Maastricht University

Deep learning applications in lung cancer imaging

Citation for published version (APA):

Hosny, A. (2022). *Deep learning applications in lung cancer imaging*. [Doctoral Thesis, Maastricht University]. ProefschriftMaken. https://doi.org/10.26481/dis.20220406ah

Document status and date:
Published: 01/01/2022

DOI:
10.26481/dis.20220406ah

Document Version:
Publisher's PDF, also known as Version of record

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 08 Mar. 2023

Deep Learning Applications in Lung Cancer Imaging

Ahmed Hosny

© copyright Ahmed Hosny, 2022

Lay-out and printing: ProefschriftMaken || www.proefschriftmaken.nl

ISBN: 978-94-6423-732-0

DOI: https://doi.org/10.26481/dis.20220406ah

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of the author or the copyright-owning journals for previous published chapters.

The work presented in this thesis is made possible by the financial support of the US National Institutes of Health (NIH-USA U24CA194354 and NIH-USA U01CA190234).

Deep Learning Applications in Lung Cancer Imaging

DISSERTATION

to obtain the doctoral degree
at 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
Wednesday
April 6th, 2022 at 16:00 hours

by

Ahmed Hosny

Supervisor
Prof. Dr. Ir. Hugo J.W.L. Aerts

Co-supervisor
Dr. Raymond H. Mak, Harvard Medical School, USA

Assessment committee
Prof. Dr. Ir. A. L.A.J. Dekker (Chair)
Prof. Dr. Ir. Dirk De Ruysscher
Dr. Ir. Stefan Klein, Erasmus Medical Center Rotterdam
Prof. Dr. Ir. Wiro J. Niessen, Erasmus Medical Center Rotterdam
Prof. Dr. Frank Verhaegen

Contents

1

Chapter 1

General Introduction and Outline

Artificial Intelligence (AI) has recently made substantial strides in perception, the interpretation of sensory information, allowing machines to better represent and interpret complex data. This has led to major advances in applications ranging from web search and self-driving vehicles to natural language processing and computer vision - tasks that up until a few years ago could only be done by humans[1]. Deep learning (DL) is a subset of machine learning (ML) that is based on a neural network structure loosely inspired by the human brain. Such structures learn discriminative features from data automatically, giving them the ability to approximate very complex nonlinear relationships. While most earlier AI methods have led to applications with sub-human performance, recent deep learning algorithms are able to match and even surpass humans in task-specific applications[2–5]. This owes to recent advances in AI research, the massive amounts of digital data now available to train algorithms, as well as modern powerful computational hardware **(Figure 1)**. Deep learning methods have been able to defeat humans in the strategy board game of Go, an achievement that was previously thought to be decades away given the highly complex game space and massive number of potential moves[6]. Following the trend towards a human-level general AI, researchers predict that AI will automate many tasks including translating languages, writing best-selling books, and performing surgery - all within the coming decades[7].

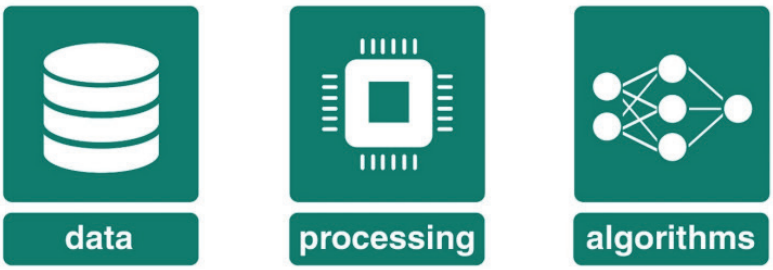

Figure 1. The recent rise of AI applications is mainly driven by the massive amounts of digital data now available to us, modern powerful computational hardware, as well as advances in AI methodologies, namely deep learning.

This thesis explores AI applications within oncology. Cancer's ever evolving nature and interaction with its surroundings continue to challenge patients, clinicians, and researchers alike. One of its deadliest forms appears in the lungs, leading to the most cancer-related mortalities worldwide[8]. Lung cancer is also the second most commonly diagnosed cancer in both men and women[9] with non-small cell lung cancer (NSCLC) comprising 85% of cases[10]. Medical imaging data is often central to diagnosing, prognosticating, and treating NSCLC patients. These non-invasive images, however, often offer information that goes beyond that captured through routine radiographic evaluation by experts. The aforementioned advances in AI methods have enabled the high-throughput extraction, and subsequent processing, of high-dimensional quantitative features from images. More specifically, this dialogue between AI and medical imaging has been recently manifested in radiomics.

Radiomics is a data-centric field involving the extraction and mining of quantitative features as a means to quantify the solid tumor radiographic phenotype[11]. It hypothesizes that radiographic phenotypes represent underlying pathophysiologies and are thus capable of discriminating between disease forms for predicting prognosis and therapeutic response[12]. Radiomics research has primarily relied on explicitly programmed algorithms that extract engineered (hand-crafted) imaging features. Such features commonly represent tumor shape, voxel intensity information (statistics), and patterns (textures). More specifically within oncology, Radiomics has demonstrated success in stratifying tumor histology[13], tumor grades[14], and clinical outcomes[11]. Additionally, associations with underlying gene expression patterns have also been reported[15]. Given these associations, radiomic features have been used to build prognostic and predictive models making use of statistical machine learning algorithms coupled with feature selection strategies[16]. More recent work, however, has shifted towards deep learning as the *de facto* machine learning approach[17].

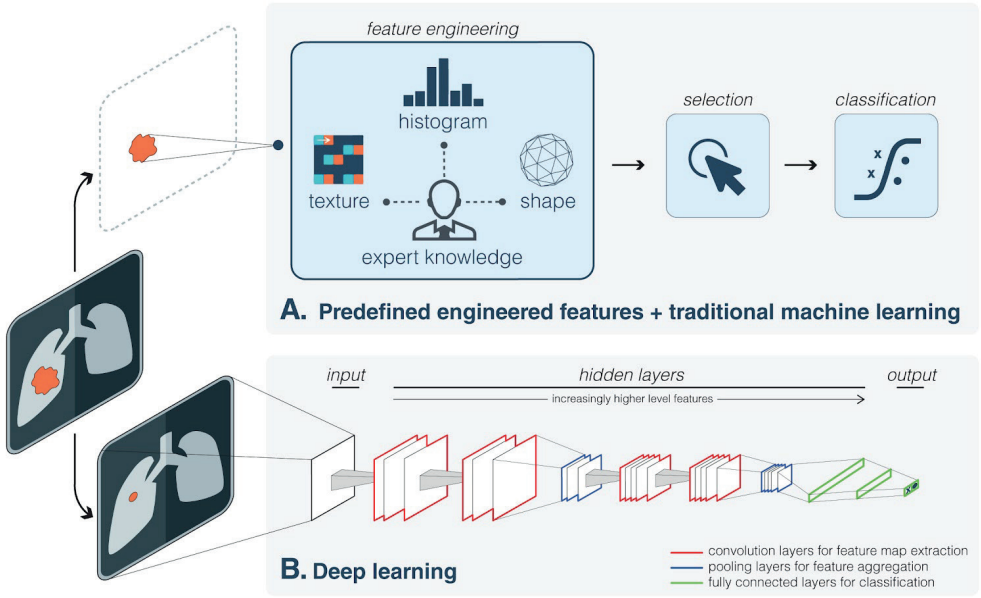

Figure 2. Artificial intelligence methods in medical imaging. This schematic outlines two methods. The first method relies on engineered features extracted from regions of interest on the basis of expert knowledge. The most robust features are selected and fed into machine learning classifiers. The second method uses deep learning where discriminatory features are learned automatically from data.

Deep learning has shown great promise in areas that rely on imaging data including radiology[18], pathology[19], dermatology[20], and ophthalmology[21] to name a few. In lieu of the often subjective visual assessment of images by trained clinicians, deep learning automatically identifies complex patterns in data and hence provides evaluations in a quantitative manner **(Figure 2)**. Compared to feature engineering approaches, crafting

and selecting the most robust features is inherent to deep learning networks and thus they require little to no human input. Deep learning methods have outperformed their engineered feature counterparts in many tasks including mammographic lesion detection[22], mortality prediction[23], and multimodal image registration[24].

This thesis explores deep learning applications in medical imaging, specifically those pertaining to NSCLC patients. Part 1 introduces a wide range of AI applications within clinical oncology, oncology-focused radiology, as well as radiotherapy. Part 2 focuses on experimental studies for tumor characterization in imaging data. Therein, the utility of deep learning in stratifying NSCLC patients from single and longitudinal images is explored. In these studies, imaging data is used to predict prognostic endpoints such as survival and distant metastasis, as well as response to main and adjuvant therapy. Additionally, therapeutic applications in radiotherapy planning are also examined, including the automated segmentation of target tumors and lymph nodes in imaging data. Part 3 highlights best practices in conducting experimental studies, both on the data science and computational methodology fronts. Finally, part 4 outlines AI applications in global health.

PART 1: Artificial Intelligence in Cancer Imaging

Chapter 2 outlines how the clinical oncology practice is experiencing rapid growth in data that are collected to enhance cancer care. Given the recent advances in the field of AI, there is now a computational basis to integrate and synthesize this growing body of multi-dimensional data, deduce patterns, and predict outcomes to improve shared patient and clinician decision-making. This chapter explores a pathway of clinical cancer care touchpoints starting from prevention and early screening to response assessment and follow up **(Figure 3)**. It then analyzes narrow task-specific AI applications in each of these touchpoints. Mapping the field in such a manner helps formalize AI interventions as a combination of three ingredients: 1. a specific clinical use case within one of the touchpoints, 2. a specific data type ranging from text to imaging, and finally 3. the appropriate ML method.

Chapter 3 discusses AI applications particularly pertaining to cancer imaging. Historically, in radiology practice, trained physicians visually assessed medical images for the detection, characterization and monitoring of tumors. AI methods excel at automatically recognizing complex patterns in imaging data and providing quantitative, rather than qualitative, assessments of radiographic characteristics. This chapter outlines two AI methods pertaining to image-based tasks. The first method relies on engineered features (e.g. volume, shape, texture, intensity) extracted from regions of interest on the basis of expert knowledge. The most robust features are selected and fed into traditional ML classifiers (e.g. support vector machines and random forests). The second method uses deep learning and comprises several layers where feature extraction, selection and ultimate classification are performed automatically and simultaneously during training. The chapter then moves to discuss three main clinical radiology tasks in oncology: abnormality detection, characterization, and subsequent monitoring of change. This is

followed by an investigation into technologies currently being utilized in the clinic and research efforts aimed at integrating AI developments into each of these tasks.

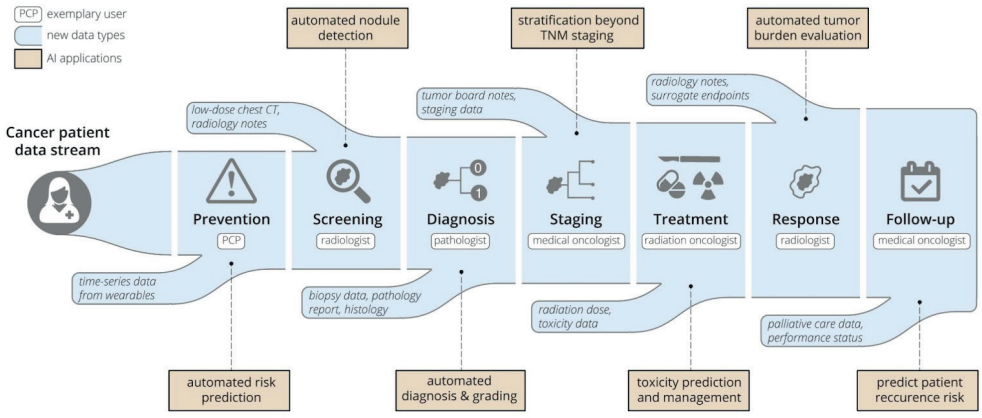

Figure 3. An example cancer patient pathway converges with an ever-increasing data stream. Potential AI applications and exemplary clinical users at each touchpoint are also illustrated.

Chapter 4 shifts the discussion to AI applications in cancer therapeutics, namely radiotherapy (RT). RT has the potential to be transformed by AI given its multifaceted, highly technical nature with heavy reliance on digital data processing and computer software. This paves the way for improved accuracy, precision, efficiency, and overall quality for cancer patients. The chapter provides a high-level overview of the RT workflow starting from patient evaluation and imaging steps to treatment planning and RT plan quality assurance. It then outlines the impact that AI may have on each of these steps **(Figure 4)**. Additionally, the roles of RT medical professionals (radiation oncologists, medical physicists, dosimetrists, therapists) are discussed, as well as how these roles may evolve alongside clinical task automation.

PART 2: Prognostic and Therapeutic Deep Learning Applications

Chapter 5 explores the utility of deep learning in the prognostication of lung cancer patients. Today, standard prognostication involves tumor staging, which in turn is based on a relatively coarse and discrete stratification. Radiographic medical images offer patient- and tumor-specific information that could be used to complement such clinical prognostic efforts.

In this chapter, an analysis setup is designed comprising seven independent datasets across five institutions totaling 1194 NSCLC patients imaged with computed tomography (CT) and treated with either radiotherapy or surgery. We evaluated the prognostic

signature of quantitative imaging features extracted through deep learning networks, and assessed its ability to stratify patients into low and high mortality risk groups as per a two-year overall survival cut off. In patients treated with surgery, deep learning networks significantly outperformed models based on predefined tumor features as well as volume and maximum diameter. In addition to highlighting image regions with prognostic influence, we also evaluated deep learning features for robustness against physiological imaging artifacts and input variability, as well as correlated them with molecular information through gene expression data.

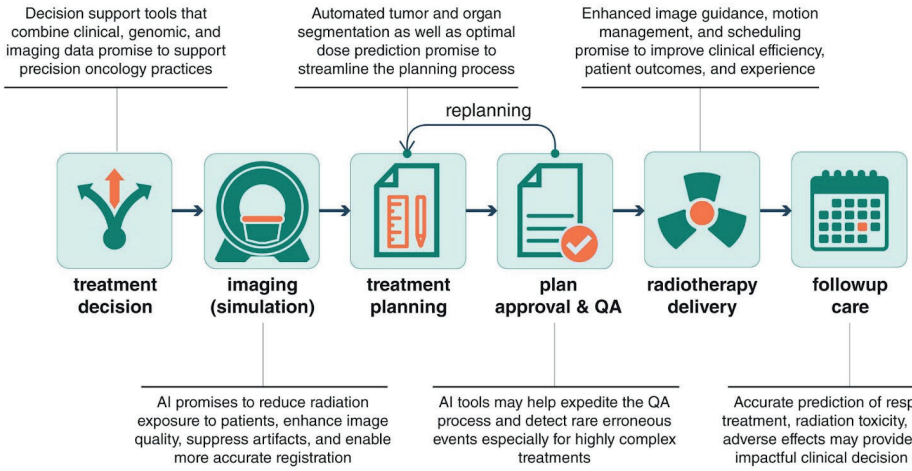

Figure 4. A general overview of the RT workflow with brief descriptions of expected AI applications in each step.

Chapter 6 proposes a deep learning approach to predicting NSCLC tumor histology from non-invasive standard-of-care CT data. Tumor histology is an important predictor of therapeutic response and outcomes in lung cancer. Tissue sampling for pathologist review is the most reliable method for histology classification, however, recent advances in deep learning for medical image analysis allude to the utility of radiologic data in further describing disease characteristics and for risk stratification.

In this chapter, we trained and validated convolutional neural networks (CNNs) on a dataset comprising 311 early-stage NSCLC patients receiving surgical treatment, with a focus on the two most common histological types: adenocarcinoma and Squamous Cell Carcinoma. The CNNs were able to predict tumor histology with an AUC of 0.71($P = 0.018$). We also found that using ML classifiers such as k-nearest neighbors and support vector machines on CNN-derived quantitative radiomics features yielded comparable discriminative performance. Our best performing CNN functioned as a robust probabilistic classifier in heterogeneous test sets, with qualitatively interpretable visual explanations to its predictions.

Chapter 7 evaluates deep learning models for predicting clinical outcomes through analyzing time-series CT images of locally advanced NSCLC patients. While qualitatively tracking lesions over space and time may be trivial, the development of clinically-relevant, automated methods that incorporate serial imaging data is far more challenging. The models, based on a combination of CNNs and recurrent neural networks (RNNs), were found to be significantly predictive of survival and cancer-specific outcomes (progression, distant metastases and local-regional recurrence). Model performance was enhanced with each additional follow-up scan. The models stratified patients into low and high mortality risk-groups, which were found to be significantly associated with overall-survival.

Chapter 8 explores DL applications in RT treatment planning. While AI methods have demonstrated great potential in streamlining clinical RT tasks, most studies are confined to *in silico* validation in small internal cohorts, lacking data on real-world clinical utility. We clinically validated DL models for localizing and segmenting primary NSCLC tumors and involved lymph nodes in CT images.

In this chapter, DL models were validated across four focus areas. *Benchmarking:* Models showed an improvement over the interobserver benchmark ($P < .01$), and were within the intraobserver benchmark. *Primary Validation*: Performance on internal data segmented by the same expert was volumetric dice (VD) 0.83 [0.82,0.85], within the interobserver benchmark. Performance on internal data segmented by other experts was VD 0.70 [0.67,0.73], worse than the interobserver benchmark ($P < .0001$). Similar results were observed on subsequent external validation data, including clinical trial and diagnostic radiology data. *Secondary Validation:* Models were found to be stable across separate images of the same subject, but tend to underestimate tumor volume by an average of 12%. *Human subject experiments:* We found no significant differences between *de novo* and AI-assisted segmentations. AI-assistance led to a 65% reduction in segmentation time ($P < .0001$).

PART 3: AI Methods and Best Practices

Chapter 9 discusses two underlying radiomics methodologies used in treatment response prediction and prognosis in RT, with broader implications across other cancer therapies. More specifically, it compares and contrasts traditional radiomics and its use of handcrafted features with deep learning radiomics where learning of relevant radiographic features is automated. The chapter explores multiple challenges shared across both methods, including reproducibility and over-fitting on small datasets. It also highlights the potential utility of deep learning interpretability efforts in decoding new insights from cancer images and non-intuitive information that is uncharted thus far.

Chapter 10 explores means to provide better transparency to datasets. Data is a fundamental ingredient in building AI models, and there are direct correlations between data quality and model robustness, fairness, and utility. This chapter introduces the Dataset Nutrition Label, a diagnostic framework providing a distilled yet comprehensive

overview of dataset "ingredients". The label is designed to be flexible and adaptable; it comprises a diverse set of qualitative and quantitative modules generated through multiple statistical and probabilistic modelling backends. Consulting such a label prior to AI model development promotes vigorous data interrogation practices, aids in recognizing inconsistencies and imbalances, provides an improved means to selecting more appropriate datasets for specific tasks, and subsequently increases the overall quality of AI models.

Chapter 11 identifies obstacles hindering transparent and reproducible AI research, including the absence of sufficiently documented methods and computer code. These shortcomings limit the evidence required for others to prospectively validate and clinically implement studies, while undermining their scientific value. The chapter also provides potential solutions with implications for the broader field.

PART 4: Beyond Cancer Imaging

Chapter 12 explores AI applications in global health, given the limited discussions around what AI can bring to medical practice in low- and middle-income countries where workforce shortages and limited resources constrain the access to and delivery of care. The chapter outlines the important role AI may play in addressing global healthcare inequities at three levels: the individual patient, health system, and population levels.

Finally, **chapter 13** provides a general discussion of the results presented in this thesis and related future perspectives.

References

1. Editors, N. Auspicious machine learning. *Nature Biomedical Engineering* **1**, 0036 (2017).
2. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
3. Moravčík, M. *et al.* DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* **356**, 508–513 (2017).
4. Xiong, W. *et al.* Toward Human Parity in Conversational Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **25**, 2410–2423 (2017).
5. Pendleton, S. D. *et al.* Perception, Planning, Control, and Coordination for Autonomous Vehicles. *Machines* **5**, 6 (2017).
6. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
7. Grace, K., Salvatier, J., Dafoe, A., Zhang, B. & Evans, O. When Will AI Exceed Human Performance? Evidence from AI Experts. http://arxiv.org/abs/1705.08807 (2017).
8. Torre, L. A. *et al.* Global cancer statistics, 2012. *CA Cancer J. Clin.* **65**, 87–108 (2015).
9. American Cancer Society: Cancer Facts and Figures 2017. Atlanta, Ga: American Cancer Society, 2017. *Am. Cancer Soc. 2014* https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2016/cancer-facts-and-figures-2016.pdf.
10. Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E. & Adjei, A. A. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin. Proc.* **83**, 584–594 (2008).
11. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).
12. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446 (2012).
13. Ganeshan, B. *et al.* Non–Small Cell Lung Cancer: Histopathologic Correlates for Texture Parameters at CT. *Radiology* **266**, 326–336 (2013).
14. Ganeshan, B., Abaleke, S., Young, R. C. D., Chatwin, C. R. & Miles, K. A. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. *Cancer Imaging* **10**, 137–143 (2010).
15. Grossmann, P. *et al.* Defining the biological basis of radiomic phenotypes in lung cancer. *Elife* **6**, (2017).
16. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci. Rep.* **5**, 13087 (2015).
17. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* (2018) doi:10.1038/s41568-018-0016-5.
18. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).

