to provide better physician-patient communication.

Chapters 4 to 7, the clinical application studies, followed as much as possible the methodological quality checklist presented in Chapter 2. For example, for improving repeatability and reproducibility, I used automatic segmentation tools to ensure consistency, and settings for the parameters were provided in the appendix of the articles. In terms of clinical utility, I used decision curve analysis or provided a nomogram, and online automated calculation tools. In the practice of these studies (Chapter 4 to 7) I found this checklist (Chapter 2) to be helpful and feasible. However, I also experienced that some items could not be implemented due to the objective conditions of the studies. For example, exploring the association of radiomics with other types of features could not be realized due to the limitation of data sources. In addition, I included as many samples as possible in the studies, for example, in Chapter 7 I used the largest dataset in the field of radiomics research for radiation pneumonitis, 701 patients, but this falls far short of the “big data” requirement, which urges us to continue to refine the algorithms presented in Chapter 3 so that they can be used in real-world studies.

8.2 Limitations of this work

Some limitations of the studies in this thesis should be noted. First, I developed predictive models for prognosis and toxicity, and users can optimize clinical decisions based on the model outputs. But the trade-off between prognosis and toxicity has not been explored in depth. Finding a balance between therapeutic efficacy and side effects that allows patients to obtain the best possible outcome while suffering less is a worthwhile but unexamined study in this thesis. This can only be achieved through both clinical and engineering efforts. From a clinical perspective, a dataset with more detailed endpoints is needed, i.e., a dataset with at least the ground truth for both prognostic and side effects endpoints. From an engineering perspective, the selection of appropriate models or algorithms is crucial and still to

be explored.

Second, this thesis did not conduct a comparison of different modeling algorithms and did not explore the models in depth. Statistical models, logistic regression or Cox proportional hazards regression, were used in chapters 3 to 6, instead of using models such as XGBoost that were considered in many studies to have stronger predictive power and to exploit the nonlinear relationship of features [7, 8]. The main reasons for this are, firstly, the logical simplicity of these algorithms I used, which have a stronger explanatory nature, due to their ability to obtain the weights of each feature. Secondly, the aim of our study was to demonstrate that radiomics or image-based biomarkers have potential for clinical applications and high predictive indicators were not our main goal. In Chapter 7, the deep learning algorithm I used is suitable, but it is not known if it is optimal.

Third, I did not provide a biological explanation for the selected features. Although I made medical knowledge-based speculations about the meaning of the features or the predictive logic of the model based on the equations of the features (Chapters 4, 5 and 6) or the feature maps (Chapters 6 and 7), this was not sufficiently rigorous. Several articles [9] have been published on biological interpretation of image-based biomarkers by methods such as proteomic and genomics as introduced in the Introduction section (Chapter 1). I agree that this is an important tool to unravel the "black box" of radiomics research, which is an objective and rigorous approach.

Finally, I did not validate the model in the real world, which is a deficiency in the vast majority of current studies. What I should acknowledge as a researcher is that the data I used, even prospectively, were influenced by the inclusion and exclusion criteria, which inevitably results in bias in the included data, even if the degree of bias is very small, the impact on the model is not known. To make it more objective and realistic, we should measure the performance of the model in the real-world and routine clinical setting.

8.3 Future perspectives

This thesis investigated the future role of radiomics and image-based biomarkers in supporting clinical decision making. Through implementation of these studies, I believe that artificial intelligence will shine in the future in the field of medical imaging. However, there are currently only a few cases where tasks such as prediction are applied to daily clinical applications other than automatic segmentation applications for clinical work. The reasons for this include many aspects such as technology, ethics, and policy. As a physician in radiotherapy who has learned some radiomics techniques, I would like to present some of my views on the future development of this field only for clinical and engineering purposes.

First, in the future, the cooperation between disciplines will be closer and the integration of resources will be the direction. This is divided into two aspects. The first aspect is that the combination of medicine and engineering will be more frequent, because both doctors and technicians have gained a deeper understanding of the application of artificial intelligence in the medical field in recent years, and are subjectively more willing to cooperate and have an understanding of the need for cooperation. The second aspect is that crossover between medical disciplines is more common, as the role of radiotherapy is currently complementary to other treatments (Chapter 1), regardless of the patient's stage.

Second, the advancement of technology and the use of reasonable technology. These are

two different aspects. Advances in technology, such as the application of federated learning will address the issue of cross-institutional data transfer to some extent, especially for cross-country data sharing where there are many policy issues, and the application of the technology will help researchers meet policy requirements. Applying the suitable technology means choosing the most appropriate technology from the perspective of clinical needs and tasks, rather than the most advanced technology. This goes hand in hand with the first point mentioned above.