19. Cruz-Roa, A. *et al.* Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci. Rep.* **7**, 46450 (2017).

20. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).

21. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* (2016) doi:10.1001/jama.2016.17216.

22. Kooi, T. *et al.* Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303–312 (2017).

23. Carneiro, G., Oakden-Rayner, L., Bradley, A. P., Nascimento, J. & Palmer, L. Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography. in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* 130–134 (2017).

24. Yang, X., Kwitt, R., Styner, M. & Niethammer, M. Quicksilver: Fast predictive image registration - A deep learning approach. *Neuroimage* **158**, 378–396 (2017).

PART I

Artificial Intelligence in Cancer Imaging

2

Chapter 2

Artificial Intelligence for Clinical Oncology

B Kann, *A Hosny*, & HJWL Aerts

Cancer Cell 2021

Abstract

Clinical oncology is experiencing rapid growth in data that are collected to enhance cancer care. With recent advances in the field of Artificial Intelligence (AI), there is now a computational basis to integrate and synthesize this growing body of multi-dimensional data, deduce patterns, and predict outcomes to improve shared patient and clinician decision-making. While there is high potential, significant challenges remain. In this perspective, we propose a pathway of clinical, cancer care touchpoints for narrow-task AI applications and review a selection of applications. We describe the challenges faced in the clinical translation of AI and propose solutions. We also suggest paths forward in weaving AI into individualized patient care, with an emphasis on clinical validity, utility, and usability. By illuminating these issues in the context of current AI applications for clinical oncology, we hope to help advance meaningful investigations that will ultimately translate to real-world clinical use.

Introduction

Over the last decade, there has been a resurgence of interest for artificial intelligence (AI) applications in medicine. This is driven by the advent of deep learning algorithms, computing hardware advances, and the exponential growth of data that are being generated and used for clinical decision making[1–3]. Oncology is particularly poised for transformative changes brought on by AI, given the proven advantages of individualized care and recognition that tumors and their response rates differ vastly from person to person[4,5]. In oncology, much like other medical fields, the overarching goal is to increase quantity and quality of life, which, from a practical standpoint, entails choosing the management strategy that optimizes cancer control and minimizes toxicity.

As multidimensional data is increasingly being generated in routine care, AI can support clinicians to form an individualized view of a patient along their care pathway and ultimately guide clinical decisions. These decisions rely on the incorporation of disparate, complex datastreams, including clinical presentation, patient history, tumor pathology and genomics, as well as medical imaging, and marrying these data to the findings of an ever-growing body of scientific literature. Furthermore, these datastreams are in a constant state of flux over the course of a patient's trajectory. With the emergence of AI, specifically *deep learning* [2], there is now a computational basis to integrate and synthesize these data, to predict where the patient's care path is headed, and ultimately improve management decisions.

While there is much reason to be hopeful, numerous challenges remain to the successful integration of AI in clinical oncology. In analyzing these challenges, it is critical to view the promise, success, and failure of AI not only in generalities, but on a clinical case-by-case basis. Not every cancer problem is a nail to AI's hammer; its value is not universal, but inextricably linked to the clinical use case[6]. The current evidence suggests that clinical translation of the vast majority of published, high-performing AI algorithms remains in a nascent stage[7]. Furthermore, we posit that the imminent value of AI in clinical oncology is in the aggregation of narrow task-specific, clinically validated and meaningful applications at clinical "touchpoints" along the cancer care pathway, rather than general, all-purpose AI for end-to-end decision-making. As the global cancer incidence increases and the financial toxicity of cancer care is increasingly recognized, many societies are moving towards value-based care systems[8,9]. With development of these systems, there will be increasing incentive for the adoption of data-driven tools - potentially powered by AI - that can lead to reduced patient morbidity, mortality, and healthcare costs[10].

Here, we will describe the key concepts of AI in clinical oncology and review a selection of AI applications in oncology from the lens of a patient moving through clinical touchpoints along the cancer care path. We will therein describe the challenges faced in the clinical translation of AI and propose solutions, and finally suggest paths forward in weaving AI into individualized patient cancer care. By illuminating these issues in the context of current AI applications for clinical oncology, we hope to provide concepts to help drive meaningful investigations that will ultimately translate to real-world clinical use.

Artificial Intelligence: From Shallow to Deep Learning

The concept of AI, formalized in the 1950's, was originally defined as the ability of a machine to perform a task normally associated with human performance[11]. Within this field, the concept of machine learning was born, which refers to an algorithm's ability to learn data and perform tasks without explicit programming[12]. Machine learning research has led to development and use of a number of "shallow" learning algorithms, including earlier generalized linear models like logistic regression, Bayesian algorithms, decision-trees, and ensemble methods[13,14]. In the simplest of these models, such as logistic regression, input variables are assumed to be independent of one another, and individual weights are learned for each variable to determine a decision boundary that optimally separates classes of labelled data. More advanced shallow learning algorithms, such as random forests, allow for the characterization and weighting of input variable combinations and relationships, thus learning decision boundaries that can fit more complex data.

Deep learning is a newer subset of machine learning, which has the ability to learn patterns from raw, unstructured input data by incorporating layered neural networks[2]. In supervised learning, which represents the most common form within medical AI, a neural network will generate a prediction from this input data and compare it to a "ground truth" annotation. This discrepancy between prediction and ground truth is encapsulated in a loss function which is then propagated back through the neural network, and over numerous cycles, the model is optimized to minimize this loss function.

For the purpose of clinical application, we can view AI as a spectrum of algorithms, the utility of which are inextricably linked to the characteristics of the task under investigation. Thorough understanding of the data stream is necessary to choose, develop, and optimize an algorithm. In general, deep learning networks offer nearly limitless flexibility in input, output, architectural and parameter design, and thus are able to fit vast quantities of heterogeneous and unstructured data never before possible[15]. Specifically, deep learning has a high propensity to learn non-linear and high-dimensional relationships in multi-modal data including time series data, pixel-by-pixel imaging data, unstructured text data, audio/video data, or biometric data. Data with significant spatial and temporal heterogeneity are particularly well-suited for DLNNs[16]. On the other hand, this power comes at the expense of limited interpretability and a proclivity for overfitting data if not trained on a large enough dataset[17]. While traditional machine learning and statistical modeling can perform quite well at certain predictive tasks, they generally struggle to fit unprocessed, unstructured, and high dimensional data compared to deep learning. Therefore, despite its limitations, deep learning has opened the door to big data analysis in oncology and promises to advance clinical oncology, so long as certain pitfalls in development and implementation can be overcome.

Cancer Care as a Mathematical Optimization Problem

To appreciate the promise surrounding AI applications for clinical oncology, it is essential to incorporate a mathematical lens to the patient care path through cancer risk prediction, screening, diagnosis and treatment. From the AI perspective, the patient path is an optimization problem, wherein heterogeneous data streams converge as inputs into a mathematical scaffold (i.e. machine learning algorithms) **(Figure 1)**. This scaffold is iteratively adjusted during training until the desired output can be reliably predicted and an action can be taken. In this setting, an ever-growing list of inputs include patient clinical presentation, past medical history, genomics, imaging, and biometrics, and can be roughly subdivided as tumor, host, or environmental factors. The complexity of the algorithms is often driven by the quantity, heterogeneity, and dimensionality of such data. Outputs are centered, most broadly, on increasing survival and/or quality of life, but are often evaluated by necessity as a series of more granular surrogate endpoints.

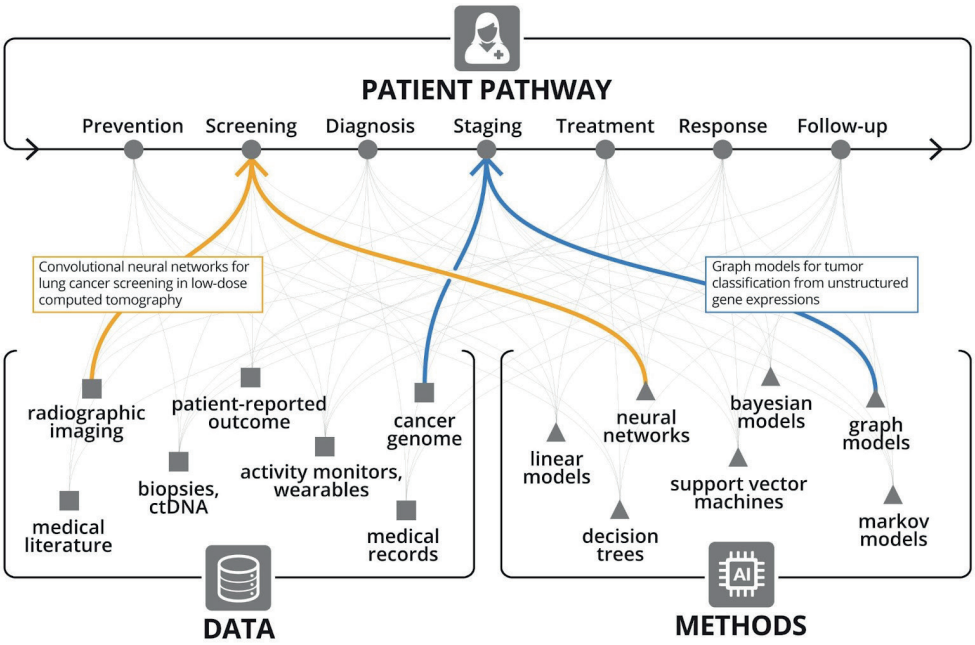

Figure 1. Narrow task-specific AI applications addressing a specific touchpoint along the patient pathway, and utilizing a specific data type and AI method.

Datastreams for Clinical Oncology

The arc of research in oncology, increasing data generation, and advances in computational technology have collectively resulted in a frameshift from low-dimensional to increasingly high-dimensional patient data representation. Earlier data and computational limitations often necessitated reducing unstructured patient data (e.g. medical images and biopsies) into a set of human-digestible discrete measures of disease extent. One notable example of such simplification lies within cancer staging systems, most prominently the AJCC TNM classification[18]. In 1977, with only three inputs commonly available - tumor size, nodal involvement, and presence of metastasis - the first edition AJCC TNM staging became standard of care for risk-stratification and decision-management in oncology. Over the subsequent decades, with the incorporation of other discrete data points, predictive nomograms could be generated using simple linear models, which have found practical use in certain situations[19–22]. More recently, improved methods to extract and analyze existing data coupled with new data streams and a growing understanding of inter- and intra-tumoral heterogeneity, have all led to the development of increasingly complex and specific stratification models. Key examples of novel data streams introduced over the past two decades are the Electronic Health Record, The Cancer Genome Atlas[23], The Cancer Imaging Archive[24], and the Project GENIE initiative[25]. Key examples of advanced risk-stratification and prediction models are the prostate cancer Decipher score[26] and breast cancer OncotypeDx score[27], which utilize discrete genomic data and shallow machine learning algorithms to form clinically validated predictive models. Useful oncology datastreams, roughly following historical order of availability, include: clinical presentation, tumor stage, histopathology, qualitative imaging, tumor genomics, patient genomics, quantitative imaging, liquid biopsies, electronic medical record mining, wearable devices, and digital behavior **(Figure 1)**. Furthermore, as a patient moves along the cancer care pathway, the number of influxing, intra-patient datastreams grows. With each step through the pathway, new data is generated out of the pathway with the potential to be reincorporated at a later time back into the pathway **(Figure 2)**.

As our biological knowledge base and datastreams grow in clinical oncology, machine learning algorithms can be deployed to learn patterns that apply to more and more precise patient groups and generate predictions to guide treatment for the next, "unseen" patient. As we assimilate more data, *optimal* cancer care, i.e. the care that results in the best survival and quality of life for a patient, inevitably becomes *precision* care, assuming we have the necessary tools to fully utilize the data. Here, at this intersection of data complexity and precision care in clinical oncology, is where the promise of AI has been so tantalizing, though as of yet, unfulfilled.

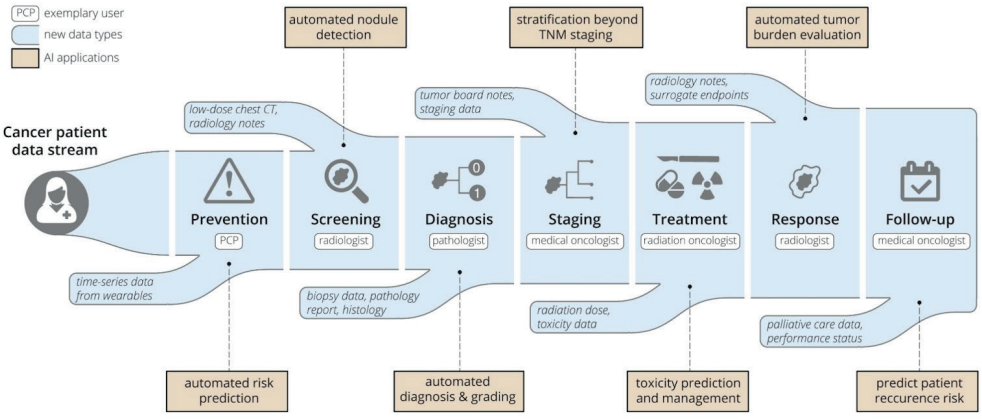

Figure 2. An example cancer patient pathway converges with an ever-increasing data stream. Potential AI applications and exemplary clinical users at each touchpoint are also illustrated.

AI Applications and Touchpoints Along the Clinical Oncology Care Path

We propose that AI development for clinical oncology should be approached from patient and clinician perspectives across the following cancer care touchpoints: Risk Prediction, Screening, Diagnosis, Prognosis, Initial Treatment, Response Assessment, Subsequent Treatment, and Follow-up **(Figure 2)**. The clinical touchpoint pathway shares features with the "cancer continuum,"[28] though it consists of more granular, patient and clinician decision-oriented points of contact for AI to add clinical benefit. Each of these touchpoints involves a critical series of decisions for oncologists and patients to make and yields a use-case for AI to provide an incremental benefit. Furthermore, touchpoint details will vary by cancer subtype. Within these touchpoints, ideal AI use-cases are ones with significant unmet need and large available datasets. In the context of supervised machine learning, these datasets require robust and accurate annotation to form a reliable "ground-truth" on which the AI system can train.

Narrow Tasks with High Reliability

As clinical oncology datastreams increase in complexity, the tools needed to discern patterns from these data are necessarily more complex, as well. Amidst this flood of heterogeneous *intra-patient* data, there is a relative dearth of *inter-patient* data which is needed to train large scale models. Therefore, to accumulate the training data required for generalizable models, it will likely be more fruitful to target and evaluate individual AI models towards specific datastreams at a particular touchpoint along the care pathway.

It is tempting to think that, given the increasing data streams that encompass multiple patient characteristics and outcomes, one could develop a unifying, dynamic model to synthesize and drive precision oncology, developing a "virtual-guide" of sorts for the oncologist and patient[29]. Analogies are often made to transformative technologies, such as self-driving cars and social media recommendations that leverage powerful neural networks on top of streams composed of billions of incoming data points, to predict real-time outcomes and continually improve performance. While in theory, this strategy could one day be deployed in a clinical setting, there are vast differences between these domains that question whether or not we *should* or even *could* pursue this strategy currently. One of the most glaring differences between the healthcare and technology domains, in terms of AI application, is the striking difference in data quality and quantity. While there has been a sea change in the collection of data within the healthcare field over the past decade, driven by the adoption of the Electronic Health Record, datasets still remain virtually siloed, intensely regulated, and, particularly in cancer care, much too small to leverage the most powerful AI algorithms available[30,31]. One of the most high-profile of these endeavors, IBM's Watson Oncology project, has attempted to develop a broad prediction machine to guide cancer care, but has been limited by suboptimal concordance with human oncologists' recommendations and subsequent distrust[32–34].

As our biological perspective has evolved, we now know that cancer is made up of thousands of distinct entities that will follow different trajectories, each with different treatment strategies[35,36]. In computational model development, there is thought to be a bare minimum number of data samples required for each model input feature[37]. As we seek to make recommendations more and more bespoke, it becomes more challenging to accrue the quantity of training data necessary to leverage complex algorithms. Fortunately, this data gap in healthcare is well-recognized, and a number of initiatives have been proposed to streamline and unify data collection[38]. However, given the innately heterogeneous, fragmented, and private nature of healthcare data, we in the oncology field may never achieve a level of data robustness enjoyed by other technology sectors. Therefore, strategies are necessary to mitigate the data problem, such as proper algorithm selection, model architecture improvements, data preprocessing, and data augmentation techniques. Above all, thoughtful selection of narrow use cases across cancer care touchpoints is paramount in order to yield clinical impact.

Once rigorously tested, these narrow-AIs could then be aggregated over the course of a patient's care to provide a measurable, clinical benefit. This sort of AI-driven dimensionality reduction of a patient's feature space allows for optimizing the development process and exporting of quality AI applications in the present environment of siloed data, expertise, and infrastructure. As of writing, there are approximately 20 FDA-approved AI applications targeted specifically for clinical oncology, and each of these performs a narrow task, utilizing a single data stream at a specific cancer care touchpoint[29,39,40] **(Table 1)**. We hypothesize that the future of AI in oncology will continue to consist of an aggregation of rigorously evaluated, narrow-task models, each one providing small, incremental benefits for patient quantity and quality of life. In the next sections, we will review select AI applications that have excelled with this narrow-task approach.

Table 1. FDA approvals to-date for deep learning applications in clinical oncology

	Name	Data type	Task	FDA summary	year
Thoracic/Liver					
1	Arterys Oncology DL	CT, MRI	segmentation of lung nodules and liver lesions, automated reporting	https://www.accessdata.fda.gov/cdrh_docs/pdf17/K173542.pdf	2017
2	Siemens AI-Rad Companion (Pulmonary)	CT	segmentation of lesions of the lung, liver, and lymph nodes	https://www.accessdata.fda.gov/cdrh_docs/pdf18/K183271.pdf	2019
3	Riverain ClearRead CT	CT	detection of pulmonary nodules in asymptomatic population	https://www.accessdata.fda.gov/cdrh_docs/pdf16/k161201.pdf	2016
4	Siemens syngo.CT Lung CAD	CT	detection of solid pulmonary nodules, alerts to overlooked regions	https://www.accessdata.fda.gov/cdrh_docs/pdf19/K193216.pdf	2020
5	GE Hepatic VCAR	CT	liver lesion segmentation and measurement	https://www.accessdata.fda.gov/cdrh_docs/pdf19/K193281.pdf	2020
6	Coreline AView LCS	CT	Characterization of nodule type, location, measurements, and Lung-RADS category	https://www.accessdata.fda.gov/cdrh_docs/pdf20/K201710.pdf	2020
7	MeVis Veolity	CT	detection of solid pulmonary nodules, alerts to overlooked regions	https://www.accessdata.fda.gov/cdrh_docs/pdf20/K201501.pdf	2021
8	Philips Lung Nodule Assessment and Comparison Option (LNA)	CT	Characterization of nodule type, location, and measurements	https://www.accessdata.fda.gov/cdrh_docs/pdf16/K162484.pdf	2017
9	NinesMeasure	CT	Characterization of nodule type, location, and measurements	https://www.accessdata.fda.gov/cdrh_docs/pdf20/K202990.pdf	2021
Breast					
10	iCAD ProFound AI	3D DBT mammography	detection of soft tissue densities and calcifications	https://www.accessdata.fda.gov/cdrh_docs/pdf19/K191994.pdf	2019
11	cmTriage	2D FFDM	triage and passive notification	https://www.accessdata.fda.gov/cdrh_docs/pdf18/K183285.pdf	2019
12	Screenpoint Transpara	FFDM	detection of suspicious soft tissue lesions and calcifications	https://www.accessdata.fda.gov/cdrh_docs/pdf19/K192287.pdf	2019
13	Zebra Medical Vision HealthMammo	2D FFDM	triage and passive notification	https://www.accessdata.fda.gov/cdrh_docs/pdf20/K200905.pdf	2020
14	Koios DS for Breast	US	classification of lesion shape, orientation, and BI-RADS category	https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190442.pdf	2019
15	Hologic Genius AI Detection	DBT mammography	detection of suspicious soft tissue lesions and calcifications	https://www.accessdata.fda.gov/cdrh_docs/pdf20/K201019.pdf	2020

Name		Data type	Task	FDA summary	year
16	Therapixel MammoScreen	FFDM	detection of suspicious findings and level of suspicion	https://www.accessdata.fda.gov/cdrh_docs/pdf19/K192854.pdf	2020
17	QuantX	MRI	image registration, automated segmentation and analysis of user-selected regions of interest	https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN170022.pdf	2020
18	ClearView cCAD	US	classification of shape and orientation of user-defined regions, and BI-RADS category	https://www.accessdata.fda.gov/cdrh_docs/pdf16/K161959.pdf	2016
Prostate					
19	Quantib Prostate	MRI	semi-automatic segmentation of anatomical structures, volume computations, automated PI-RADS category	https://www.accessdata.fda.gov/cdrh_docs/pdf20/K202501.pdf	2020
20	GE PROView	MRI	prediction of PI-RADS category	https://www.accessdata.fda.gov/cdrh_docs/pdf19/K193306.pdf	2020
CNS					
21	Cortechs NeuroQuant	MRI	automated segmentation and volumetric quantification of brain lesions	https://www.accessdata.fda.gov/cdrh_docs/pdf17/K170981.pdf	2017

Narrow-Task AI Examples Across the Clinical Oncology Touchpoints

T1. Risk Prediction and Prevention. Given the burden to people and healthcare systems of cancer diagnosis and management, there is a significant opportunity for AI to help predict an individual's risk of developing cancer, and thereby target screening and early interventions effectively and efficiently. In a mathematical sense, the patient's entire personal history up until diagnosis makes up a vast and extremely heterogeneous datastream to be evaluated, positioning deep learning to have an impact. This is evidenced by the steady development of tools that leverage computational modeling to refine cancer risk. In the past few years, several DL algorithms have been investigated to further tailor risk prediction beyond traditional models. Some of these algorithms utilize novel datastreams that were not available until recently: satellite imagery[41], internet search history[42], and wearable devices[43]. Others maximize the utility of pre-existing datastreams, including patient genomics, routine imaging, unstructured health record data, and deeper family history to improve predictions[44].

T2. Screening. Cancer screening involves the input and evaluation of data at a distinct time-point to determine whether or not additional diagnostic testing and procedures are warranted. Datastreams can be in the form of serum markers, medical imaging, or visual or endoscopic examination. Each of these modalities provides opportunities for the integration of AI to improve prediction of cancer. For serum markers, such as prostate specific antigen (PSA), early research suggests that machine learning algorithms modeling PSA at different timepoints, in conjunction with other serum markers, may be able to better predict the presence of prostate cancer than PSA alone[45]. Perhaps more than in any other application, AI has found high impact use in medical imaging screening. Narrow-task models have been developed to localize lesions and predict risk of malignancy on lung cancer CT[46] and breast cancer mammography[47], with applications that have been shown to perform on par, or sometimes better than expert diagnosticians[48]. In these applications, raw pixel data of the image is utilized as input into a deep learning convolutional neural network that is trained based on radiologist-labelled ground-truth outputs. Importantly, while the algorithms demonstrate impressive results in terms of area under the curve, sensitivity and specificity, they do not evaluate direct clinical endpoints, such as cancer mortality, healthcare costs, or quality of life. Outside of medical imaging, AI has found utility in screening endoscopy for colorectal carcinoma, with an application that guides biopsy site selection[49,50]. Furthermore, there are opportunities to improve diagnostic yield for other malignancies for which screening has been traditionally difficult and unproven. This could be accomplished by AI improving analysis of pre-existing datastreams, such as abdominal CT or MRI imaging, or via its ability to integrate multi-modal datastreams, like EHR and genomic data. While currently the United States Preventive Services Task Force (USPSTF) recommends against screening for many cancers[51], there are a number of ongoing investigations to determine if incorporation of AI into screening criteria and technology may allow screening to be utilized in a wider array of disease sites, such as pancreatic cancer.

T3. Diagnosis. Diagnosing involves the exclusion of other benign disease processes and the characterization of cancer by primary site, histopathology, and increasingly, genomic classification. Diagnosis represents an AI touchpoint for these three domains by analyzing their respective datastreams: including clinical exam and medical imaging (i.e. Radiomics), digital pathology, and genomic sequencing. A key study that revealed the promise of deep learning for cancer diagnosis showed that convolutional neural networks could achieve dermatologist level accuracy in the classification of skin cancers utilizing digital photographs[15]. Other promising areas of investigation in this realm include non-invasive brain tumor diagnosis[52] and prostate cancer Gleason grading[53] via MRI, automated histopathologic diagnosis for breast cancer[54] and prostate cancer[55] , and utilization of radiographic and histopathologic data to predict underlying genomic classification[56]. Thus far, the *Screening and Diagnosis* touchpoints account for nearly all FDA-approved AI applications for clinical oncology, with three algorithms focusing on mammography and three focusing on CT-based lesion diagnosis[39] .

T4. Risk Stratification and Prognosis. Historically, risk-stratification consisted of TNM staging, though increasingly additional datastreams such as genomics, advanced imaging, and serum markers have allowed for more precise risk stratification. Given the vast heterogeneity in cancer risk, risk-stratification presents a highly attractive use case for AI. Over the past two decades, genomic classifiers, developed with machine learning, have been integrated into risk-stratification for a number of malignancies. Classifiers such as OncotypeDx for breast cancer, a logistic regression based classifier, and the Decipher score, a random forest-based classifier, have demonstrated the ability to improve prognostication[57] and guide treatment[58]. The Decipher score genomic classifier is based on 22 genomic expression markers input into a random forest model that was trained to predict metastasis after prostatectomy for patients with prostate cancer at a single institution[26]. This classifier has been subsequently validated in several external settings, and is now undergoing investigation in several randomized control trials (NCT04513717; NCT02783950). Deep learning strategies have been explored to integrate multi-omic data sources into risk-stratification models utilizing combinations of diagnostic imaging[59], EHR data[43,60], and genomic information[61]. Furthermore, there is the potential for deep learning to better risk-stratify patients based on large population databases, such as the Surveillance, Epidemiology, and End Results Program, by learning non-linear relationships between database variables, though preliminary efforts require validation[62].

T5. Initial Treatment Strategy. The formulation of initial treatment strategy is arguably the most pivotal touchpoint for AI in the cancer pathway, as it directly influences patient management. The last two decades have seen exponential growth in the number and complexity of initial treatment options for common cancers[3]. A common predicament for initial treatment is what combination of systemic therapy, radiotherapy, and surgery is optimal for a given patient. Machine learning methods utilizing genomic[63] and radiomic data[64] have been investigated to predict radiation sensitivity. While immunotherapy has been adopted in an increasing number of disease settings, it remains difficult to predict response based on currently available biomarkers, and machine learning algorithms with radiomic input have demonstrated the ability to improve response prediction[65].

Furthermore, deep learning has demonstrated the ability to analyze multi-modal datastreams within the genomic realm: a recent analysis demonstrated that integration of tumor mutational burden, copy number alteration, and microsatellite instability code can help predict response to immunotherapy[66]. AI is also enabling more accurate "evidence-based treatment". Natural language processing and powerful language models can help analyze published scientific works and utilize existing oncology literature e.g. extracting medical oncology concepts from EHR and linking these to a literature corpus[67].

T6. Response Assessment. Assessment of response to treatment generally includes radiographic and clinical assessments. Quantitative response assessment criterias like RECIST and RANO have long been established as reproducible ways to assess response to therapy, though in the age of targeted immunotherapies, validity has been questioned[68]. As targeted therapeutics and immunotherapies have entered the clinic, however, it has become clear that response assessment via RECIST is inadequate, due to phenomena such as pseudoprogression[69]. Detailed response assessment is often a time intensive process that requires a high degree of human expertise and experience, not to mention high intra- and inter-reader variability. Additionally, despite periodic review and revision of these criteria, they remain inapt at capturing edge cases, such as variable lesion response, in the case of patients receiving immunotherapy. Deep learning has demonstrated potential for automated response assessment, including automated RANO assessment[70] and RECIST response in patients undergoing immunotherapy [71].

T7. Subsequent Treatment Strategy. When approaching AI algorithm development for subsequent treatment strategy, there are a number of specific considerations that generate complexity as compared to from initial treatment strategy. Firstly, there are additional datastreams to consider, such as prior treatments, treatment-related toxicity, restaging imaging, and often multiple tissue specimens. Given the heterogeneity in datastreams and the shrinking patient populations from which to build these models, subsequent treatment strategy is a challenging space for evidence-based decision-making, and in turn, for reliable AI applications. Algorithms that utilize longitudinal follow-up information may help here. In one example, AI has demonstrated the ability to synthesize serial CT follow-up imaging for lung cancer patients post-chemoradiation, and demonstrated the ability to predict later recurrence[72]. An intervention such as this could guide selection for patients to undergo consolidative treatments like surgery or immunotherapy.

T8. Follow-up. Another underexplored area for AI oncologic applications is development of tools to guide precision follow-up. Diagnostic and screening algorithms may often be transferable to the follow-up setting, but will require retraining and validation for the task of interest. Similar to T7, the effect of prior cancer treatment on the datastream will often shift things significantly. For example, radiomic features extracted from the same tumor, pre- and post-treatment, show significant discrepancies[73]. These "delta" features could be used to predict patient recurrence risk and late toxicity, helping to tailor follow-up plans[74]. Appropriately triaging patients for escalated follow-up and attention can promote decreased morbidity and more efficient healthcare resource utilization; AI leveraging EHR data has demonstrated the ability to accomplish this, by

selecting patients at high risk for acute care visit while undergoing cancer therapy and assigning them to an escalated preventative care strategy[75]. In cases where patients have untreatable relapse, end-of-life care becomes an extremely important and challenging process. AI has shown potential here as well, as a way to triage patients at high risk of mortality and nudge physicians to converse with patients regarding their values, wishes, and quality of life options[76].

Challenges for Clinical Translation: Beyond Performance Validation

While tremendous strides have been made in the development of oncologic AI, as evidenced by the surge in publications and published datasets in recent years, there remains a large gap between evidence for AI performance and evidence for clinical impact. While there have been thousands of published studies of deep learning algorithm performance[77], a recent systematic review, found only nine prospective trials, and two published randomized clinical trials of deep learning in medical imaging[7].

As alluded to above, perhaps the defining barrier to development of clinical AI applications in oncology, and healthcare overall, is data limitation, both in quality and quantity. The problems with data curation, aggregation, transparency, bias, and reliability have been well-described[78,79]. Additionally, the lack of AI model interpretability, trust, reproducibility, and generalizability have received ample, and well-justified attention[80]. While all of these challenges must be overcome for successful AI development, here we will introduce several concepts specific to clinical translation of models that have already succeeded in preliminary stages of development and validation: *clinical validity, utility, and usability* **(Figure 3)**. Incorporation of these concepts into model design and evaluation is easy to overlook, yet is critical to move clinical AI beyond the research and development stage into real-world cancer care.

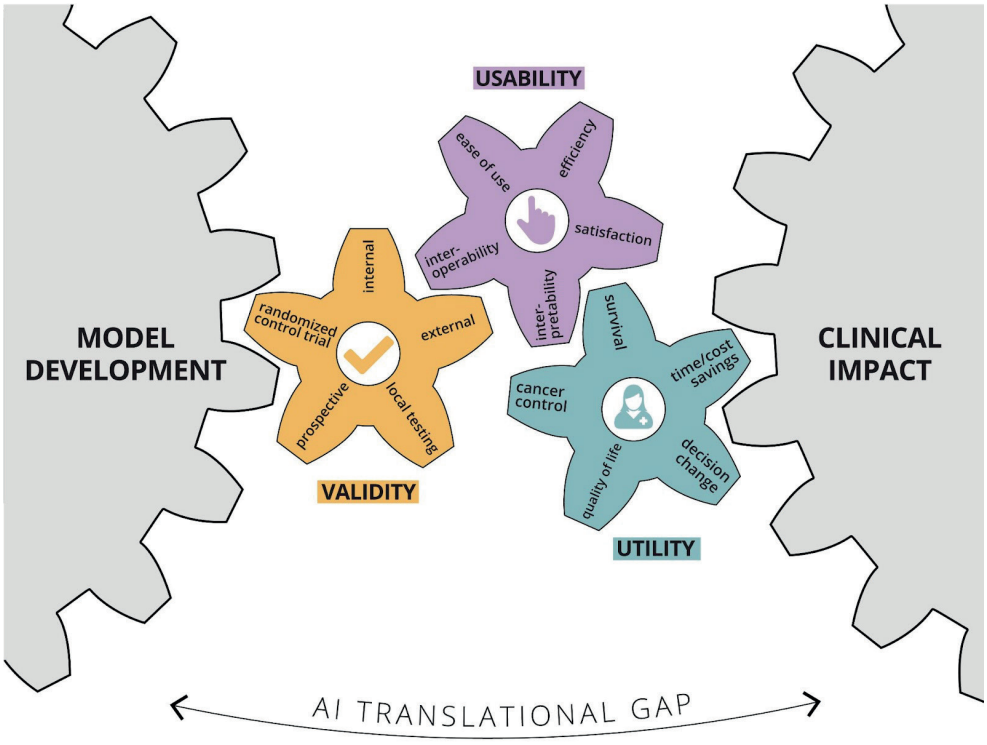

Figure 3. Bridging the AI translational gap between initial model development and routine clinical cancer care by emphasizing and demonstrating three essential concepts: clinical validity, utility, and usability.

To demonstrate *clinical validity*, a model is often evaluated in the following general sequence: internal validation, external validation, prospective testing, and local testing in the real-world population of interest[81]. Recently developed guidelines such as FAIR data, CONSORT/SPIRTAI, and the (in development) TRIPOD-AI checklists should be followed to ensure reproducibility, transparency, and methodologic rigor[82]. These guidelines are an important step forward in standardizing AI model development pathways and establishing a basis to determine AI study methodological rigor. While the vast majority of AI published reports include an internal, blinded test set, far fewer utilize an external validation set, and an even smaller proportion employ prospective testing and benchmark comparisons with human experts[83]. Given the lack of hypothesize-driven feature selection in most AI models, performance in real-world scenarios can vary dramatically if the test data distribution varies from the training data[84]. For this reason, multiple external validation sets are of utmost importance. Beyond this, it is often difficult to predict how a model will perform on edge cases - those that were under-represented in training data[85]. In the practice of oncology, detection of rare findings can be critical to safe cancer care, and thus must be taken into account to demonstrate a model is clinically valid. One way to mitigate the risk of model failure in real-world

use, is to conduct trial, run-in periods of "silent" prospective testing in the scenario of interest[86]. If a model performs well in the run-in period, there is some assurance that it will be safe to utilize, though its performance on extremely rare cases may be still difficult to presume.

Demonstrating *clinical utility*, requires clinical validity as a prerequisite, but goes beyond performance validation to the testing of clinically meaningful endpoints. High performance on commonly used endpoints, such as area under the receiver operating characteristic curve, sensitivity, or specificity, may suffice for certain diagnostic applications, but real-world impact will require validation of clinical endpoints as appropriate for each touchpoint along the care pathway. In the case of oncology, this includes overall survival, disease control, toxicity reduction, quality of life improvement, and decrease of healthcare resource utilization. Testing of these endpoints should be ideally performed in the setting of a randomized trial. The gold standard would be randomizing patients to the AI intervention and directly comparing clinical endpoints. A few of these trials have been completed, with one notable example involving testing accuracy for polyp detection rate on colonoscopy[87]. In this study, the primary outcome was adenoma detection rate. Despite demonstrating the superiority of the AI systems, downstream clinical benefit in terms of quality of life or survival requires yet further investigation. Another approach to AI clinical trials is to apply a validated model to all patients for risk-stratification, and then to apply randomized interventions. This was pursued successfully in a trial that utilized EHR data to predict patients at high-risk for emergency department (ED) visits during radiotherapy[75]. High-risk patients were then randomized to usual care, or extra preventative provider visits. It was found that high-risk patients randomized to extra visits had significantly fewer ED and hospital admissions, while low-risk patients had uniformly low rates of ED and hospital admissions without extra care. While providing a lower level of clinical utility evidence than a true randomized trial, this type of study strategy is attractive and practical for AI-based risk-prediction models, which make up a large proportion of AI models in development. Randomized clinical trials are notoriously difficult and time-consuming to execute, and AI interventions have unique characteristics that make such undertakings even more daunting. Notably, AI models are able to adapt to new data and improve over time; how would one take this into account in a traditional randomized trial? While we need AI to embrace randomized trials to truly prove clinical utility, it may be time to recognize that a re-imagining of the traditional randomized clinical trial may be necessary to appropriately study the benefits of AI applications[88].

Beyond validation of clinically meaningful endpoints, demonstrating *clinical usability* involves study of the AI model in a real-world setting, where it interfaces with clinical practitioners and patients. Evaluation of effects of the model on timed tasks, user satisfaction, and acceptance of AI recommendations should be performed[89]. A mechanism of feedback should be integrated into the design of the platform to identify weak points and opportunities for improved interface[90]. Additionally, interoperability between systems at the facility-to-facility, intra-facility, and point-of-care levels are crucial to streamline workflow[91]. Usability issues are also specific to the datastreams being analyzed. New datastreams such as mobile health data and wearable activity

monitors each present unique challenges to usability and adoption[43]. A key component of promoting usability is interpretability of the AI algorithm. As data streams become more dimensional, it is increasingly difficult to discern a biological or clinical rationale supporting an algorithm's predictions. This "black-box" effect may be acceptable in certain consumer electronics industries, but due to the consequential and medicolegal nature of healthcare decision-making, lack of interpretability poses a tremendous barrier to clinical use[92,93]. Fortunately, there is a growing research field dedicated to investigation of interpretability issues, and several techniques, such as saliency maps, hidden-states analysis, variable importance metrics, and feature visualizations can illuminate some aspects of AI prediction rationale[94,95]. Beyond this, an appreciation of advances in Human Factors research and collaboration with appropriate experts can help streamline the adoption of otherwise clinically validated algorithms. Finally, translating algorithms into clinical usable solutions requires robust information technology support services that may require dedicated investment from clinical institutions and departments.

Another key concept related to clinical usability is addressing the challenges that emerge when multiple AI models are deployed sequentially or simultaneously at a given touchpoint or series of touchpoints. *Orchestration* of these situations, which are expected to become more common, require attention to end-user responsibilities, interoperability, access, and training. As a patient moves through the oncology care path, they interact (directly or indirectly) with many different care providers who may be the primary users of a given AI application **(Figure 2)**. These users may have a primarily diagnostic or therapeutic role (or both). From a simplified perspective, the primary diagnosticians of the cancer care path are pathologists and radiologists, while the therapeutic clinicians tend to be medical, radiation, and surgical oncologists. Multidisciplinary touchpoints along the pathway, e.g. tumor boards, represent opportunities to collate and orchestrate disparate AI applications. In addition to physicians, there are numerous advanced practice providers such as nurses and physician assistants, as well as therapists, social workers, and medical students, who may be users of a specific AI application. If, for example, a patient receives a CT scan with an AI-generated prediction of malignancy, and this prediction is subsequently utilized as input for another algorithm to recommend surgery as treatment, who is the "designated user" primarily responsible for utilizing and disseminating that information? A further issue, which logically follows, is who is legally liable for decisions based on the use of the model. Specific solutions have not yet been developed to address these issues, and are, unfortunately, likely to arise on an ad hoc, case by case basis. This clinical orchestration of AI models merits further resources, investigation, and guidelines aimed at medical AI developers and cancer care providers to navigate these complex issues.

Despite the vanishingly few FDA-approved AI applications for oncologic indications, with numerous applications in the pipeline, there is high interest in streamlining ways to bridge the gap between development and clinical translation. Accordingly, the FDA is in the process of devising AI and machine learning-specific guidelines for approved clinical use. The recently released action plan incorporates the above clinical concepts and sets the stage for further defining a framework for safe AI translation to the clinic[96].

Conclusions

Increasing datastreams and advances in computational algorithms have positioned artificial intelligence to improve clinical oncology via rigorously evaluated, narrow-task applications interacting at specific touchpoints along the cancer care path. While there are a number of promising artificial intelligence applications for clinical oncology in development, substantial challenges remain to bridge the gap to clinical translation. The most successful models have leveraged large-scale, robustly annotated datasets for narrow tasks at specific cancer care touchpoints. Further development of artificial intelligence applications for cancer care should emphasize clinical validity, utility, and usability. Successful incorporation of these concepts will require bringing a patient-provider, clinical decision-centric focus to model development and evaluation.

Acknowledgements

The authors acknowledge financial support from NIH (HA: NIH-USA U24CA194354, NIH-USA U01CA190234, NIH-USA U01CA209414, and NIH-USA R35CA22052; BK: NIH-K08:DE030216), the European Union - European Research Council (HA: 866504), as well as the Radiological Society of North America (BK: RSCH2017).

References

1. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
2. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* vol. 521 436–444 (2015).
3. Kann, B. H., Johnson, S. B., Aerts, H. J. W. L., Mak, R. H. & Nguyen, P. L. Changes in Length and Complexity of Clinical Practice Guidelines in Oncology, 1996-2019. *JAMA Netw Open* **3**, e200841 (2020).
4. Schilsky, R. L. Personalized medicine in oncology: the future is now. *Nat. Rev. Drug Discov.* **9**, 363–366 (2010).
5. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* **12**, 323–334 (2012).
6. Maddox, T. M., Rumsfeld, J. S. & Payne, P. R. O. Questions for Artificial Intelligence in Health Care. *JAMA* **321**, 31–32 (2019).
7. Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).
8. Porter, M. E. A strategy for health care reform--toward a value-based system. *N. Engl. J. Med.* **361**, 109–112 (2009).
9. Yousuf Zafar, S. Financial Toxicity of Cancer Care: It's Time to Intervene. *J. Natl. Cancer Inst.* **108**, (2016).
10. Kuznar, W. The Push Toward Value-Based Payment for Oncology. *Am Health Drug Benefits* **8**, 34 (2015).
11. Russell, I. & Haller, S. Introduction: Tools and Techniques of Artificial Intelligence. *International Journal of Pattern Recognition and Artificial Intelligence* vol. 17 685–687 (2003).
12. Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* **3**, 210–229 (1959).
13. Bhattacharyya, R. *et al.* Personalized Network Modeling of the Pan-Cancer Patient and Cell Line Interactome. *Cold Spring Harbor Laboratory* 806596 (2019) doi:10.1101/806596.
14. Richens, J. G., Lee, C. M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun.* **11**, 3923 (2020).
15. Esteva, A. *et al.* Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **546**, 686 (2017).
16. Zhong, G., Ling, X. & Wang, L. From shallow feature learning to deep learning: Benefits from the width and depth of deep architectures. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **9**, e1255 (2019).
17. Zhu, X., Vondrick, C., Fowlkes, C. & Ramanan, D. Do We Need More Training Data? *arXiv [cs.CV]* (2015).
18. Amin, M. B. *et al.* The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more 'personalized' approach to cancer staging. *CA Cancer J. Clin.* **67**, 93–99 (2017).
19. Stephenson, A. J. *et al.* Predicting the outcome of salvage radiation therapy for recurrent prostate cancer after radical prostatectomy. *J. Clin. Oncol.* **25**, 2035–2041 (2007).