Third, based on the above two points, the establishment of a platform that can integrate technology and clinical needs. I established the prototype of such a platform for daily clinical applications in Chapter 7. In the future, as technology evolves, more and more easy-to-use and efficient platforms will emerge. This will lower the threshold of technical knowledge for users (clinical decision makers), while the modular platform can be personalized and changed according to the needs of users. This will facilitate the practical application of image-based biomarkers.

References

1. Hwang JK, Page BJ, Flynn D, Passmore L, McCaul E, Brady J, et al. Validation of the Eighth Edition TNM Lung Cancer Staging System. J Thorac Oncol. 2020;15:649-54. doi:10.1016/j.jtho.2019.11.030.

2. Chansky K, Detterbeck FC, Nicholson AG, Rusch VW, Vallières E, Groome P, et al. The IASLC Lung Cancer Staging Project: External Validation of the Revision of the TNM Stage Groupings in the Eighth Edition of the TNM Classification of Lung Cancer. J Thorac Oncol. 2017;12:1109-21. doi:10.1016/j.jtho.2017.04.011.

3. Vinod SK, Hau E. Radiotherapy treatment for lung cancer: Current status and future directions. Respirology. 2020;25 Suppl 2:61-71. doi:10.1111/resp.13870.

4. Taylor C, Correa C, Duane FK, Aznar MC, Anderson SJ, Bergh J, et al. Estimating the Risks of Breast Cancer Radiotherapy: Evidence From Modern Radiation Doses to the Lungs and Heart and From Previous Randomized Trials. J Clin Oncol. 2017;35:1641-9. doi:10.1200/jco.2016.72.0722.

5. Niu S, Zhang Y. Applications and therapeutic mechanisms of action of mesenchymal stem cells in radiation-induced lung injury. Stem Cell Res Ther. 2021;12:212. doi:10.1186/s13287-021-02279-9.

6. Käsmann L, Dietrich A, Staab-Weijnitz CA, Manapov F, Behr J, Rimner A, et al. Radiation-induced lung toxicity - cellular and molecular mechanisms of pathogenesis, management, and literature review. Radiation Oncology (London, England). 2020;15:214. doi:10.1186/s13014-020-01654-9.

7. Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Sci Rep. 2021;11:6968. doi:10.1038/s41598-021-86327-7.

8. Luo L, Lin H, Huang J, Lin B, Huang F, Luo H. Risk factors and prognostic nomogram for patients with second primary cancers after lung cancer using classical statistics and machine learning. Clin Exp Med. 2022. doi:10.1007/s10238-022-00858-5.

9. Tomaszewski MR, Gillies RJ. The Biological Meaning of Radiomic Features. Radiology. 2021;298:505-16. doi:10.1148/radiol.2021202553.

Appendices

Summary

A large number of medical images are acquired during the management of radiotherapy patients, including in pre-radiotherapy diagnosis, during the treatment of radiotherapy, monitoring of side effects and efficacy. In recent years, many studies have demonstrated that quantitatively extracted features from such images can be used as biomarkers to assist in clinical decision making. One of the widely studied approaches is radiomics, where features are quantitatively extracted from images non-invasively, and biomarkers based on these features are screened and modeled by machine learning approaches. However, there are still methodological and clinical application challenges, and accordingly, this thesis investigated the following three aspects: a) methodological quality assessment of radiomics studies. B) prognostic value of image-extracted biomarkers in lung cancer patients undergoing radiotherapy. C) prediction of a radiotherapy side effect, radiation pneumonitis, by image-based machine learning models.

In this thesis, an objective methodological quality assessment of current radiomics research was presented, based on which a methodological assessment checklist was proposed (Chapter 2). Difficulties faced in research such as insufficient sample size may be alleviated by methods such as deep learning (Chapter 3). This thesis also demonstrated the prognostic (Chapters 4 and 5) and toxicity prediction (Chapters 6 and 7) capabilities of image-derived biomarkers and compared them to benchmark models commonly used in clinical settings. The results demonstrated that image-derived biomarkers have the potential for clinical application and that combining multi-modality images and multi-dimensional information can improve the power of the models (Chapters 4 and 6). Selection of regions of interest (Chapter 5) and model building algorithms (Chapter 8) based on clinical needs is critical.

Overall, this thesis demonstrated the potential for future applications of image-derived biomarkers for the management of radiotherapy patients and to support clinical decision making.