20. Bari, A. *et al.* Prognostic models for diffuse large B-cell lymphoma in the rituximab era: a never-ending story. *Ann. Oncol.* **21**, 1486–1491 (2010).

21. Mittendorf, E. A. *et al.* Incorporation of Sentinel Lymph Node Metastasis Size Into a Nomogram Predicting Nonsentinel Lymph Node Involvement in Breast Cancer Patients With a Positive Sentinel Lymph Node. *Annals of Surgery* vol. 255 109–115 (2012).

22. Creutzberg, C. L. *et al.* Nomograms for Prediction of Outcome With or Without Adjuvant Radiation Therapy for Patients With Endometrial Cancer: A Pooled Analysis of PORTEC-1 and PORTEC-2 Trials. *International Journal of Radiation Oncology*Biology*Physics* **91**, 530–539 (2015).

23. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113 (2013).

24. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).

25. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).

26. Erho, N. *et al.* Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS One* **8**, e66855 (2013).

27. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).

28. Chambers, D. A., Vinson, C. A. & Norton, W. E. *Advancing the Science of Implementation across the Cancer Continuum*. (Oxford University Press, 2018).

29. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).

30. Bi, W. L. *et al.* Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J. Clin.* **69**, 127–157 (2019).

31. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).

32. Somashekhar, S. P. *et al.* Abstract S6-07: Double blinded validation study to assess performance of IBM artificial intelligence platform, Watson for oncology in comparison with Manipal multidisciplinary tumour board – First study of 638 breast cancer cases. *Cancer Res.* **77**, S6–07–S6–07 (2017).

33. Lee, W.-S. *et al.* Assessing Concordance With Watson for Oncology, a Cognitive Computing Decision Support System for Colon Cancer Treatment in Korea. *JCO Clin Cancer Inform* **2**, 1–8 (2018).

34. Gyawali, B. Does global oncology need artificial intelligence? *Lancet Oncol.* **19**, 599–600 (2018).

35. Polyak, K. Heterogeneity in breast cancer. *J. Clin. Invest.* **121**, 3786–3788 (2011).

36. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).

37. Mitsa, T. How Do You Know You Have Enough Training Data? *Towards Data Science* https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee (2019).

38. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).

39. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* **3**, 118 (2020).

40. Hamamoto, R. *et al.* Application of Artificial Intelligence Technology in Oncology: Towards the Establishment of Precision Medicine. *Cancers* **12**, (2020).

41. Bibault, J.-E., Bassenne, M., Ren, H. & Xing, L. Deep Learning Prediction of Cancer Prevalence from Satellite Imagery. *Cancers* **12**, (2020).

42. White, R. W. & Horvitz, E. Evaluation of the Feasibility of Screening Patients for Early Signs of Lung Carcinoma in Web Search Logs. *JAMA Oncol* **3**, 398–401 (2017).

43. Beg, M. S., Gupta, A., Stewart, T. & Rethorst, C. D. Promise of Wearable Physical Activity Monitors in Oncology Practice. *Journal of Oncology Practice* vol. 13 82–89 (2017).

44. Ming, C. *et al.* Machine learning-based lifetime breast cancer risk reclassification compared with the BOADICEA model: impact on screening recommendations. *Br. J. Cancer* **123**, 860–867 (2020).

45. Nitta, S. *et al.* Machine learning methods can more efficiently predict prostate cancer compared with prostate-specific antigen density and prostate-specific antigen velocity. *Prostate Int* **7**, 114–118 (2019).

46. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).

47. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).

48. Salim, M. *et al.* External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncol* **6**, 1581–1588 (2020).

49. Zhou, D. *et al.* Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer. *Nat. Commun.* **11**, 2961 (2020).

50. Guo, L. *et al.* Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). *Gastrointest. Endosc.* **91**, 41–51 (2020).

51. USPSTF: A and B Recommendations. https://www.uspreventiveservicestaskforce.org/uspstf/recommendation-topics/uspstf-and-b-recommendations.

52. Chang, K. *et al.* Residual Convolutional Neural Network for the Determination of IDH Status in Low- and High-Grade Gliomas from MR Imaging. *Clin. Cancer Res.* **24**, 1073–1081 (2018).

53. Schelb, P. *et al.* Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. *Radiology* **293**, 607–617 (2019).

54. Ehteshami Bejnordi, B. *et al.* Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* **318**, 2199–2210 (2017).

55. Nagpal, K. *et al.* Development and Validation of a Deep Learning Algorithm for Gleason Grading of Prostate Cancer From Biopsy Specimens. *JAMA Oncol* **6**, 1372–1380 (2020).

56. Lu, C.-F. *et al.* Machine Learning–Based Radiomics for Molecular Subtyping of Gliomas. *Clin. Cancer Res.* **24**, 4429–4436 (2018).

57. Spratt, D. E. *et al.* Individual Patient-Level Meta-Analysis of the Performance of the Decipher Genomic Classifier in High-Risk Men After Prostatectomy to Predict Development of Metastatic Disease. *J. Clin. Oncol.* **35**, 1991–1998 (2017).

58. Sparano, J. A. *et al.* Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. *N. Engl. J. Med.* **379**, 111–121 (2018).

59. Kann, B. H. *et al.* Multi-Institutional Validation of Deep Learning for Pretreatment Identification of Extranodal Extension in Head and Neck Squamous Cell Carcinoma. *J. Clin. Oncol.* **38**, 1304–1311 (2020).

60. Manz, C. R. *et al.* Validation of a Machine Learning Algorithm to Predict 180-Day Mortality for Outpatients With Cancer. *JAMA Oncol* (2020) doi:10.1001/jamaoncol.2020.4331.

61. Qiu, Y. L., Zheng, H., Devos, A., Selby, H. & Gevaert, O. A meta-learning approach for genomic survival analysis. *Nat. Commun.* **11**, 6350 (2020).

62. She, Y. *et al.* Development and Validation of a Deep Learning Model for Non-Small Cell Lung Cancer Survival. *JAMA Netw Open* **3**, e205842 (2020).

63. Scott, J. G., Harrison, L. B. & Torres-Roca, J. F. Genomic biomarkers for precision radiation medicine – Authors' reply. *The Lancet Oncology* vol. 18 e239 (2017).

64. Lou, B. *et al.* An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction. *The Lancet Digital Health* vol. 1 e136–e147 (2019).

65. Sun, R. *et al.* A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol.* **19**, 1180–1191 (2018).

66. Xie, C. *et al.* Immune Checkpoint Blockade in Combination with Stereotactic Body Radiotherapy in Patients with Metastatic Pancreatic Ductal Adenocarcinoma. *Clin. Cancer Res.* **26**, 2318–2326 (2020).

67. Simon, G. *et al.* Applying artificial intelligence to address the knowledge gaps in cancer care. *Oncologist* **24**, 772 (2019).

68. Villaruz, L. C. & Socinski, M. A. The clinical viewpoint: definitions, limitations of RECIST, practical considerations of measurement. *Clin. Cancer Res.* **19**, 2629–2636 (2013).

69. Gerwing, M. *et al.* The beginning of the end for conventional RECIST — novel therapies require novel imaging approaches. *Nature Reviews Clinical Oncology* vol. 16 442–458 (2019).

70. Kickingereder, P. *et al.* Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* **20**, 728–740 (2019).

71. Arbour, K. C. *et al.* Deep Learning to Estimate RECIST in Patients with NSCLC Treated with PD-1 Blockade. *Cancer Discov.* (2020) doi:10.1158/2159-8290.CD-20-0419.

72. Xu, Y. *et al.* Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clin. Cancer Res.* **25**, 3266–3275 (2019).

73. van Dijk, L. V. *et al.* Delta-radiomics features during radiotherapy improve the prediction of late xerostomia. *Sci. Rep.* **9**, 12483 (2019).

74. Chang, Y. *et al.* An investigation of machine learning methods in delta-radiomics feature analysis. *PLOS ONE* vol. 14 e0226348 (2019).

75. Hong, J. C. *et al.* System for High-Intensity Evaluation During Radiation Therapy (SHIELD-RT): A Prospective Randomized Study of Machine Learning–Directed Clinical Evaluations During Radiation and Chemoradiation. *Journal of Clinical Oncology* vol. 38 3652–3661 (2020).

76. Ramchandran, K. J. *et al.* A predictive model to identify hospitalized cancer patients at risk for 30-day mortality based on admission criteria via the electronic medical record. *Cancer* vol. 119 2074–2080 (2013).

77. Kann, B. H., Thompson, R., Thomas, C. R., Jr, Dicker, A. & Aneja, S. Artificial Intelligence in Oncology: Current Applications and Future Directions. *Oncology* **33**, 46–53 (2019).

78. Norgeot, B. *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.* **26**, 1320–1324 (2020).

79. Thompson, R. F. *et al.* Artificial intelligence in radiation oncology: A specialty-wide disruptive transformation? *Radiother. Oncol.* **129**, 421–426 (2018).

80. Beam, A. L., Manrai, A. K. & Ghassemi, M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA* **323**, 305–306 (2020).

81. Park, Y. *et al.* Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* **3**, 326–331 (2020).

82. Liu, X., Faes, L., Calvert, M. J., Denniston, A. K. & CONSORT/SPIRIT-AI Extension Group. Extension of the CONSORT and SPIRIT statements. *The Lancet* vol. 394 1225 (2019).

83. Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y. & Park, S. H. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. *Korean J. Radiol.* **20**, 405–410 (2019).

84. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V. & Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognit.* **45**, 521–530 (2012).

85. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Ré, C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *Proc ACM Conf Health Inference Learn (2020)* **2020**, 151–159 (2020).

86. Kang, J., Morin, O. & Hong, J. C. Closing the Gap Between Machine Learning and Clinical Cancer Care—First Steps Into a Larger World. *JAMA Oncol* **6**, 1731–1732 (2020).

87. Wang, P. *et al.* Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* **68**, 1813–1819 (2019).

88. Haring, A. In the age of machine learning randomized controlled trials are unethical. *Towards Data Science* https://towardsdatascience.com/in-the-age-of-machine-learning-randomized-controlled-trials-are-unethical-74acc05724af (2019).

89. Kumar, A. *et al.* Usability of a Machine-Learning Clinical Order Recommender System Interface for Clinical Decision Support and Physician Workflow. *medRxiv* (2020).

90. Cutillo, C. M. *et al.* Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *npj Digital Medicine* vol. 3 (2020).

91. He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).

92. Doshi-Velez, F. & Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv [stat.ML]* (2017).

93. Wang, F., Kaushal, R. & Khullar, D. Should Health Care Demand Interpretable Artificial Intelligence or Accept 'Black Box' Medicine? *Ann. Intern. Med.* **172**, 59–60 (2020).

94. Olah, C. *et al.* The building blocks of interpretability. *Distill* **3**, (2018).

95. Guo, T., Lin, T. & Antulov-Fantulin, N. Exploring Interpretable LSTM Neural Networks over Multi-Variable Data. *arXiv [cs.LG]* (2019).

96. FDA. Artificial Intelligence/Machine Learning (AI/ML)-Based.:Jf/<X Software as a Medical Device (SaMD) Action Plan. (2021).

3

Chapter 3

Artificial Intelligence in Radiology

A Hosny, C Parmar, J Quackenbush, LZ Schwartz & HJWL Aerts

Nature Reviews Cancer 2018

Abstract

Artificial intelligence (AI) algorithms, particularly deep learning, have demonstrated remarkable progress in image-recognition tasks. Methods ranging from convolutional neural networks to variational autoencoders have found myriad applications in the medical image analysis field, propelling it forward at a rapid pace. Historically, in radiology practice, trained physicians visually assessed medical images for the detection, characterization and monitoring of diseases. AI methods excel at automatically recognizing complex patterns in imaging data and providing quantitative, rather than qualitative, assessments of radiographic characteristics. In this opinion article, we establish a general understanding of AI methods, particularly those pertaining to image-based tasks. We explore how these methods could impact multiple facets of radiology, with a general focus on applications in oncology, and demonstrate ways in which these methods are advancing the field. Finally, we discuss the challenges facing clinical implementation and provide our perspective on how the domain could be advanced.

Introduction

Artificial Intelligence **[G]** (AI) has recently made substantial strides in perception, the interpretation of sensory information, allowing machines to better represent and interpret complex data. This has led to major advances in applications ranging from web search and self-driving vehicles to natural language processing and computer vision - tasks that up until a few years ago could only be done by humans[1]. Deep learning is a subset of machine learning **[G]** that is based on a neural network structure loosely inspired by the human brain. Such structures learn discriminative features from data automatically, giving them the ability to approximate very complex nonlinear relationships **(Box 1)**. While most earlier AI methods have led to applications with sub-human performance, recent deep learning algorithms are able to match and even surpass humans in task-specific applications[2–5] **(Figure 1)**. This owes to recent advances in AI research, the massive amounts of digital data now available to train algorithms, as well as modern powerful computational hardware. Deep learning methods have been able to defeat humans in the strategy board game of Go, an achievement that was previously thought to be decades away given the highly complex game space and massive number of potential moves[6]. Following the trend towards a human-level general AI, researchers predict that AI will automate many tasks including translating languages, writing best-selling books, and performing surgery - all within the coming decades[7].

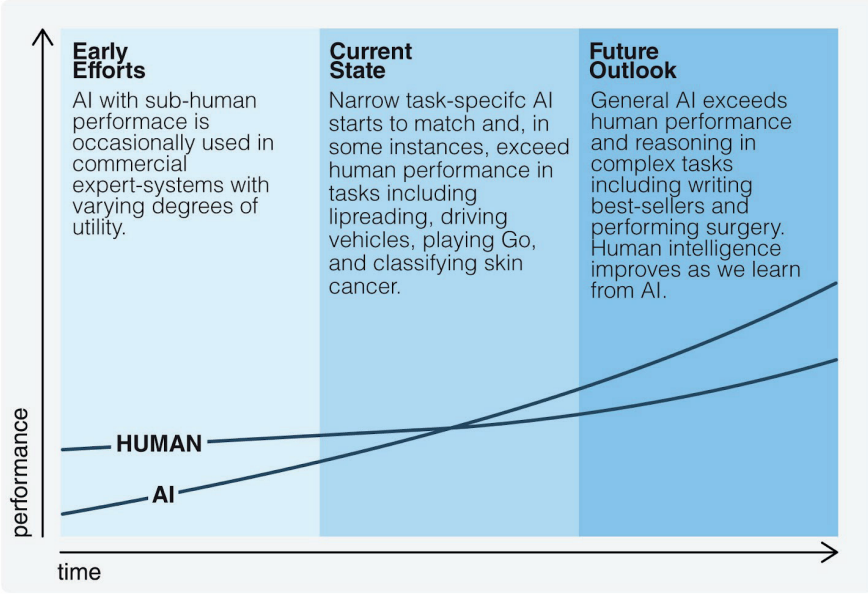

Figure 1. Artificial versus human intelligence. This plot outlines the performance levels of artificial intelligence (AI) and human intelligence starting from the early computer age and extrapolating into the future. Early AI came with a subhuman performance and varying degrees of success. Currently, we are witnessing narrow task-specific AI applications that are able to match and occasionally surpass human intelligence. It is expected that general AI will surpass human performance in specific applications within the coming years. Humans will potentially benefit from the human-AI interaction, bringing them to higher levels of intelligence.

Within healthcare, AI is becoming a major constituent of many applications including drug discovery, remote patient monitoring, medical diagnostics and imaging, risk management, wearables **[G]**, virtual assistants, and hospital management. Many domains with big data components such as the analysis of DNA and RNA sequencing data[8] are also expected to benefit from the use of AI. Medical fields that rely on imaging data, including radiology, pathology, dermatology[9], and ophthalmology[10], have already begun to benefit from the implementation of AI methods **(Box 2)**. Within radiology, trained physicians visually assess medical images and report findings to detect, characterize, and monitor diseases. Such assessment is often based on education and experience and can be, at times, subjective. In contrast to such qualitative reasoning, AI excels at recognizing complex patterns in imaging data and can provide a quantitative assessment in an automated fashion. More accurate and reproducible radiology assessments can then be made when AI is integrated into the clinical workflow as a tool to assist physicians.

Machine learning algorithms based on predefined engineered features

Traditional artificial intelligence (AI) methods rely, largely, on predefined engineered feature algorithms **(Figure 2A)** with explicit parameters based on expert knowledge. Such features are designed to quantify specific radiographic characteristics, such as the 3D shape of a tumor or the intratumoral texture and distribution of pixel intensities (histogram). A subsequent selection step ensures that only the most relevant features are used. Statistical machine learning models are then fit to this data to identify potential imaging-based biomarkers. Examples of these models include support vector machines and random forests.

Deep learning algorithms

Recent advances in AI research have given rise to new non-deterministic deep learning algorithms that do not require explicit feature definition, representing a fundamentally different paradigm in machine learning[11–13]. The underlying methods of deep learning have existed for decades. However, only in recent years, sufficient data and computational power have become available. Without explicit feature predefinition or selection, these algorithms learn directly by navigating the data space giving them superior problem-solving capabilities. While various deep learning architectures have been explored to address different tasks, convolutional neural networks (CNN) are the most prevalent deep learning architecture typologies in medical imaging today[14]. A typical CNN comprises a series of layers that successively map image inputs to desired endpoints, while learning increasingly higher level imaging features **(Figure 2B)**. Starting from an input image, 'hidden layers' within CNNs usually include a series of convolution and pooling operations extracting feature maps and performing feature aggregation respectively. These hidden layers are then followed by fully connected layers providing high level reasoning before an output layer produces predictions. CNNs are often trained end-to-end with labelled data for supervised learning. Other architectures, such as deep autoencoders[15] and generative adversarial networks[16], are more suited to unsupervised learning tasks on unlabeled data. Transfer learning, or using pre-trained networks on other datasets, is often utilized when dealing with scarce data[17].

Box 1: Artificial Intelligence methods in Medical Imaging

As imaging data are collected during routine clinical practice, large datasets are - in principle - readily available, thus offering an incredibly rich resource for scientific and medical discovery. Radiographic images, coupled with data on clinical outcomes, has led to the emergence and rapid expansion of radiomics **[G]** as a field of medical research[18–20]. Early radiomics studies were largely focused on mining images for a large set of predefined engineered features **[G]** that describe radiographic aspects of shape, intensity, and texture. More recently, radiomics studies have incorporated deep learning techniques to learn feature representations automatically from example images[14] hinting at the significant clinical relevance of many of these radiographic features. Within oncology, multiple efforts have successfully explored radiomics tools for assisting clinical decision making related to the diagnosis and risk stratification of different cancers[21,22]. For example, studies in non-small-cell lung cancer (NSCLC) used radiomics to predict distant metastasis in lung adenocarcinoma[23], tumor histological subtypes[24], as well as disease recurrence[25], somatic mutations[26], gene-expression profiles[27], and overall survival rates[28]. Such findings have motivated exploring the clinical utility of AI-generated biomarkers based on standard-of-care radiographic images[29] - with the ultimate hope of better supporting radiologists in disease diagnosis, imaging quality optimization, data visualization, response assessment, and report generation **[G]**.

In this opinion article, we start by establishing a general understanding of AI methods particularly pertaining to image-based tasks. We then explore how up-and-coming AI methods will impact multiple radiograph-based practices within oncology. Finally, we discuss the challenges and hurdles facing the clinical implementation of these methods.

AI in Medical Imaging

The primary driver behind the emergence of AI in medical imaging has been the desire for greater efficacy and efficiency in clinical care. Radiological imaging data continues to grow at a disproportionate rate when compared with the number of available trained readers, and the decline in imaging reimbursements has forced healthcare providers to compensate by increasing productivity[30]. These factors have contributed to a dramatic increase in radiologists' workload. Studies report that, in some cases, an average radiologist must interpret one image every 3-4 seconds in an 8-hour workday to meet workload demands[31]. As radiology involves visual perception as well as decision-making under uncertainty[32], errors are inevitable - especially under such constrained conditions.

Radiology-based

Thoracic Imaging. Lung cancer is one of the most common and deadly of tumors. Lung cancer screening can help identify pulmonary nodules, with early detection being lifesaving in many cases. AI can help to automatically identify these nodules and also assist in categorizing them as benign or malignant.

Abdominal and Pelvic Imaging. With the rapid growth in medical imaging, especially computed tomography (CT) and magnetic resonance imaging (MRI), more incidental findings including liver lesions are identified. AI may aid in characterizing these lesions as benign or malignant and prioritizing follow up evaluation for these patients.

Colonoscopy. Colonic polyps that are undetected or misclassified pose a potential risk for colorectal cancer. While most polyps are initially benign, they can become malignant over time[33]. Hence, early detection and consistent monitoring with robust AI-based tools is critical.

Mammography. Screening mammography is technically challenging to expertly interpret. AI can assist in the interpretation, in part by identifying and characterizing microcalcifications (small deposits of calcium in the breast).

Brain Imaging. Brain tumors are characterized by abnormal growth of tissue and can either be benign, malignant, primary or metastatic[34].

Radiation Oncology. Radiation treatment planning can be automated by segmenting tumors for radiation dose optimization. Furthermore, assessing response to treatment by monitoring over time is essential to evaluate the success of radiation therapy efforts. AI is able to perform these assessments, thereby improving accuracy and speed.

Non-Radiology-based

Dermatology. Diagnosing skin cancer requires trained dermatologists to visually inspect suspicious areas. With the large variability in sizes, shades and textures, skin lesions are rather challenging to interpret[9]. The massive learning capacity of deep learning algorithms qualifies them to handle such variance and detect characteristics well beyond those considered by humans.

Pathology. The quantification of digital whole slide images of biopsies is vital in the accurate diagnosis of many types of cancers. With the large variation in imaging hardware, slide preparation, magnification and staining techniques, traditional AI methods often require considerable tuning to address this problem. More robust AI is able to more accurately perform mitosis detection, segment histologic primitives (such as nuclei, tubules and epithelium), count events as well as characterize and classify tissue[35–38].

DNA and RNA sequencing. The ever increasing amounts of available sequencing data continues to provide opportunities for utilizing genomic endpoints in cancer diagnosis and care. AI-based tools are able to identify and extract high level features correlating somatic point mutations and cancer types[39] as well as predict the effect of mutations on sequence specificities of RNA- and DNA-binding proteins[40].

Box 2. Examples of Clinical Application Areas of Artificial Intelligence in Oncology

A seamlessly integrated AI component within the imaging workflow would increase efficiency, reduce errors, and achieve objectives with minimal manual input by providing trained radiologists with pre-screened images and identified features. Therefore, substantial efforts and policies are being put forward to facilitate technological advances related to AI in medical imaging. Almost all image-based radiology tasks are contingent upon the quantification and assessment of radiographic characteristics from images. These characteristics can be important for the clinical task at hand, that is, for the detection, characterization, or monitoring of diseases. The application of logic and statistical pattern recognition to problems in medicine have long been proposed since the early 1960s[41,42]. As computers became more prevalent in the 1980s, the AI-powered automation of many clinical tasks has shifted radiology from a perceptual subjective craft to a quantitatively computable domain[43,44]. The rate at which AI is evolving radiology is parallel to that in other application areas and is proportional to the rapid growth of data and computational power.

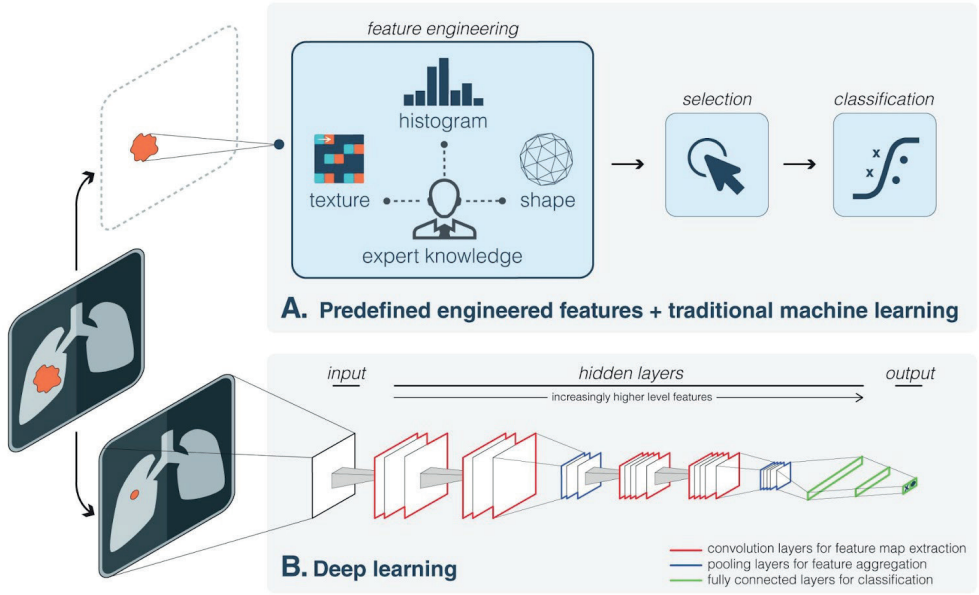

Figure 2. Artificial intelligence methods in medical imaging. This schematic outlines two artificial intelligence (AI) methods for a representative classification task, such as the diagnosis of a suspicious object as either benign or malignant. A. The first method relies on engineered features extracted from regions of interest on the basis of expert knowledge. Examples of these features in cancer characterization include tumour volume, shape, texture, intensity and location. The most robust features are selected and fed into machine learning classifiers. B. The second method uses deep learning and does not require region annotation — rather, localization is usually sufficient. It comprises several layers where feature extraction, selection and ultimate classification are performed simultaneously during training. As layers learn increasingly higher-level features **(Box 1)**, earlier layers might learn abstract shapes such as lines and shadows, while other deeper layers might learn entire organs or objects. Both methods fall under radiomics, the data-centric, radiology-based research field.

There are two classes of AI methods that are in wide use today **(Box 1**, **Figure 2)**. The first uses handcrafted engineered features that are defined in terms of mathematical equations (such as tumor texture) and can thus be quantified using computer programs[45]. These features are used as input to state-of-the-art machine learning models that are trained to classify patients in ways that can support clinical decision making. While such features are perceived to be discriminative, they rely on expert definition and hence do not necessarily represent the most optimal feature quantification approach for the discrimination task at hand. Moreover, predefined features are often unable to adapt to variations in imaging modalities, such as computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI)**,** and their associated signal to noise characteristics.

The second method, deep learning, has gained considerable attention in recent years. Deep learning algorithms can automatically learn feature representations from data without the need for prior definition by human experts. This data driven approach allows for more abstract feature definitions making it more informative and generalizable. Deep learning can thus automatically quantify phenotypic characteristics of human tissues[46], promising substantial improvements in diagnosis and clinical care. Deep learning has the added benefit of reducing the need for manual preprocessing steps. For example, to extract predefined features, accurate segmentation **[G]** of diseased tissues by experts is often needed[47]. Because deep learning is data driven **(Box 1)**, with enough example data it can automatically identify diseased tissues and hence avoid the need for expert defined segmentations. Given its ability to learn complex data representations, deep learning is also often robust against unwanted variation such as inter-reader variability, and can hence be applied to a large variety of clinical conditions and parameters. In many ways, deep learning can mirror what trained radiologists do, that is, identify image parameters, but also weigh up the importance of these parameters based on other factors to arrive at a clinical decision.

Given the growing number of applications of deep learning in medical imaging[14], several efforts have compared deep learning methods with their predefined feature-based counterparts, and have reported substantial performance improvements with deep learning[48,49]. Studies have also shown that deep learning technologies are on par with radiologists' performance for both detection[50] and segmentation[51] tasks in ultrasound and MRI respectively. For the classification tasks of lymph node metastasis in PET–CT, deep learning had higher sensitivities but lower specificities than radiologists[52]. As these methods are iteratively refined and tailored for specific applications, a better command of the sensitivity:specificity trade-off is expected. Deep learning can also enable faster development times as it depends solely on curated data and its corresponding metadata rather than domain expertise. On the other hand, traditional predefined feature systems have shown plateauing performance over recent years and hence do not generally meet the stringent requirements for clinical utility. As a result, only a few have been translated into the clinic[53]. It is expected that high performance deep learning methods will surpass the threshold for clinical utility within the near future, and can therefore be expeditiously translated into the clinic.

Impact on Oncology Imaging

In this section, we focus on three main clinical radiology tasks that specifically pertain to oncology: abnormality detection followed by characterization and subsequent monitoring of change **(Figure 3)**. These tasks require a diversified set of skills: medical, in terms of disease diagnosis and care, as well as technical for capturing and processing radiographic images. Both these skills hint at the ample opportunities where up and coming AI technologies can positively impact clinical outcomes by identifying phenotypic characteristics in images. In addition to performing these tasks on radiographic cancer images, such as in thoracic imaging and mammography, they are also common to other oncology subspecialties where non-radiographic images are used **(Box 2)**. For each of these tasks, we investigate technologies currently being utilized in the clinic and provide highlights of research efforts aimed at integrating state-of-the-art AI developments in these practices.

Detection

Within the manual detection workflow, radiologists rely on manual perceptive skills to identify possible abnormalities, followed by cognitive skills to either confirm or reject the findings. Radiologists visually scan through stacks of images while periodically adjusting viewing planes and window width and level settings. Relying on education, experience and an understanding of the healthy radiograph, radiologists are trained to identify abnormalities based on changes in imaging intensities or the appearance of unusual patterns. These criteria, and many more, fall within a somewhat subjective decision matrix that enables reasoning in problems ranging from detecting lung nodules to breast lesions and colon polyps. As dependence on computers has increased, automated methods for the identification and processing of these predefined features - collectively known as computer aided detection (CADe) - have long been proposed and occasionally utilized in the clinic[45]. Radiologist-defined criteria are distilled into a pattern recognition problem where computer vision algorithms highlight conspicuous objects within the image[54]. However, these algorithms are often task-specific and do not generalize across diseases and imaging modalities. Additionally, the accuracy of traditional predefined feature-based CADe systems remains questionable with ongoing efforts to reduce false positives. It is often the case that outputs have to be assessed by radiologists to decide if a certain automated annotation merits further investigation, thereby making it labor intensive. In examining mammograms, some studies have reported that radiologists rarely altered their diagnostic decisions after viewing results and that predefined feature-based CADe integration had no statistical significance on the radiologist's performance within a clinical setting[55,56]. This is owing, in part, to the sub-human performance of these systems. Recent efforts have explored deep learning-based CADe to detect pulmonary nodules in CT[57] and prostate cancer in multiparametric imaging **[G]**, specifically multiparametric MRI[58]. In detecting lesions in mammograms, early results show that utilizing convolutional neural networks **(Deep learning algorithms; Box 1)** in CADe outperforms traditional CADe systems at low sensitivity while performing comparably at high sensitivity, and shows similar performance compared to human

readers[59]. These findings hint at the utility of deep learning in developing robust high-performance CADe systems.

Characterization

Characterization is an umbrella term referring to the segmentation, diagnosis, and staging of a disease. These tasks are accomplished by quantifying radiological characteristics of an abnormality, such as the size, extent, as well as internal texture. While handling routine tasks of examining medical images, humans are simply not capable of accounting for more than a handful of qualitative features. This is exacerbated by the inevitable variability across human readers, with some performing better than others. Automation through AI can, in principle, consider a large number of quantitative features together with their degrees of relevance - while performing the task at hand in a reproducible manner every time. For instance, it is difficult for humans to accurately predict the status of malignancy in the lung due to the similarity between benign and malignant nodules in CT scans. AI can automatically identify these features, and many others, while treating them as imaging biomarkers. Such biomarkers could hence be used to predict malignancy likelihood amongst other clinical endpoints including risk assessment, differential diagnosis, prognosis, and response to treatment.

Within the initial segmentation step, whilst non-diseased organs can be segmented with relative ease, identifying the extent of diseased tissue is potentially orders of magnitude more challenging. Typical practices of tumor segmentation within clinical radiology today are often limited to high-level metrics such as the largest in-plane diameter. However, in other clinical cases, a higher specificity and precision are vital. For instance, in clinical radiation oncology, the extents of both tumor and non-tumor tissues have to be accurately segmented for radiation treatment planning. Attempts at automating segmentation have made their way into the clinic, with varying degrees of success[60]. Segmentation finds its roots in earlier computer vision research carried out in the 1980s[61] with continued refinement over the past decades. Simpler segmentation algorithms used clustered imaging intensities to isolate different areas, or utilized region growing where regions are expanded around user-defined seed points within objects until a certain homogeneity criterion is no longer met[62]. A second generation of algorithms saw the incorporation of statistical learning and optimization methods to improve segmentation precision, such as the watershed algorithm where images are transformed into topological maps with intensities representing heights[63]. More advanced systems incorporate prior knowledge into the solution space, as in the use of a probabilistic atlas **[G]** - often an attractive option when objects are ill-defined in terms of their pixel intensities. Such atlases have enabled more accurate automated segmentations as they contain information regarding the expected locations of tumors across entire patient populations[60]. Applications of probabilistic atlases include segmenting brain MRI for locating diffuse low-grade glioma[64], prostate MRI for volume estimation[65] and head and neck CT for radiotherapy treatment planning[66], to name a few.

Recently proposed deep learning architectures for segmentation include fully convolutional networks, networks comprised of convolutional layers only, that output segmentation probability maps across entire images[67]. Other architectures, such as the U-net[68], have been specifically designed for medical images. Studies have reported that a single deep learning system is able to perform diverse segmentation tasks across multiple modalities and tissue types: brain MRI, breast MRI and cardiac CT angiography (CTA), without task-specific training[69]. Others describe deep learning methods for brain MRI segmentation that completely eliminate the need for image registration, a required preprocessing step in atlas-based methods[70].

Multiple radiographic characteristics are also employed in subsequent diagnosis tasks. These are critical to identify, for instance, if a lung nodule is solid or if it contains non-solid areas, also known as ground-glass opacity **[G]** (GGO) nodules. GGOs are rather challenging to diagnose and often require special management protocols, mainly due to the lack of associated characteristics of malignancy or invasiveness in radiographs[71]. Generally, tumor radiographic characteristics may include information regarding size, maximum diameter, sphericity, internal texture, and margin definition. The logic for diagnosis is based on these, often subjective, characteristics enabling the stratification of objects into classes indicative of being benign or malignant. Methods to automate diagnoses are collectively referred to as computer aided diagnosis (CADx) systems. Similar to CADe, they often rely on predefined engineered discriminative features. Several systems are already in clinical use, as is the case with screening mammograms[72]. They usually serve as a second opinion in complementing a radiologist's assessment[73] and their perceived successes have led to the development of similar systems for other imaging modalities including ultrasound and MRI[74]. For instance, traditional CADx systems have been used on ultrasound images to diagnose cervical cancer in lymph nodes, where they have been found to improve the performance of particularly inexperienced radiologists as well as reduce variability amongst them[75]. Other application areas include prostate cancer in multiparametric MRI where a malignancy probability map is first calculated for the entire prostate, followed by automated segmentation for candidate detection[76].

The accuracy of traditional predefined feature-based CADx systems is contingent upon several factors, including the accuracy of prior object segmentations. It is often the case that errors are magnified as they propagate through the various image-based tasks within the clinical oncology workflow. We also find that some traditional CADx methods fail to generalize across different objects. For instance, while the measurement of growth rates over time is considered as a major factor in assessing risk, pulmonary nodule CADx systems designed around this criterion are often unable to accurately diagnose special nodules such as cavity and GGO nodules[77]. Such nodules require further descriptors for accurate detection and diagnosis - descriptors that are not discriminative when applied to the more common solid nodules[78]. This eventually leads to multiple solutions that are tailored for specific conditions with limited generalizability. Without explicit predefinition of these discriminative features, deep learning-based CADx is able to automatically learn from patient populations and form a general understanding of

variations in anatomy - thus allowing it to capture a representation of common and uncommon cases alike.

Architectures such as CNNs are well suited for supervised diagnostic classification tasks **(Figure 2B)**. For both the breast lesion and lung nodule classification tasks, studies report a substantial performance gain of deep learning-based CADx - specifically those utilizing stacked denoising auto-encoders - over its traditional state-of-the-art counterparts. This is mainly owing to the automatic feature exploration mechanism and higher noise tolerance of deep learning. Such performance gain is assessed using multiple metrics including the area under receiver operating characteristic curve **[G]** (AUC), accuracy, sensitivity and specificity to name a few[49].

Staging systems, such as tumor–node–metastasis (TNM) in oncology, rely on preceding information gathered through segmentation and diagnosis to classify patients into multiple predefined categories[79]. This enables a well-informed choice of the type of treatment and aids in predicting survival likelihood and prognosis. Staging has generally seen little to no automation since it relies on qualitative descriptions that are often difficult to quantitatively measure. The automated staging of primary tumor size (T), nearby lymph nodes (N), and distant metastasis (M) all require different feature sets and approaches. While traditional machine learning might have relied on ensemble methods where multiple distinct models are combined, deep learning has the ability to learn joint data representations simultaneously[80] - making it well suited for such multi-faceted classification problems. Most deep learning efforts to detect lymph node involvement and distant metastasis - and ultimately obtain an accurate staging - have been carried out on pathology images[81,82]. However, more work on radiographic images is expected to appear in the near future.

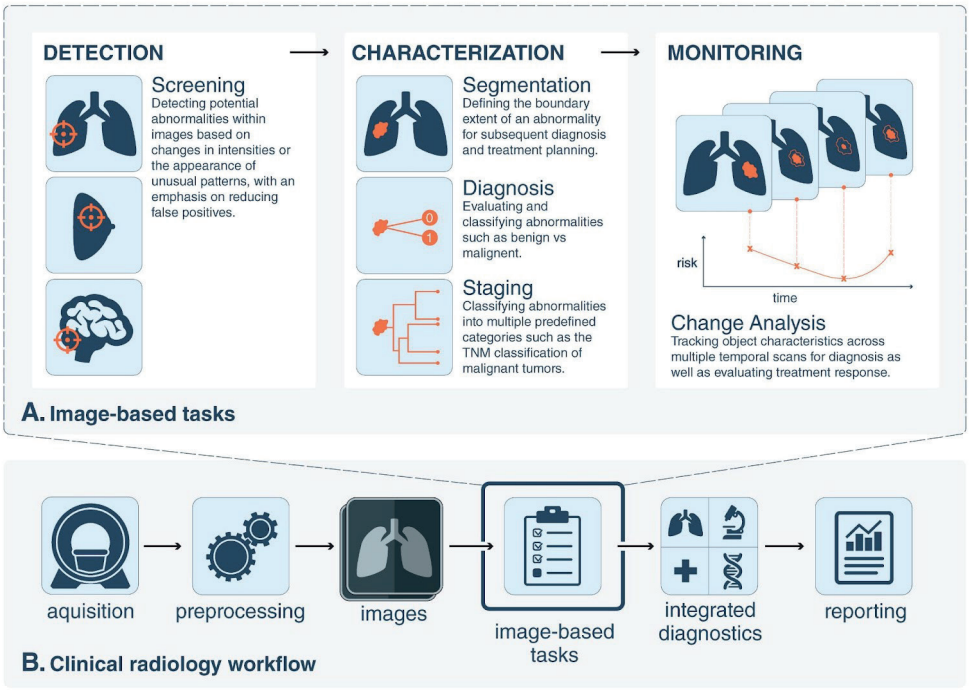

Figure 3. Artificial intelligence impact areas within oncology imaging. This schematic outlines the various tasks within radiology where artificial intelligence (AI) implementation is likely to have a large impact. A. The workflow comprises the following steps: preprocessing of images after acquisition, image-based clinical tasks (which usually involve the quantification of features either using engineered features with traditional machine learning or deep learning), reporting results through the generation of textual radiology reports and, finally, the integration of patient information from multiple data sources. B. AI is expected to impact image-based clinical tasks, including the detection of abnormalities; the characterization of objects in images using segmentation, diagnosis and staging; and the monitoring of objects for diagnosis and assessment of treatment response. TNM, tumour–node–metastasis.

Monitoring

Disease monitoring is essential for diagnosis as well as evaluating treatment response. The workflow involves an image registration preprocess where the diseased tissue is aligned across multiple scans, followed by evaluating simple metrics on them using predefined protocols - very similar to diagnosis tasks on single time-point images. A simple data comparison protocol follows and is used to quantify change. In oncology, for instance, examples of these protocols include Response Evaluation Criteria in Solid Tumors (RECIST) and World Health Organization (WHO)[83] and define information regarding tumor size. Here, we find that the main goal behind such simplification is reducing the amount of effort and data a human reader must interact with while performing the task. However, it is often based on incorrect assumptions regarding isotropic tumor growth. While some change characteristics are directly identifiable by

humans such as relatively large variations in object size, shape, and cavitation, others are not. These could include subtle variations in texture and heterogeneity within the object. Poor image registration, dealing with multiple objects as well as physiological changes over time all contribute to more challenging change analyses. Moreover, the inevitable interobserver variability[84] remains to be a major weakness in the process. Computer aided change analysis is considered a relatively younger field compared with CADe and CADx systems, one that has not yet achieved as much of a widespread adoption[85]. Early efforts in automating change analysis workflows relied on the automated registration of multiple images followed by subtracting them from one another, where changed pixels are highlighted and presented to the reader. Other more sophisticated methods perform a pixel-by-pixel classification - based on predefined discriminative features - to identify changed regions and hence produce a more concise map of change[86]. As the predefined features used for registration differ from those used for the subsequent change analysis, a multi-step procedure is required combining different feature sets. This could compromise the change analysis step as it becomes highly sensitive to registration errors. With computer aided change analysis based on deep learning, feature engineering is eliminated and a joint data representation can be learned. Deep learning architectures, such as recurrent neural networks, are very well suited for such temporal sequence data formats and are expected to find ample applications in monitoring tasks.