Samenvatting

Een aanzienlijke hoeveelheid medische beelden worden gemaakt tijdens de radiotherapeutische behandeling van patiënten, zoals gedurende de diagnose voor de radiotherapie, tijdens de behandeling met radiotherapie, en gedurende monitoring van bijwerkingen en effectiviteit na de behandeling. De laatste jaren hebben vele studies aangetoond dat kwantitatief uit beelden geëxtraheerde kenmerken kunnen worden gebruikt als biomarkers ter ondersteuning van de klinische besluitvorming. Deze benadering heet Radiomics waarbij "machine learning" wordt gebruikt om kenmerken te screenen en te combineren in modellen.. Voor de toepassing van Radiomics zijn er echter nog enkele methodologische en klinische uitdagingen, en daarom werden in dit proefschrift de volgende drie aspecten onderzocht: a) methodologische kwaliteitsbeoordeling van radiomics-studies. b) prognostische waarde van uit beelden geëxtraheerde biomarkers bij longkankerpatiënten die radiotherapie ondergaan. c) voorspelling van bijwerkingen van radiotherapie, stralingspneumonitis, door beeldgebaseerde machine-learning modellen.

In dit proefschrift wordt een objectieve methodologische kwaliteitsbeoordeling van het huidige radiomics-onderzoek gepresenteerd, op basis waarvan een methodologische beoordelingschecklist is voorgesteld (hoofdstuk 2). Uitdagingen voor dergelijk onderzoek, zoals onvoldoende steekproefgrootte, kunnen worden verlicht door methoden als deep learning (hoofdstuk 3). In dit proefschrift werden ook de prognostische (hoofdstukken 4 en 5) en toxiciteit-voorspellende waarde (hoofdstukken 6 en 7) van uit beelden afgeleide biomarkers aangetoond en vergeleken met de benchmarkmodellen die gewoonlijk in klinische settings worden gebruikt. De resultaten toonden aan dat van beelden afgeleide biomarkers potentieel hebben voor klinische toepassing en dat het combineren van multimodale beelden en multidimensionale informatie de kracht van de modellen kan verbeteren (hoofdstukken 4 en 6). Selectie van de interessante gebieden op een beeld (hoofdstuk 5) en van model-algoritmen (hoofdstuk 8) op basis van klinische behoeften is van cruciaal belang.

In het algemeen toonde dit proefschrift het potentieel aan voor toekomstige toepassingen van uit beelden afgeleide biomarkers voor de radiotherapeutische behandeling van patiënten en ter ondersteuning van klinische besluitvorming.

Impact

1. Clinical impact

This thesis examined radiomics and image-based biomarkers in the context of clinical needs. The outcomes studied are the most important for radiation oncologists to consider when making clinical decisions. There are three implications for the clinic. First, different sources of information used in our daily work, such as clinical parameters, tumor metabolic information, and anatomical imaging information can be valuable in the field of artificial intelligence (Chapters 4, 6, and 7). Clinical cognition is the basis of AI research in the medical field. In Chapter 5, I optimized the ROI based on clinical experience and demonstrated that clinical knowledge could guide the optimization of models. Second, the efficacy of parameters based on radiomics, for example, may meet or even exceed the benchmark parameters currently used in the clinical practice. In Chapter 6, I demonstrated that dosiomics predictive power outperforms current benchmark DVH parameters. Third, artificial intelligence tools have the potential to be embedded in daily clinical practice. The application platforms presented in Chapter 7 evidence the potential for future applications of artificial intelligence.

Overall, this thesis contributes to the application of radiomics and artificial intelligence to assist clinical work and update clinical tools.

2. Technological impact

Although I do not propose new algorithms or invent new hardware in this paper, there are several lessons learned from the application of the technology that can help technologists working in the field. First, problems that are considered clinically intractable can be accomplished using the appropriate technology needed for the clinical task. As discussed previously, predictive power beyond the benchmark model can be achieved using simple artificial intelligence models. Again, the approach presented in Chapter 7 adapts commonly used algorithms to specific tasks, dynamically combining CT images and radiation dose images to achieve results that are difficult to accomplish with non-artificial intelligence approaches. Second, trials and studies from a technical perspective should take full account of clinical experience and clinical needs. For example, the image preprocessing methods and the choice of algorithms, should be adapted to the task context. Third, I follow the tenet of open science and made our code, configuration files and data as open as possible. This can be made available to future technologists for reference.

3. Impact on patients

Although the users of the model developed in this thesis are physicians, it is the patients who are ultimately the recipient of the clinical intervention. The methods and models presented in this thesis have practical implications for patients.