Other Opportunities

In addition to the three primary clinical tasks mentioned above, AI is expected to impact other image-based tasks within the clinical radiology workflow. These include the preprocessing steps following image acquisition as well as subsequent reporting and integrated diagnostics **(Figure 3A)**.

Starting at the outset of the workflow, the first of these tasks to be improved is reconstruction. We find a widening gap between advancements in image acquisition hardware and image reconstruction software, a gap potentially addressed by new deep learning methods for suppressing artifacts and improving overall quality. For instance, CT reconstruction algorithms have seen little to no change in the past 25 years[87]. Additionally, many filtered-back projection image reconstruction algorithms are computationally expensive, signifying that a tradeoff between distortions and runtime is inevitable[88]. Recent efforts report deep learning's flexibility in learning reconstruction transformations for various MRI acquisition strategies, by treating the reconstruction process as a supervised learning **[G]** task where a mapping between the scanner sensors and resultant images is derived[89]. Other efforts employ novel AI methods to correct for artifacts as well as address certain imaging modality-specific problems such as the limited angle problem in CT[90] - a missing data problem where only a portion of the scanned space can be reconstructed, due to the scanner's inability to perform full 180° rotations around objects. Studies have also utilized CNNs and synthetically generated artifacts to combine information from original and corrected images as a means to suppress metal artifacts[91]. More work is needed to investigate the accuracy of deep learning-based reconstruction algorithms and their ability to recreate rare unseen structures, as initial

errors propagated throughout the radiology workflow can have adverse effects on patient outcome.

Another preprocessing task to be enhanced is registration, as touched upon previously in the monitoring section. This process is often based on predefined similarity criteria such as landmark and edge-based measures. In addition to the computational power and time consumed by these predefined feature-based methods, some are sensitive to initializations **[G]** , chosen similarity features, and the reference image[92]. Deep learning methods could handle complex tissue deformations through more advanced non-rigid registration algorithms whilst providing better motion compensation for temporal image sequences. Studies have shown that deep learning leads to generally more consistent registrations and is an order of magnitude faster compared with more conventional methods[93]. Additionally, deep learning is multi-modal in nature where a single shared representation between imaging modalities can be learned[94]. Multimodal images in cancer have enabled the association of multiple quantitative functional measurements as in the PET hybrids: PET–MRI and PET–CT, thus improving the accuracy of tumor characterization and assessment[95]. With robust registration algorithms based on deep learning, the utility of multimodal imaging can be further explored without concerns regarding registration accuracy.

Radiology reports lie at the intersection of radiology and multiple oncology subspecialties. However, the generation of these textual reports can be a laborious and routine time-consuming task. When compared to conventional dictation, even structured reporting systems with bulleted formatting have been shown not to improve attending physicians' perception of report clarity[96]. As the report generation task falls towards the end of the radiology workflow, it is the most sensitive to errors from preceding steps. Additionally, the current radiologist–oncologist communication model has not been found to be optimally coordinated - especially with regards to monitoring lesions over time[97]. Due to the often different formats in which data is recorded by medical professionals, AI-run automatic report generation tools can pave the way for a more standardized terminology - an area that currently lacks stringent standards as well as an agreed upon understanding of what constitutes a 'good' report[98]. Such tools could also replace the traditional qualitative text-based approach with a more interactive quantitative one, which has been shown to enhance and promote collaboration between different parties[99]. Within lung cancer screening, this could include quantified information about the size and location of a nodule, probability of malignancy and associated confidence level. These well-structured reports are also immensely beneficial to population sciences and big data mining efforts. Following deep learning advances in the automatic caption generation **[G]** from photographic images[100], recent efforts have explored means to diagnose abnormalities in chest x-rays and automatically annotate it in a textual format[101]**.**

After carrying out various clinical tasks and generating radiology reports **(Figure 3A)**, AI-based integrated diagnostics could potentially enable healthcare-wide assimilation of data from multiple streams, thus capitalizing on all data types pertaining to a particular patient. In addition to radiology reports describing findings from medical images and their associated metadata, other data could be sourced from the clinic or from pathology

or genomics testing. Data from wearables, social media, and other lifestyle quantifying sources could all potentially offer valid contributions to such a comprehensive analysis. This will be crucial in providing AI biomarkers with robust generalizability towards different endpoints. Such consolidation of standard medical data, using traditional AI methods, has already demonstrated the ability to advance clinical decision-making in lung cancer diagnosis and care[27].

AI Challenges in Medical Imaging

We are currently witnessing a major paradigm shift in the design principles of many computer-based tools used in the clinic. There is great debate about the speed with which newer deep learning methods will be implemented in clinical radiology practice[102], with speculations for the time needed to fully automate clinical tasks ranging from a few years to decades. The development of deep learning-based automated solutions will begin with tackling the most common clinical problems where sufficient data is available. These problems could involve cases where human expertise is in high demand or data is far too complex for human readers - examples of these include the reading of lung screening CTs, mammograms, and images from virtual colonoscopy. A second wave of efforts is likely to address more complex problems such as multiparametric MRI imaging. A common trait amongst current AI tools is their inability to address more than one task, as is the case with any narrow intelligence. A comprehensive AI system able to detect multiple abnormalities within the entire human body is yet to be developed.

Data continues to be the most central and crucial constituent for learning AI systems. With one out of four Americans receiving a CT examination[103] and one out of ten receiving an MRI examination[104] annually, millions of medical images are produced each year. Additionally, recent well-implemented advances in US-based digital health systems - such as the Picture Archiving and Communication System (PACS) - have ensured medical images are electronically organized in a systematic manner[105,106], with parallel efforts in Europe[107] and developing countries[108]. It is clear that large amounts of medical data are indeed available, and are stored in such a manner that enables relative ease in access and retrieval. However, such data is rarely curated and this represents a major bottleneck in attempting to learn any AI model. Curation can refer to patient cohort selection relevant for a specific AI task, but can also refer to segmenting objects within images. Curation ensures that training data adheres to a defined set of quality criteria and is clear of compromising artifacts. It can also help avoid unwanted variance in data due to differences in data acquisition standards and imaging protocols, especially across institutions, such as the time between contrast agent administration and actual imaging. An example of data curation within oncology could include assembling a cohort of patients with specific stages of disease and tumor histology grades. While photographic images can be labelled by non-experts, using for instance crowdsourcing approaches, medical images do require domain knowledge. Hence, it is imperative that such curation is performed by a trained reader to ensure credibility - making the process expensive. It is also very time consuming, although utilizing recent deep learning algorithms promises to reduce annotation time substantially: meticulous slice-by-slice

segmentation can potentially be substituted by single seed points within the object and from which full segmentations could be automatically generated. The amount of data which needs to be curated is another limiting factor and is highly dependent on the AI approach - with deep learning methods being more prone to overfitting and hence often require more data.

The suboptimal performance of many automated and semi-automated segmentation algorithms[60] has hindered their utility in curating data as human readers are almost always needed to verify accuracy. More complications arise with rare diseases where automated labelling algorithms are non-existent. The situation is exacerbated when only a limited number of human readers have prior exposure and are capable of verifying these uncommon diseases. One solution that enables automated data curation is unsupervised learning **[G]**. Recent advances in unsupervised learning, including generative adversarial networks[16] and variational autoencoders[15] amongst others, show great promise as discriminative features are learned without explicit labelling. Recent studies have explored unsupervised domain adaptation using adversarial networks to segment brain MRI, leading to a generalizability and accuracy close to that of supervised learning methods[109]. Others employ sparse autoencoders to segment breast density and score mammographic texture in an unsupervised manner[110]. Self-supervised learning **[G]** efforts have also utilized spatial context information as supervision for recognizing body parts in CT and MRI volumes through the use of paired CNNs[111]. Nevertheless, public repositories such as The Cancer Imaging Archive (TCIA)[112] offer unparalleled open-access to labelled medical imaging data allowing immediate AI model prototyping, and thus eliminating lengthy data curation steps.

Albeit intuitively leading to higher states of intelligence, the recent paradigm shift from programs based on well-defined rules to others that learn directly from data has brought certain unforeseen concerns to the spotlight. A strong theoretical understanding of deep learning is yet to be established[113], in spite of the reported successes across many fields - explaining why deep learning layers that lie between inputs and outputs are labelled as 'hidden layers' **(Box 1, Figure 2B)**. Identifying specific features of an image that contribute to a predicted outcome is highly hypothetical causing a lack of understanding of how certain conclusions are drawn by deep learning. This lack of transparency makes it difficult to predict failures, isolate the logic for a specific conclusion, or troubleshoot inabilities to generalize to different imaging hardware, scanning protocols and patient populations. Not surprisingly, many uninterpretable AI systems with applications in radiology have been dubbed 'black-box medicine'[114].

From a regulatory perspective, discussions are underway regarding the legal right of regulatory entities to interrogate AI frameworks on the mathematical reasoning for an outcome[115]. While such questioning is possible with explicitly programmed mathematical models, new AI methods such as deep learning have opaque inner workings as mentioned above. Sifting through hundreds of thousands of nodes in a neural network, and their respective associated connections, to make sense of their stimulation sequence is unattainable. An increased network depth and node count brings more complex decision-making together with a much more challenging system to take apart

and explore. On the other hand, we find that many safe and effective US Food and Drug Administration (FDA)-approved drugs have unknown mechanisms of action[116,117]. From that perspective and despite the degree of uncertainty surrounding many AI algorithms, the FDA has already approved high-performance software solutions albeit having somewhat obscure working mechanisms. Regulatory bodies, such as the FDA, have been regulating CADe and CADx systems that rely on machine learning and pattern recognition techniques since the earliest days of computing. However, it is the shift to deep learning that now poses new regulatory challenges and requires new guidance for submissions seeking approval. Even after going to market, deep learning methods evolve over time as more data is processed and learned from. Thus, it is crucial to understand the implications of such lifelong learning in these adaptive systems. Periodic testing over specific time intervals could potentially ensure that learning and its associated prediction performance are following forecasted projections. Additionally, such benchmarking tests need to adapt to AI specifics such as the sensitivity of prediction probabilities in CNNs.

Other ethical issues may arise from the use of patient data to train these AI systems. Data is hosted within networks of medical institutions, often lacking secure connections to state-of-the-art AI systems hosted elsewhere. More recently, Health Insurance Portability and Accountability Act **[G]** (HIPAA)-compliant storage systems have paved the way for more stringent privacy preservation. Studies have explored systems that enable multiple entities to jointly train AI models without sharing their input datasets - only sharing the trained model[118,119]. Other efforts use a decentralized 'federated' learning approach[120]. During training, data remains local while a shared model is learnt by combining local updates. Inference is then performed locally on live copies of the shared model, eliminating data sharing and privacy concerns. 'Cryptonets' are deep learning networks trained on encrypted data, even making encrypted predictions that can only be decrypted by the owner of a decryption key - thus ensuring complete confidentiality throughout the entire process[121]. All these solutions, albeit still in early developmental stages, promise to create a sustainable 'data to AI' ecosystem - without undermining privacy and HIPAA compliance.

Future Perspectives

From the early days of X-ray imaging in the 1890s to more recent advances in CT, MR and PET scanning, medical imaging continues to be a pillar of medical treatment. Current advances in imaging hardware - in terms of quality, sensitivity and resolution - enable the discrimination of minute differences in tissue densities. Such differences are, in some cases, difficult to recognize by a trained eye and even by some traditional AI methods used in the clinic. These methods are thus not fully on par with the sophistication of imaging instruments, yet another motivation to pursue this paradigm shift towards more powerful AI tools. Moreover, and in contrast to traditional methods based on predefined features, we find that deep learning algorithms scale with data, that is, as more data is generated every day and with ongoing research efforts, we expect to see relative improvements in performance. All these advances promise an increased accuracy and reduction in the number of routine tasks that exhaust time and effort.

Aligning research methodologies is crucial in accurately assessing the impact of AI on patient outcome. In addition to the undeniable importance of reproducibility and generalizability, utilizing agreed-upon benchmarking datasets, performance metrics, standard imaging protocols and reporting formats will level the experimentation field and enable unbiased indicators. It is also important to note that AI is unlike human intelligence in many ways; excelling in one task does not necessarily imply excellence in others. Therefore the promise of up and coming AI methods should not be overstated. Almost all state-of-the-art advances in the field of AI fall under the narrow AI category, where AI is trained for one task, and one task only - with only a handful exceeding human intelligence. While such advances excel in interpreting sensory perceptual information in a bottom-up fashion, they lack higher level, top-down knowledge of contexts as well as fail to make associations the way a human brain does. Thus, it is evident that the field is still in its infancy and overhyped excitement surrounding it should be replaced with rational thinking and mindful planning. It is also evident that AI is unlikely to replace radiologists within the near or even distant future. The roles of radiologists will expand as they become more connected to technology and have access to better tools. They are also likely to emerge as critical elements in the AI training process, contributing knowledge and overseeing efficacy. As different forms of AI exceed human performance, we expect it to evolve into a valuable educational resource. Human operators will not only be overseeing outcomes, but will also seek to interpret the reasoning behind them - as a means of validation as well as potentially discovering hidden information that might have been overlooked **(Figure 1)**.

In contrast to traditional AI algorithms locked within proprietary commercial packages, we find that the most popular deep learning software platforms available today are open-source. This has, and continues to, foster experimentation on a massive scale. In terms of data, AI efforts are expected to shift from processed medical images to raw acquisition data. Raw data is almost always downsampled and optimized for human viewers. This simplification and loss of information are both avoidable when the analyses are run by machines, but are associated with caveats including reduced interpretability and impeded human validation. As more data is generated, more signal is available for training. However, more noise is also present. We expect the process of discerning signal from noise to become more challenging over time. With difficulties in curating and labelling data, we foresee a major push towards unsupervised learning techniques to fully utilize the vast archives of unlabeled data.

Open questions include the ambiguity of who controls AI and is ultimately responsible for its actions, the nature of the interface between AI and healthcare and whether implementation of a regulatory policy too soon will cripple AI application efforts. Enabling interoperability amongst the multitude of AI applications that are currently scattered across healthcare will result in a network of powerful tools. This AI web will not only function at the inference level, but also at the life-long training level. We join the many calls[122] that advocate for creating an interconnected network of deidentified patient data from across the world. Using such data to train AI on a massive scale will enable a robust AI that is generalizable across different patient demographics, geographic

regions, diseases, and standards of care. Only then will we see a socially responsible AI benefiting the many and not the few.

Glossary

Area under receiver operating characteristic curve	(AUC). A sensitivity versus specificity metric for measuring the performance of binary classifiers that can be extended to multi-class problems. The area under the curve is equal to the probability that a randomly chosen positive sample ranks above a randomly chosen negative one or is regarded to have a higher probability of being positive.
Artificial intelligence	(AI). A branch of computer science involved with the development of machines that are able to perform cognitive tasks that would normally require human intelligence.
Caption generation	The often automated generation of qualitative text describing an illustration or image and its contents.
Ground-glass opacity	(GGO). A visual feature of some subsolid pulmonary nodules that is characterized by focal areas of slightly increased attenuation on computed tomography. Underlying bronchial structures and vessels are often visually preserved (being even more recognizable owing to increased contrast), thus making the detection and diagnosis of such nodules somewhat challenging.
Health Insurance Portability and Accountability Act	(HIPAA). A US act that sets provisions for protecting and securing sensitive patient medical data.
Image registration	A process that involves aligning medical images either in terms of spatial or temporal characteristics, mostly intramodality and occasionally intermodality.
Imaging modalities	A multitude of imaging methods that are used to non-invasively generate visualizations of the human anatomy. Examples of these include computed tomography (CT), computed tomography angiography (CTA), magnetic resonance imaging (MRI), mammography, ultrasonography (echocardiography) and positron emission tomography (PET).
Initializations	Within optimization problems, constantly adjusted parameters during run time need to be initialized to some value before the start of the process. Good initialization techniques aid models in converging faster and hence speed up the iteration process.
Machine learning	A branch of artificial intelligence and computer science that enables computers to learn without being explicitly programmed.

Multiparametric imaging	Medical imaging in which two or more parameters are used to visualize differences between healthy and diseased tissue. In multiparametric magnetic resonance imaging (MRI), these parameters include T2-weighted MRI, diffusion-weighted MRI and dynamic contrast enhanced MRI, among others.
Predefined engineered features	A set of context-based human-crafted features designed to represent knowledge regarding a specific data space.
Probabilistic atlas	A single composite image formed by combining and registering pre-segmented images of multiple patients that thus contains knowledge on population variability.
Radiomics	A data-centric field investigating the clinical relevance of radiographic tissue characteristics automatically quantified from medical images.
Report generation	The communication of assessments and findings in both image and text formats among medical professionals.
Segmentation	The partitioning of images to produce boundary delineations of objects of interest. Such a boundary is defined by pixels and voxels (3D pixels) when performed in 2D and 3D, respectively.
Self-supervised learning	A type of supervised learning where labels are determined by the input data as opposed to being explicitly provided.
Supervised learning	A type of machine learning where functions are inferred from labelled training data. Example data pairs consist of the input together with its desired output or label.
Unsupervised learning	A type of machine learning where functions are inferred from training data without corresponding labels.
Wearables	A collective term describing health-monitoring devices, smartwatches and fitness trackers that have recently been integrated into the health-care ecosystem as a means to remotely track vitals and adhere to treatment plans.

Acknowledgements

The authors acknowledge financial support from the US National Institutes of Health (NIH-USA U24CA194354 and NIH-USA U01CA190234).

References

1. Editors, N. Auspicious machine learning. *Nature Biomedical Engineering* **1**, 0036 (2017).
2. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
3. Moravčík, M. *et al.* DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* **356**, 508–513 (2017).
4. Xiong, W. *et al.* Toward Human Parity in Conversational Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **25**, 2410–2423 (2017).
5. Pendleton, S. D. *et al.* Perception, Planning, Control, and Coordination for Autonomous Vehicles. *Machines* **5**, 6 (2017).
6. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
7. Grace, K., Salvatier, J., Dafoe, A., Zhang, B. & Evans, O. When Will AI Exceed Human Performance? Evidence from AI Experts. http://arxiv.org/abs/1705.08807 (2017).
8. Rusk, N. Deep learning. *Nat. Methods* **13**, 35–35 (2015).
9. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
10. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* (2016) doi:10.1001/jama.2016.17216.
11. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
12. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* (2017) doi:10.1093/bib/bbx044.
13. Kevin Zhou, S., Greenspan, H. & Shen, D. *Deep Learning for Medical Image Analysis*. (Academic Press, 2017).
14. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
15. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv [stat.ML]* (2013).
16. Goodfellow, I. *et al.* Generative Adversarial Nets. in *Advances in Neural Information Processing Systems 27* (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q.) 2672–2680 (Curran Associates, Inc., 2014).
17. Shin, H.-C. *et al.* Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).
18. Aerts, H. J. W. L. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. *JAMA Oncol* (2016) doi:10.1001/jamaoncol.2016.2631.
19. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn. Reson. Imaging* **30**, 1234–1248 (2012).
20. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446 (2012).

21. Kolossváry, M., Kellermayer, M., Merkely, B. & Maurovich-Horvat, P. Cardiac Computed Tomography Radiomics: A Comprehensive Review on Radiomic Techniques. *J. Thorac. Imaging* (2017) doi:10.1097/RTI.0000000000000268.

22. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).

23. Coroller, T. P. *et al.* CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother. Oncol.* **114**, 345–350 (2015).

24. Wu, W. *et al.* Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology. *Front. Oncol.* **6**, 71 (2016).

25. Huynh, E. *et al.* Associations of Radiomic Data Extracted from Static and Respiratory-Gated CT Scans with Disease Recurrence in Lung Cancer Patients Treated with SBRT. *PLoS One* **12**, e0169172 (2017).

26. Rios Velazquez, E. *et al.* Somatic mutations drive distinct imaging phenotypes in lung cancer. *Cancer Res.* (2017) doi:10.1158/0008-5472.CAN-17-0122.

27. Grossmann, P. *et al.* Defining the biological basis of radiomic phenotypes in lung cancer. *Elife* **6**, (2017).

28. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci. Rep.* **5**, 13087 (2015).

29. O'Connor, J. P. B. *et al.* Imaging biomarker roadmap for cancer studies. *Nat. Rev. Clin. Oncol.* **14**, 169–186 (2017).

30. Boland, G. W. L., Guimaraes, A. S. & Mueller, P. R. The radiologist's conundrum: benefits and costs of increasing CT capacity and utilization. *Eur. Radiol.* **19**, 9–11; discussion 12 (2009).

31. McDonald, R. J. *et al.* The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad. Radiol.* **22**, 1191–1198 (2015).

32. Fitzgerald, R. Error in radiology. *Clin. Radiol.* **56**, 938–946 (2001).

33. Shin, Y. & Balasingham, I. Comparison of hand-craft feature based SVM and CNN based deep learning framework for automatic polyp classification. in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 3277–3280 (2017).

34. Orringer, D. A. *et al.* Rapid intraoperative histology of unprocessed surgical specimens via fibre-laser-based stimulated Raman scattering microscopy. *Nat Biomed Eng* **1**, (2017).

35. Albarqouni, S. *et al.* AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE Trans. Med. Imaging* **35**, 1313–1321 (2016).

36. Djuric, U., Zadeh, G., Aldape, K. & Diamandis, P. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *npj Precision Oncology* **1**, 22 (2017).

37. Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Inform.* **7**, 29 (2016).

38. Bejnordi, B. E. *et al.* Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* 929–932 (2017).

39. Yuan, Y. *et al.* DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. *BMC Bioinformatics* **17**, 476 (2016).

40. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).

41. Ledley, R. S. & Lusted, L. B. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science* **130**, 9–21 (1959).

42. Lodwick, G. S., Keats, T. E. & Dorst, J. P. THE CODING OF ROENTGEN IMAGES FOR COMPUTER ANALYSIS AS APPLIED TO LUNG CANCER. *Radiology* **81**, 185–200 (1963).

43. Ambinder, E. P. A history of the shift toward full computerization of medicine. *J. Oncol. Pract.* **1**, 54–56 (2005).

44. Haug, P. J. Uses of diagnostic expert systems in clinical care. *Proc. Annu. Symp. Comput. Appl. Med. Care* 379–383 (1993).

45. Castellino, R. A. Computer aided detection (CAD): an overview. *Cancer Imaging* **5**, 17–19 (2005).

46. Shen, D., Wu, G. & Suk, H.-I. Deep Learning in Medical Image Analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).

47. Veeraraghavan, H. MO-A-207B-01: Radiomics: Segmentation & Feature Extraction Techniques. *Med. Phys.* **43**, 3694–3694 (2016).

48. Paul, R. *et al.* Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma. *Tomography* **2**, 388–395 (2016).

49. Cheng, J.-Z. *et al.* Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Sci. Rep.* **6**, 24454 (2016).

50. Chen, H., Zheng, Y., Park, J.-H., Heng, P.-A. & Kevin Zhou, S. Iterative Multi-domain Regularized Deep Learning for Anatomical Structure Detection and Segmentation from Ultrasound Images. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* 487–495 (Springer, Cham, 2016).

51. Ghafoorian, M. *et al.* Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities. *Sci. Rep.* **7**, 5110 (2017).

52. Wang, H. *et al.* Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images. *EJNMMI Res.* **7**, 11 (2017).

53. van Ginneken, B., Schaefer-Prokop, C. M. & Prokop, M. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology* **261**, 719–732 (2011).

54. Nagaraj, S., Rao, G. N. & Koteswararao, K. The role of pattern recognition in computer-aided diagnosis and computer-aided detection in medical imaging: a clinical validation. *Int. J. Comput. Appl.* **8**, 18–22 (2010).

55. Cole, E. B. *et al.* Impact of computer-aided detection systems on radiologist accuracy with digital mammography. *AJR Am. J. Roentgenol.* **203**, 909–916 (2014).

56. Lehman, C. D. *et al.* Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Intern. Med.* **175**, 1828–1837 (2015).

57. Huang, X., Shan, J. & Vaidya, V. Lung nodule detection in CT using 3D convolutional neural networks. in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* 379–383 (2017).

58. Tsehay, Y. K. *et al.* Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images. in *Medical Imaging 2017: Computer-Aided Diagnosis* vol. 10134 1013405 (International Society for Optics and Photonics, 2017).

59. Kooi, T. *et al.* Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303–312 (2017).

60. Sharma, N. & Aggarwal, L. M. Automated medical image segmentation techniques. *J. Med. Phys.* **35**, 3–14 (2010).

61. Haralick, R. M. & Shapiro, L. G. Image segmentation techniques. *Computer Vision, Graphics, and Image Processing* **29**, 100–132 (1985).

62. Pham, D. L., Xu, C. & Prince, J. L. Current methods in medical image segmentation. *Annu. Rev. Biomed. Eng.* **2**, 315–337 (2000).

63. Grau, V., Mewes, A. U. J., Alcañiz, M., Kikinis, R. & Warfield, S. K. Improved watershed transform for medical image segmentation using prior information. *IEEE Trans. Med. Imaging* **23**, 447–458 (2004).

64. Parisot, S. *et al.* A Probabilistic Atlas of Diffuse WHO Grade II Glioma Locations in the Brain. *PLoS One* **11**, e0144200 (2016).

65. Ghose, S. *et al.* A coupled schema of probabilistic atlas and statistical shape and appearance model for 3D prostate segmentation in MR images. in *2012 19th IEEE International Conference on Image Processing* 541–544 (2012).

66. Han, X. *et al.* Atlas-based auto-segmentation of head and neck CT images. *Med. Image Comput. Comput. Assist. Interv.* **11**, 434–441 (2008).

67. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3431–3440 (2015).

68. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241 (Springer, Cham, 2015).

69. Moeskops, P. *et al.* Deep Learning for Multi-task Medical Image Segmentation in Multiple Modalities. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* 478–486 (Springer, Cham, 2016).

70. de Brebisson, A. & Montana, G. Deep neural networks for anatomical brain segmentation. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* 20–28 (2015).

71. Cioffi, U., Raveglia, F., De Simone, M. & Baisi, A. Ground-glass opacities: A curable disease but a big challenge for surgeons. *J. Thorac. Cardiovasc. Surg.* **154**, 375–376 (2017).

72. Champaign, J. L. & Cederbom, G. J. Advances in breast cancer detection with screening mammography. *Ochsner J.* **2**, 33–35 (2000).

73. Shiraishi, J., Li, Q., Appelbaum, D. & Doi, K. Computer-aided diagnosis and artificial intelligence in clinical imaging. *Semin. Nucl. Med.* **41**, 449–462 (2011).

74. Ayer, T., Ayvaci, M. U., Liu, Z. X., Alagoz, O. & Burnside, E. S. Computer-aided diagnostic models in breast cancer screening. *Imaging Med.* **2**, 313–323 (2010).

75. Zhang, J., Wang, Y., Yu, B., Shi, X. & Zhang, Y. Application of Computer-Aided Diagnosis to the Sonographic Evaluation of Cervical Lymph Nodes. *Ultrason. Imaging* **38**, 159–171 (2016).

76. Giannini, V. *et al.* A fully automatic computer aided diagnosis system for peripheral zone prostate cancer detection using multi-parametric magnetic resonance imaging. *Comput. Med. Imaging Graph.* **46 Pt 2**, 219–226 (2015).

77. El-Baz, A. *et al.* Computer-aided diagnosis systems for lung cancer: challenges and methodologies. *Int. J. Biomed. Imaging* **2013**, 942353 (2013).

78. Edey, A. J. & Hansell, D. M. Incidentally detected small pulmonary nodules on CT. *Clin. Radiol.* **64**, 872–884 (2009).

79. Mirsadraee, S., Oswal, D., Alizadeh, Y., Caulo, A. & van Beek, E., Jr. The 7th lung cancer TNM classification and staging system: Review of the changes and implications. *World J. Radiol.* **4**, 128–134 (2012).

80. Sohn, K., Shang, W. & Lee, H. Improved Multimodal Deep Learning with Variation of Information. in *Advances in Neural Information Processing Systems 27* (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q.) 2141–2149 (Curran Associates, Inc., 2014).

81. Litjens, G. *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **6**, 26286 (2016).

82. Cruz-Roa, A. *et al.* Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci. Rep.* **7**, 46450 (2017).

83. Jaffe, C. C. Measures of response: RECIST, WHO, and new alternatives. *J. Clin. Oncol.* **24**, 3245–3251 (2006).

84. Thiesse, P. *et al.* Response rate accuracy in oncology trials: reasons for interobserver variability. Groupe Français d'Immunothérapie of the Fédération Nationale des Centres de Lutte Contre le Cancer. *J. Clin. Oncol.* **15**, 3507–3514 (1997).

85. Khorasani, R., Erickson, B. J. & Patriarche, J. New opportunities in computer-aided diagnosis: change detection and characterization. *J. Am. Coll. Radiol.* **3**, 468–469 (2006).

86. Patriarche, J. W. & Erickson, B. J. Part 1. Automated change detection and characterization in serial MR studies of brain-tumor patients. *J. Digit. Imaging* **20**, 203–222 (2007).

87. Pan, X., Sidky, E. Y. & Vannier, M. Why do commercial CT scanners still employ traditional, filtered back-projection for image reconstruction? *Inverse Probl.* **25**, 1230009 (2009).

88. Pipatsrisawat, T., Gacic, A., Franchetti, F., Puschel, M. & Moura, J. M. F. Performance analysis of the filtered backprojection image reconstruction algorithms. in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* vol. 5 v/153–v/156 Vol. 5 (2005).

89. Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R. & Rosen, M. S. Image reconstruction by domain-transform manifold learning. *Nature* **555**, 487–492 (2018).

90. Hammernik, K., Würfl, T., Pock, T. & Maier, A. A Deep Learning Architecture for Limited-Angle Computed Tomography Reconstruction. in *Bildverarbeitung für die Medizin 2017* 92–97 (Springer Berlin Heidelberg, 2017).

91. Gjesteby, L. *et al.* Deep learning methods for CT image-domain metal artifact reduction. in *Developments in X-Ray Tomography XI* vol. 10391 103910W (International Society for Optics and Photonics, 2017).

92. El-Gamal, F. E.-Z. A., Elmogy, M. & Atwan, A. Current trends in medical image registration and fusion. *Egyptian Informatics Journal* **17**, 99–124 (2016).

93. Yang, X., Kwitt, R., Styner, M. & Niethammer, M. Quicksilver: Fast predictive image registration - A deep learning approach. *Neuroimage* **158**, 378–396 (2017).

94. Ngiam, J. *et al.* Multimodal deep learning. in *Proceedings of the 28th international conference on machine learning (ICML-11)* 689–696 (2011).

95. Yankeelov, T. E., Abramson, R. G. & Quarles, C. C. Quantitative multimodality imaging in cancer research and therapy. *Nat. Rev. Clin. Oncol.* **11**, 670–680 (2014).

96. Johnson, A. J., Chen, M. Y. M., Zapadka, M. E., Lyders, E. M. & Littenberg, B. Radiology report clarity: a cohort study of structured reporting compared with conventional dictation. *J. Am. Coll. Radiol.* **7**, 501–506 (2010).

97. Levy, M. A. & Rubin, D. L. Tool support to enable evaluation of the clinical response to treatment. *AMIA Annu. Symp. Proc.* 399–403 (2008).

98. European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). *Insights Imaging* **2**, 93–96 (2011).

99. Folio, L. R. *et al.* Quantitative Radiology Reporting in Oncology: Survey of Oncologists and Radiologists. *AJR Am. J. Roentgenol.* **205**, W233–43 (2015).

100. Karpathy, A. & Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 664–676 (2017).

101. Shin, H.-C. *et al.* Learning to read chest X-rays: recurrent neural cascade model for automated image annotation. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2497–2506 (2016).

102. Lee, J.-G. *et al.* Deep Learning in Medical Imaging: General Overview. *Korean J. Radiol.* **18**, 570–584 (2017).

103. Statistics / Health care use / Computed tomography (CT) exams. doi:10.1787/3c994537-en.

104. OECD. Magnetic resonance imaging (MRI) exams. (2015) doi:10.1787/1d89353f-en.

105. Bryan, S. *et al.* Radiology report times: impact of picture archiving and communication systems. *AJR Am. J. Roentgenol.* **170**, 1153–1159 (1998).

106. Mansoori, B., Erhard, K. K. & Sunshine, J. L. Picture Archiving and Communication System (PACS) implementation, integration & benefits in an integrated health system. *Acad. Radiol.* **19**, 229–235 (2012).

107. Lemke, H. U. PACS developments in Europe. *Comput. Med. Imaging Graph.* **27**, 111–120 (2003).

108. Mendel, J. B. & Schweitzer, A. L. PACS for the developing world. *Journal of Global Radiology* **1**, 5 (2015).

109. Kamnitsas, K. *et al.* Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks. in *Information Processing in Medical Imaging* 597–609 (Springer, Cham, 2017).

110. Kallenberg, M. *et al.* Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring. *IEEE Trans. Med. Imaging* **35**, 1322–1331 (2016).

111. Zhang, P., Wang, F. & Zheng, Y. Self supervised deep representation learning for fine-grained body part recognition. in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* 578–582 (2017).

112. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).

113. Wang, G. A Perspective on Deep Imaging. *IEEE Access* **4**, 8914–8924 (2016).

114. Ford, R. A., Price, W. & Nicholson, I. I. Privacy and Accountability in Black-Box Medicine. *Mich. Telecomm. & Tech. L. Rev.* **23**, 1 (2016).

115. Selbst, A. D. & Powles, J. Meaningful information and the right to explanation. *International Data Privacy Law* **7**, 233–242 (2017).

116. Imming, P., Sinning, C. & Meyer, A. Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.* **5**, 821–834 (2006).

117. Drugs with Unknown Antiparasitic Mechanism of Action. in *Encyclopedia of Parasitology* (ed. Dr., H. M.) 400–402 (Springer Berlin Heidelberg, 2008).

118. Shokri, R. & Shmatikov, V. Privacy-Preserving Deep Learning. in *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security* 1310–1321 (ACM, 2015).

119. Phong, L. T., Aono, Y., Hayashi, T., Wang, L. & Moriai, S. Privacy-Preserving Deep Learning: Revisited and Enhanced. in *Applications and Techniques in Information Security* (eds. Batten, L., Kim, D. S., Zhang, X. & Li, G.) vol. 719 100–110 (Springer Singapore, 2017).

120. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-Efficient Learning of Deep Networks from Decentralized Data. in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2017).

121. Gilad-Bachrach, R. *et al.* CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy. in *International Conference on Machine Learning* 201–210 (2016).

122. Cahan, A. & Cimino, J. J. A Learning Health Care System Using Computer-Aided Diagnosis. *J. Med. Internet Res.* **19**, e54 (2017).

4

Chapter 4

Artificial Intelligence in Radiation Oncology

E Huynh, *A Hosny*, C Guthier, D Bitterman, S Petit, DA Haas-Kogan, B Kann, HJWL Aerts & RH Mak

Nature Reviews Clinical Oncology 2020

Abstract

Artificial Intelligence (AI) has the potential to fundamentally alter the way medicine is practiced, as it excels in recognizing complex patterns in medical data and provides a quantitative, rather than qualitative, assessment of clinical conditions. In particular, the field of radiation oncology has the potential to be transformed by AI given its multifaceted, highly technical nature with heavy reliance on digital data processing and computer software, with the potential to improve the accuracy, precision, efficiency, and overall quality of radiation therapy (RT) for cancer patients. In this perspective article, we begin with a general description of AI methods, then furnish the reader with a high-level overview of the RT workflow and the impact that AI may have on each of these steps. Lastly, the challenges of clinical development and implementation of AI in RT are discussed, and our perspective on the changing roles of RT medical professionals.

Introduction

Radiation therapy (RT) plays a critical role in the treatment of cancer, and is indicated in ~50% of cancer patients. However, it is estimated that millions of patients currently lack access to this vital treatment[1–5], due to barriers such as resource scarcity (e.g. facilities, treatment machines, treatment planning systems, etc.), and trained staff[6]. Furthermore, RT has become increasingly complex over the past few decades requiring near complete reliance on human-machine interaction including both software and hardware.

Despite technological advances, much of the RT workflow still requires time-consuming, manual labor by a team of medical staff including radiation oncologists, medical physicists, medical dosimetrists, and radiation therapists. The growing complexity of the human-machine interactions in conjunction with the increasing incidence of cancer have created workforce shortages throughout the world and increasingly variable quality of care. In fact, variations in the treatment planning process have been shown to negatively impact overall survival even in clinical trials where extra care is given to standardizing approaches[7,8]. Furthermore, the RT knowledge and experience gap between adequately- and under-resourced health care environments poses an enormous public health challenge, and represents one of the great global inequities in cancer care.

Artificial intelligence (AI) is transforming multiple fields of medicine, and has the potential to address many of the challenges faced in RT to improve access to and the quality of cancer care throughout the world. Here, we provide an overview of the potential for AI to transform the field of RT by walking through each step of the workflow and highlighting examples where AI may increase efficiency, accuracy and quality of RT, thereby enhancing value-based cancer care delivery in today's resource-limited healthcare environment. While the breadth of applications of AI in RT is widespread, we have not covered all applications in this article, as we aimed to provide a glimpse of the transformative potential of AI in RT and our perspective on the future of the radiation oncology workforce.

Artificial Intelligence (AI) Methods

Early AI applications relied on rule-based reasoning, a set of human expert-defined steps and procedures to be followed by a computer system[9,10]. However, these methods often failed to generalize to variation in input data and task scope given that lack of intelligent components for dealing with edge cases not explicitly described in the knowledge base[11]. Rule-based AI systems found varying degrees of clinical utility[12] until the 2010's when there was a fundamental shift in the algorithms powering the automation of image-based tasks. This shift was marked by the revival of neural networks, a class of machine learning algorithms loosely based on our presumed understanding of how the human brain functions.

Research in neural networks has evolved from the mathematical developments of backpropagation in the 1960's[13] - the main mechanism in training neural networks -

towards simple networks in the 1980's[14,15]. The large amounts of data available today, increased computational power, and advances in algorithm development have all revived interest in the subject leading to "deeper" neural networks with multiple intermediate hidden layers between inputs and outputs. The utilization of such algorithms has obviated the need for rule predefinition, as a mapping between inputs and outputs can be learned from training data automatically. This approach provides deep learning algorithms a larger learning capacity than its predecessors and subsequently an ability to approximate very complex non-linear relationships in data. Deep learning can therefore begin to approximate human capabilities for highly complex tasks, and has been applied in several medical scenarios[16].

The RT workflow contains a multitude of complex tasks, including tumor and organ segmentation, dose optimization, outcome prediction, and quality assurance, which have seen varying degrees of digitization and consequent automation over the years. This heterogeneity is also reflected in the data types used, ranging from radiographic images and dose maps to hardware calibration log files and maintenance records. A non-exhaustive list of recent AI algorithms and examples of tasks addressed in radiation oncology can be found in Table 1. The multimodal nature of deep learning architectures[17] may allow for cross-modality learning, generalizability, and aggregation across these different data streams for improved clinical decision making and better quality of cancer care for all patients[18].

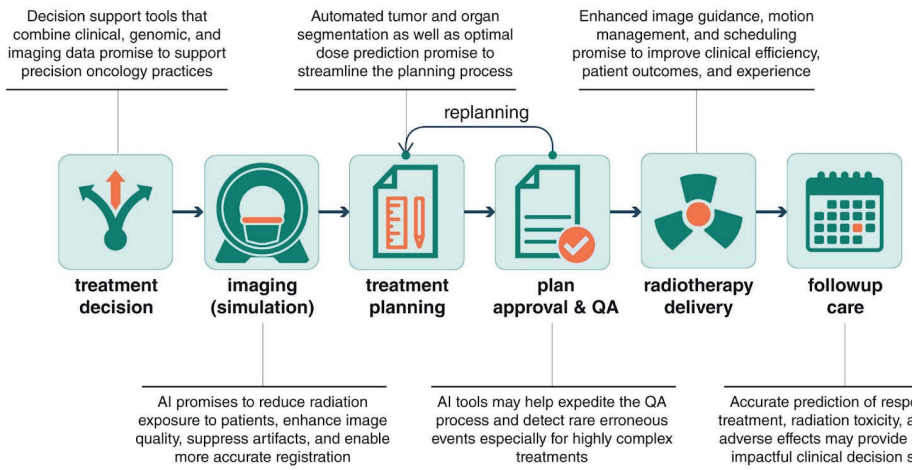

Figure 1. A general overview of the radiation therapy (RT) workflow with brief descriptions of expected AI applications in each step. The RT workflow begins with the decision to treat the patient with RT, followed by a simulation appointment where images are acquired for treatment planning. After the patient-specific treatment plan is created, the plan is approved, reviewed and has quality assurance (QA) measures performed on it prior to delivery to the patient. The patient is then seen for follow-up care.

Impact of AI on the Radiation Therapy Workflow

The RT workflow can be divided into several steps including treatment decision, preparation, planning, delivery and follow-up **(Figure 1)**. Here we described key tasks of each step, the staff involved, and provide a few examples of how AI may have an impact. For steps in the workflow where we do not foresee AI to have a significant impact (e.g. the actual delivery of radiation), we have not provided examples.

A. Evaluation and Clinical Plan

I. Patient evaluation

The first step of the RT clinical workflow is patient intake and evaluation, which is typically a consultation by the radiation oncologist and includes reviewing the patient's symptoms, medical history, physical exam, pathological and genomic data, diagnostic studies, prognostication, comorbidities, and potential toxicity from RT, and making a recommendation for a treatment plan based on a synthesis of these data. An emerging challenge for clinicians is the continuing growth of data available that are orders of magnitude beyond what the human can rapidly identify and interpret. AI-based methods that can automatically extract key features that are clinically actionable will be critical to building decision support tools for the clinician at the initial point of care. Advances in AI approaches for medical imaging[19] and natural language processing for electronic medical records,[20,21] have shown initial promise in guiding treatment selection and/or clinical management. These AI-built models have been reported to potentially improve prognostication[22,23] and predict outcomes after treatment[21,24–26] but have not seen clinical implementation yet.

II. Dose prescription

The prescribed dose to the tumor and dose constraints to the organs are determined by the radiation oncologist prior to treatment planning based on nationally accepted standards and clinical trial data. However, variation in tumor biology may result in substantially different radiation sensitivity, even for a given cancer type. Furthermore, depending on the geometrical arrangement of the tumor and organs, the desired dose may not be achievable, which is often not known until after the planning process is near completion. AI may enable personalization of RT by predicting the tumor's radiation sensitivity[27], what is achievable in a treatment plan based on contours of the tumor and organs, and an optimal dose prescription.