First, in terms of practical benefits to patients, the approach proposed in this thesis can help optimize clinical decision making, thereby prolonging patient prognosis (Chapters 4 and 5) and reducing patient suffering (Chapters 6 and 7). They may also be used to inform patients better of their expected outcomes and ultimately in shared decision making, where appropriate.

Second, from a financial burden perspective, as a result of accurate screening of high-risk groups, physicians will be able to target patients for closer follow-up or recommend certain treatments or medications. Considering the whole population, this will reduce the overall economic burden. All of the models developed in this thesis are based on routine examinations without the need to undertake expensive tests such as genetic sequencing. Hardware such as computers are reusable. Therefore, patients do not have to bear additional costs. This is important for society as a whole, but especially important in countries, such as China, in which patients themselves have to pay a significant part of the treatment cost.

4. Societal impact

In this thesis, we demonstrated the potential of image-derived biomarkers for clinical applications. From a societal perspective, effective support for clinical decision making can reduce the financial burden on patients and insurance expenditures, thereby increasing the effectiveness of health insurance utilization.

The application of the clinical prediction models presented in the paper has the potential to provide better treatment protocols for patients, reduce the incidence of side effects, and improve prognosis. As a result, the workload of physicians can be reduced to some extent and more medical resources can be freed up to serve the society.

At the same time, this thesis provides an explanation of the clinical applications of AI, which may improve physicians' acceptance of AI and thus contribute to the future application of AI tools in the real world.

Curriculum Vitae

Zhen Zhang was born on June 1990 in China. He holds a bachelor's degree in Science in Medicine from Shandong First Medical University and a master's degree in Oncology from Tianjin Medical University in China. In 2014, he joined the department of Radiotherapy at Tianjin medical university cancer hospital and institution. In 2020, he joined Maastro Clinic and started as a PhD student at Maastricht University at the Faculty of Health, Medicine and Life Sciences, under the supervision of Prof. Andre Dekker.

His main research interests are the applications of radiomics and artificial intelligence in radiotherapy. During his PhD he studied programming and artificial intelligence and combined it with clinical knowledge and experience. He has published several articles and oral presentations at ESTRO and other conferences and received ESMO travel grant. He is a reviewer for International journal of radiation oncology biology physics, Radiotherapy and Oncology, Frontiers in oncology, Frontiers in public health, Thoracic cancer, Journal of Digital Imaging, Translational cancer research and Academy of medicine.

List of Publications

Published research on international journals (1 (shared) first author, * (shared) corresponding author)

1. Zhenwei Shi [1,] *, **Zhen Zhang** [1,] *, Zaiyi Liu, Lujun Zhao, Zhaoxiang Ye, Andre Dekker and Leonard Wee. Methodological Quality of Machine Learning-Based Quantitative Imaging Analysis Studies in Esophageal Cancer: A Systematic Review of Clinical Outcome Prediction after Concurrent Chemoradiotherapy. Eur J Nucl Med Mol Imaging 2021. https://doi.org/10.1007/s00259-021-05658-9.

2. **Zhen, Zhang** [1], Zhixiang Wang, Meng Yan, Jiaqi Yu, Andre Dekker, Lujun Zhao and Leonard Wee. Radiomics and Dosiomics Signature from Whole Lung Predicts Radiation Pneumonitis: A Model Development Study with Prospective External Validation and Decision-Curve Analysis. Int J Radiat Oncol Biol Phys 2022, S0360-3016(22)03189-3. https://doi.org/10.1016/j.ijrobp.2022.08.047.

3. **Zhen, Zhang** [1], Zhixiang Wang [1], Tianchen Luo, Meng Yan, Andre Dekker, Dirk De Ruysscher, Alberto Traverso, Leonard Wee and Lujun Zhao. Computed tomography and radiation dose images-based deep-learning model for predicting radiation pneumonitis in lung cancer patients after radiation therapy: A pilot study with external validation. Radiotherapy and Oncology, https://doi.org/10.1016/j.radonc.2023.109581.

4. Zhixiang Wang [1], **Zhen Zhang** [1], Ying Feng, Lizza E. L. Hendriks, Razvan L. Miclea, Hester Gietema, Janna Schoenmaekers, Andre Dekker, Leonard Wee and Alberto Traverso. Generation of Synthetic Ground Glass Nodules Using Generative Adversarial Networks (GANs). Eur Radiol Exp 2022, 6 (1), 59. https://doi.org/10.1186/s41747-022-00311-y.

5. Junzhuo Liu, Zhixiang Wang, Ye Zhang, Alberto Traverso, Andre Dekker, **Zhen Zhang** * and Qiaosong Chen *. CycleGAN Clinical Image Augmentation Based on Mask Self-Attention Mechanism. IEEE Access 2022, 10, 105942–105953. https://doi.org/10.1109/ACCESS.2022.3211670.

6. Tohidinezhad Fariba, Dennis Bontempi, **Zhen Zhang**, Anne-Marie Dingemans, Joachim Aerts, Gerben Bootsma, Johan Vansteenkiste, Sayed Hashemi, Egbert Smit, Hester Gietema, Hugo JWL. Aerts, Andre Dekker, Lizza E.

L. Hendriks, Alberto Traverso, and Dirk De Ruysscher. Computed Tomography-Based Radiomics for the Differential Diagnosis of Pneumonitis in Stage IV Non-Small Cell Lung Cancer Patients Treated with Immune Checkpoint Inhibitors. European Journal of Cancer 2023, 183:142–51. https://doi.org/10.1016/j.ejca.2023.01.027.

7. Ying Feng, Zhixiang Wang, Meizhu Xiao, Jinfeng Li, Yuan Su, Bert Delvoux, **Zhen Zhang**, Andre Dekker, Sofia Xanthoulea, Zhiqiang Zhang, Alberto Traverso, Andrea Romano, Zhenyu Zhang, Chongdong Liu, Huiqiao Gao, Shuzhen Wang and Linxue Qian. An Applicable Machine Learning Model Based on Preoperative Examinations Predicts Histology, Stage, and Grade for Endometrial Cancer. Front Oncol 2022, 12, 904597. https://doi.org/10.3389/fonc.2022.904597.

8. Ying Feng, Zhixiang Wang, Ran Cui, Meizhu Xiao, Huiqiao Gao, Huimin Bai, Bert Delvoux, **Zhen Zhang**, Andre Dekker, Andrea Romano, Shuzhen Wang, Alberto Traverso, Chongdong Liu and Zhenyu Zhang. Clinical analysis and artificial intelligence survival prediction of serous ovarian cancer based on preoperative circulating leukocytes. J Ovarian Res 2022, 15 (1), 64. https://doi.org/10.1186/s13048-022-00994-2.

9. Hao Yu, Jiaqi Zhang, **Zhen Zhang**, Youyou Wang, Guangying Xu, Liming Xu, Ningbo Liu, Lujun Zhao and Ping Wang. One Cycle of Concurrent Chemotherapy vs. Two Cycles of Concurrent Chemotherapy With Radiation Therapy in Patients With Limited-Stage Small Cell Lung Cancer. Frontiers in Oncology 2021, 11, 785022. https://doi.org/10.3389/fonc.2021.785022.

10. Zhixiang Wang, Glauco Lorenzut, **Zhen Zhang**, Andre Dekker, and Alberto Traverso. Applications of Generative Adversarial Networks (GANs) in Radiotherapy: Narrative Review. Precision Cancer Medicine 2022, 5 (0). https://doi.org/10.21037/pcm-22-28.

Submitted research

1. **Zhen Zhang [1],** Lu Yu [1], et al. A PET/CT Radiomics Model for Predicting Distant metastasis in Early-Stage Non–Small Cell Lung Cancer Patients Treated with Stereotactic Body Radiotherapy: A Multicentric Study. (Cancer, under review)

2. **Zhen Zhang [1],** Meng Yan [1], et al. Combining tumor radiomics features and whole-lung radiomics features to predict prognosis in locally advanced

non-small cell lung cancer treated with curative radiotherapy. (In preparation)

Acknowledgements

First of all, I would like to thank my supervisors: Andre Dekker, Leonard Wee, Alberto Traverso. They have given me a lot of support, both in life and in my research. They are all very kind, humble and knowledgeable people. During my PhD studies, they taught me how to think as a researcher to consider scientific questions and find reasonable solutions. This will be of great help to me throughout my life.

Then, I would like to thank my friends at Maastricht University: Zhixiang Wang, Shenlun Chen, who helped me a lot, especially in programming, and who helped me to go from knowing nothing about programming to being able to program independently.

I would like to thank my supervisor in China, Lujun Zhao, and my students for their support and assistance in providing me with a lot of clinical help. This helped me to complete my research with high efficiency.

I would like to thank my family for their unwavering support so that I can concentrate fully on my research and studies.

Finally, I would like to thank the Netherlands government and the Chinese government for their support, which has enabled me to have a very good study time.