B. Treatment Preparation

I. Treatment Simulation (imaging for planning)

In preparation for treatment planning, planning appointments take place where a patient is immobilized to prevent significant motion during treatment. In most cases, medical images are acquired in this position, which are then used for treatment planning. Depending on the disease site, this process can be very complex and user-dependent, often requiring physician and physicist involvement. For example, special consideration is taken to evaluate the potential interference between areas in the immobilization device and treatment beam angles or patient-specific issues that may result in collision with the treatment machine. Similar to how AI has been able to expedite treatment planning

based on a patient's anatomy[28–30], we speculate that AI could have a role in identifying challenges that may be encountered at treatment simulation based on the patient's anatomy, and offer solutions input from the algorithm training data, thus expediting and optimizing the planning process.

i. Patient image acquisition
Many patients treated with RT require multiple medical images for treatment planning such as computed tomography (CT) for calculating radiation dose, and MRI for segmentation of tumors. Typically, these images are acquired in different patient positions (CT in the treatment position, other modalities acquired for diagnostic imaging in a different position), which introduces uncertainty when aligning the images. One method to minimize this uncertainty is to eliminate the need for a CT and acquire an MRI that can also provide electron density information (i.e. synthetic CT). AI has been employed to develop synthetic CTs from MRIs of the brain[31,32] and pelvis[33], with minimal dose differences in the treatment plans compared to the actual CT[31,33]. This could also improve clinical efficiency by reducing the number of appointments patients need to attend and radiation exposure from CT scans.

Advances in technology have led to the emerging role of MRI in RT, with the installation of integrated MRI-RT treatment units[34–36]. High resolution and low noise MRIs require long acquisition times, and a compromise is made between the resolution and signal-to-noise ratios necessary to suppress image noise and perform clinical tasks and image acquisition time. AI has the potential to reduce MR scan times by enabling reconstruction of fine structures from undersampled MRIs and has been developed for the generation of high resolution, high contrast and low noise brain[37–39] and cardiac MRI[40]. Due to the complexities of integrating MRI with a treatment machine, current systems are built with low strength magnets, typically 0.35-1.5 T[41–43] which reduces image quality compared to high resolution MR scans. AI could enable the reconstruction of high signal, high resolution images from low field strength images; for example, 7T-like images of the brain from 3T MRI[44] to improve the visualization of tumors throughout treatment.

ii. Image processing and registration
Image registration is an integral part of the RT workflow where data from multimodality and longitudinal images are used in treatment planning, delivery and monitoring radiation delivery. Commercially available automatic registration algorithms are typically designed to perform well only for modality-specific registration problems and are sensitive to image artifacts which compromises accuracy, and often requires additional manual edits to achieve a clinically acceptable registration.

AI tools have been trained to determine the sequence of motion actions to result in optimal image alignment, and shown better accuracy and robustness than several state-of-the-art methods45 and are generalizable across multiple imaging modalities45,46. Furthermore, AI has been able to improve registration robustness against imaging artifacts, such as with x-ray images of the spine that contained metal artifacts from metal screws and guide wires47 and motion artifacts, such as with fetal MRI48. AI tools have

been developed for initial applications in image registration with MRI49, x-ray50,4751, CT/MRI52 and MRI/PET registration53. Although many of these algorithms for image registration have not been developed in the context of RT, challenges they address are also faced in RT and could be applied here to improve the RT workflow.

II. Dosimetric Treatment Planning

i. Tumor segmentation for targeting radiation

Currently, one of the most time-consuming but critical steps, for the radiation oncologist is the manual segmentation of the primary tumor and affected lymph nodes. The accuracy of segmentation can directly impact outcomes; an incorrectly delineated tumor may lead to under- or over-dosing, resulting in a decrease in the likelihood of tumor control or increased toxicity, respectively. There is inter-observer variation in tumor segmentation, even among expert radiation oncologists[54,55], which can lead to differences in treatment plan quality and directly impact survival[7,8,56,577,56,57]. Current semi-automated segmentation tools that incorporate prior knowledge, such as segmentation atlases, have been unreliable or inaccessible to many radiation oncologists due to costs and still require significant manual effort[58,59].

AI approaches have the potential to dramatically increase the efficiency, reproducibility, and quality of RT planning by developing automated segmentation approaches, such as those developed for nasopharyngeal carcinomas[60], primary lung tumors[61], and oropharyngeal carcinomas[62]. The quality of these segmentations performed similarly against human experts. Further studies are required to directly compare the impact of AI approaches on efficiency, reproducibility and quality against the current clinical standard within the RT clinical workflow.

ii. Organ segmentation

Organs adjacent to the tumor are segmented in order to measure and restrict the radiation dose to those critical organs within safe limits during the planning process. Early AI approaches have demonstrated promise in the ability to delineate a variety of organs throughout the body including the complex anatomy of the head and neck region[63], thoracic organs[64], kidneys[65], liver[66,67], and cardiac substructures[68], however, these studies are limited by small training sets and potential over-fitting. The largest scale example of this approach involves an academic-industry partnership between Google DeepMind and partnering with the RT Department at University College London Hospitals to develop an algorithm capable of segmenting organs in the head and neck region with performance comparable to human experts using a training data set of 663 patients[69]. With commercially-available AI-based auto-segmentation tools starting to become available in treatment planning systems, there is a need for tools to perform quality assurance on these AI processes to identify errors from auto-segmentation[70].

iii. Treatment plan generation (dose optimization)

With medical images, segmentations, and dose prescription provided, the medical dosimetrist aims to generate the most optimal treatment plan for the patient with the goal of maximizing the dose delivered to the tumor while sparing surrounding organs. Treatment planning is a time-intensive, iterative process where the dosimetrist designs

the dose distribution, making the necessary changes in a trial-and-error based fashion to achieve the goals outlined in the prescription. The treatment plan is then evaluated by the radiation oncologist before approval for treatment. The plan quality achieved depends on several different human factors resulting in a large variation among treatment plans both intra- and inter-institutionally[71].

Current strategies that aim to standardize and improve efficiency involve automation of hard coded rules to perform repetitive tasks or optimization of plan parameters with pre-defined plan objectives using statistical methods[72,73]. These methods are designed for specific treatment sites, and have difficulty handling varying ranges of plan complexity, and patient-specific tradeoffs.

AI tools for automating treatment planning have two main steps: 1) predicting the optimal dose distribution, 2) identifying the treatment machine parameters to achieve that distribution. Several studies showed the ability of deep learning algorithms to predict optimal dose distributions for patients based on their anatomy[28–30] and to accelerate dose calculations[74]. In order for AI-based treatment planning algorithms to generate a high quality plan, the algorithms require information regarding the complex decision-making process to be included in the model, similar to those used to play ATARI games[75] or the board game Go[76]. Recent studies have applied these gamification concepts to automatically generate treatment plans for cervical cancer[77], and lung cancer[78]. Overall, AI techniques have the potential to substantially improve this critical step in the RT workflow by providing prediction of what radiation dose distributions can be safely achieved in advance so that clinicians can select the optimal treatment approach and then generating the treatment plan to deliver the optimal radiation dose. Thus, AI has the potential in the near-term to fully automate the treatment planning process.

C. Pre-treatment Review and Verification

After the clinician approves the treatment plan, medical physicists perform plan checks and other QA checks to ensure that all the technical components involved in treatment delivery are functioning and set correctly to deliver the intended dose to the patient. To reduce repetitive, time-consuming manual measurements, and improve efficiency, AI has been developed for some QA activities, such as patient-specific and machine QA measurements.

Patient-specific QA involves assessment of treatment plans to detect human error and potential anomalies in software and hardware machine performance. These include checking plan and treatment parameters, and verifying the patient's planned dose against the delivered dose. AI tools have been developed to expedite this process, and to detect rare events. For example, a physical measurement is currently performed for highly complex treatments, to compare the planned and delivered dose. While the majority of plans pass this QA, in the rare case that a plan fails, there are many contributing factors that require investigation and may delay patient treatment. An AI algorithm was designed to predict QA passing rates from the plan itself and identifying the potential sources of error, eliminating the need for physical dose measurement[79,80].

Machine QA evaluates the accuracy and precision of treatment machine characteristics and are conducted on a daily, weekly, monthly and annual basis. The plethora of data acquired has provided the means to develop AI algorithms that are capable of predicting trends and errors, such as multi-leaf collimator positional errors[81], beam symmetry trends[82], and to automatically detect imaging artifacts[83].

D. Treatment

I. Treatment set-up and delivery

i. Scheduling

Patients enter the department for several appointments including consultation, radiation planning, treatment, and follow-ups, all of which can have varying durations and wait times. Extended wait times impact both the efficiency of the clinic and patient anxiety and satisfaction[84]. AI has the potential to identify the most critical factors that contribute to wait time duration (such as time of day, fraction number, median past duration of treatments, number of treatment fields and previous treatment duration[85]), and predict wait times, enabling optimization of clinic flow and efficiency. Appointment scheduling may be further optimized based on the treatment site, immobilization and treatment technique used to decrease the room turnover time between patients; thereby increasing efficiency and accommodating a higher patient load.

ii. Image guidance and motion management

A key part of RT delivery is setting up the patient in the same position that the treatment plan was created. Currently, the most common on-treatment imaging method uses the treatment machine's cone beam CT (CBCT) to set-up the patient, which suffers from severely decreased image quality compared to the planning CT. AI has been applied to improve the image quality of CBCT for better patient set-up[86]. Increasingly complex and multi-modality imaging techniques are being incorporated into image-guided RT including the use of on-board MRI, ultrasound, optical surface imaging, and represent a unique opportunity for imaging-based AI methods to enhance and/or synthesize complex data at the point of care.

Patient or organ motion throughout treatment can result in increased dose to normal tissue in order to ensure that the tumor volume is adequately treated. Motion management methods aim to reduce, capture and/or monitor the extent of motion from respiration and/or digestion. Variability in motion exists between and within individuals in magnitude, amplitude and frequency, and movement relative to other organs requiring predictive modelling of tumor motion. AI may accommodate these factors by generating patient-specific models that can adapt to changes in motion patterns to improve tumor tracking. Thus far, research in this area has largely focussed on the prediction of respiratory motion using data collected from external surrogates[87–89] as inputs for the models. These algorithms could be automatically adjusted in real-time for complex breathing patterns[87].

II. Adaptive treatment
Significant deviations in a patient's anatomy between the planning appointment and actual treatment (days to weeks later) or throughout treatment (over several weeks) may warrant re-planning. These deviations are often due to tumor shrinkage/growth or anatomical variations that can result in varying dose to the tumor and organs. Adaptive treatments involve creating a new treatment plan based on an updated image of the patient's anatomy. Currently the physician must decide when anatomic changes are large enough to be clinically relevant based on their qualitative clinical assessment of a patient's clinical parameters and images. AI may provide the tools to predict which patients will require adaptation and the ideal time point at which it should occur, such as AI tools developed for head and neck patients to predict geometric changes throughout treatment[90,91]. Similar approaches have been applied for lung cancer patients to identify the need for plan adaptation based on changes between the initial and on-treatment images[92] to maximize tumor local control and reduce radiation-induced pneumonitis[78].

E. Completion
I. Response assessment and follow-up care
The Response Evaluation Criteria in Solid Tumors[93] is the most widely adopted system for evaluating treatment response of solid tumors based on the size and presence of the tumor. AI has the potential to provide more detailed information about the tumor's response to radiation throughout the course of treatment, such as changes in the tumor phenotype captured in imaging features, that may provide better assessment and prediction of response. Early studies have used AI with pre- and post-treatment imaging for early assessment of response in lung[27,94,95], bladder[96] pancreatic cancer[97,98].

Furthermore, the presence of radiation-induced organ damage can obfuscate the detection of disease recurrence. Early studies have shown that AI has the potential to detect early changes in the lung that were associated with local recurrence and may be overlooked by physicians as radiation-induced fibrosis[99]. This additional information would enable earlier, personalized treatment interventions to improve outcomes.

II. Toxicity Prediction and Management
Managing acute and late toxicities in patients is difficult due to the unpredictability of its presence and/or severity. Predictive models of radiation toxicity may be generated from risk factors, including clinical data, germline genomics and dose distribution, and imaging data to guide treatment planning. To date, most approaches have focused on subsets of this data and/or extrapolation of radiobiological modeling from pre-clinical and observational studies.

AI may be poised to analyze these data streams more comprehensively to build more robust models[100], that incorporate co-morbidities, dose and pre-treatment imaging data to provide clinical decision support for anticipatory management and secondary prevention. For example, AI-based normal tissue complication probability models were developed for head and neck cancer patients to predict the severity of acute dysphagia[101], xerostomia[102], and oral mucositis[103]. Other studies have developed models for predicting

radiation-induced pneumonitis[104–106], esophagitis[107], rectal toxicity[108], and epilepsy[109] in other disease sites.

On-treatment clinical data can also be used to provide guidance on potentially severe toxicity. AI methods trained on clinical data from electronic medical records have been demonstrated to accurately predict the risk of acute toxicities, leading to emergency room visits and hospital admission for patients receiving chemoradiation[21]. The integration of multiple clinical datastreams to provide advanced forecasting of adverse events during RT is a representative example of the power of AI to provide real-time, impactful clinical decision support at the point of care.

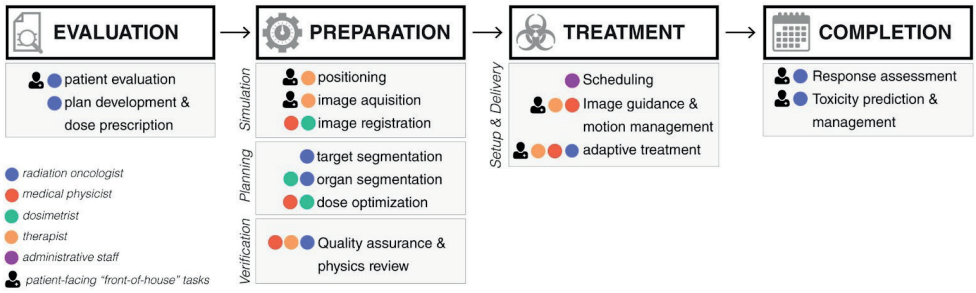

Figure 2. Detailed breakdown of the RT workflow with indications of staff involvement and patient-facing steps. The RT workflow can be broken down into evaluation and clinical plan, treatment preparation, treatment and completion. Within each of these categories, various steps are outlined with staff involved in each step, such as the radiation oncologist, medical physicists, dosimetrists, therapists and administrative staff.

Development Challenges

Multiple challenges lie ahead of developing clinical AI solutions, with data arguably being the most critical component. The amount of data needed for high accuracy AI applications strongly depends on the application and the nature of the outcome data. The wealth of data generated with every patient often requires laborious curation before it can be utilized in developing AI models, especially given the lack of standards in its generation. Areas that suffer from weak standard definitions include organ definitions and anatomical extents[110], treatment techniques, tumor recurrence, toxicity severity grading, and the concepts and metrics used to evaluate treatment plans[111,112]. This inhibits the sharing and aggregation of data across institutions- a prerequisite for developing AI models that accurately capture the full breadth of clinical variation while avoiding bias toward local standards. While medical data repositories such as The Cancer Imaging Archive[113] have helped promote sharing practices and professional organizations have attempted to standardize ontology[114,115], more work remains ahead.

The proprietary nature of the treatment planning software for their optimization algorithms is another hurdle facing the development of AI solutions. This challenge has

been alleviated as some vendors start to release application programming interfaces that allow research efforts to communicate with and integrate into clinical software, albeit with restricted scopes.

Furthermore, early research has focused on easily measured outcomes, such as overall survival, which may not be the best outcome of interest for all patients treated with RT. Instead, AI solutions will begin to move toward outcomes that are more directly pertinent to RT, such as the prediction of radiation-specific outcomes (e.g. tumor control, radiation toxicities), however, the collection of robust outcomes continues to be a challenge.

Clinical Implementation Challenges

Clinical adoption is a key barrier to realizing the potential of AI in RT as the introduction of AI tools will require an upfront investment of time and effort to understand their utility and limitations, and redesign current clinical workflows. Many research studies remain at the proof-of-concept stage and lack external validation[116], resulting in a slow translation into routine practice[117] where demonstration of generalizability and effectiveness becomes unattainable. Establishing trust in AI systems is also crucial, given the black box nature of many machine learning algorithms and specifically deep learning. Despite active research in AI interpretability[118], this lack of transparency hinders our ability to understand AI outputs, predict failures, and troubleshot generalizability issues. Without actively monitoring deployed AI performance, as well as continuous assessment of training data fit to the problem at hand, errors may increase as systematic biases are introduced into these systems.

Current AI tools are not perfectly accurate, and thus, three criteria can be used to evaluate the potential for clinical implementation: 1) time and ability for the user to judge the accuracy of the result, 2) correct the erroneous result, and, 3) consequence of it on a patient. Even in the case of severe consequences, clinical implementation can be fairly straightforward as long as inaccuracies by the model are detected by staff and corrected before moving on to the next step in the RT workflow. However, if the time and ability required for the user to judge the accuracy of the result outweighs the efficiency or accuracy gains of using an AI-tool, the potential for clinical implementation will be lower. Furthermore, for applications where the user cannot judge the correctness of the result, for instance when a tumor is not visible on the image and an AI tool is used for auto-segmentation, the risk-benefit ratio of using the AI-based tool is much more challenging. Tasks assisted or completed by AI that could have a significant impact on a patient's treatment will face a greater challenge with clinical implementation because of the potential consequence to the patient.

From a legal standpoint, means of governing algorithmic decision-making are yet to be fully developed, including the right to be given an explanation for algorithm outputs as well as implications of data protection laws[119,120]. While AI has the potential to reduce medical errors, it is also expected to alter the legal landscape around clinical liabilities

and responsibilities[121]. In terms of ethics, algorithms used for facial detection[122] and predicting offenders' risk of recidivism[123] have already demonstrated inherent racial bias, with applications in health care already starting to suffer similar obstacles[124]. The increased utilization of AI will change the dynamics of the patient-doctor relationship, while unethical AI may be developed by parties with ulterior motives and skew results toward financial gain[125]. All these challenges must be addressed ahead of effective widespread adoption.

Regulation and Clinical Evaluation

Currently, AI technologies are classified as software as a medical device (SaMD) by the US Food & Drug Administration and international regulatory bodies[126]. Many of its applications in RT will fall under these regulatory standards, such as treatment planning decision-support software which has been explicitly identified as a SaMD[127,128]. While much discussion has recently focused on timing of re-evaluation of new devices and locked versus continuously learning AI algorithms[129], clearer standards for clinical evaluation to determine the utility of these devices are needed. AI tools can have implications for patient outcomes in ways that can only be identified with robust retrospective or prospective studies carried out in representative populations.

While randomized clinical trials are the gold standard for evaluation of oncology therapies, this is neither feasible nor necessary for all AI tools. Tools that have the potential to affect outcomes, costs and efficiency should be considered for prospective clinical evaluation[130]. Given the rapid proliferation of these technologies, master protocols evaluating multiple technologies of a single class across a range of malignancies may make such efforts more efficient and feasible[130]. While Phase I/II studies may be adequate for low-risk devices that remain under provider surveillance, Phase III studies will be needed for high-risk tools that are used without standard clinical oversight. Post-market surveillance will be critical to assess the value of AI-based RT devices as they interact with other hardware and software which may affect their function. High-quality, risk-stratified clinical validation can establish the value of, and engender trust in these devices, which is particularly important for these black-box systems that can have an impact on cancer care.

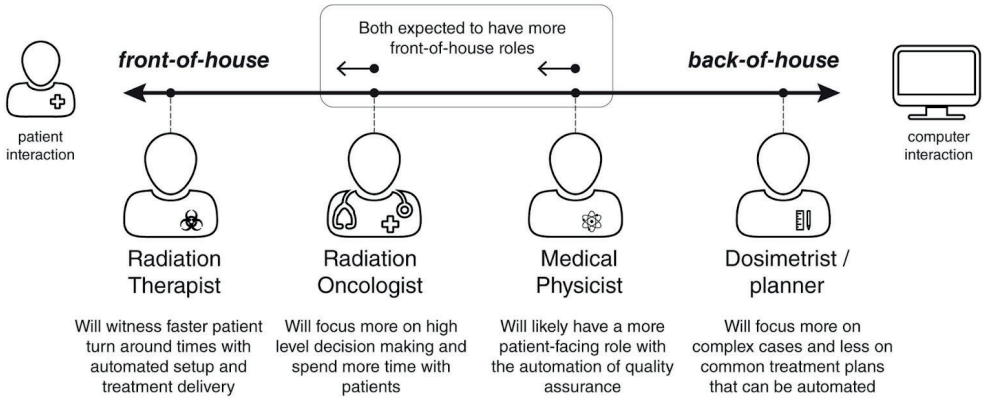

Figure 3. Members of the RT workforce (therapists, radiation oncologists, medical physicists and dosimetrists) are shown along a spectrum of interactions with patients and computers. Our projection of how each profession is expected to evolve with the clinical integration of AI tools is shown and described.

Perspective: Impact of AI on the Radiation Oncology Workforce

As the shift toward AI integration into RT clinics unfolds over the next few decades, the role of existing staff will be redefined, especially those that spend time on repetitive manual tasks. AI will predominantly impact staff members that perform "back-of-house" activities, including the technical aspects of RT such as segmentation, plan design, and QA, with less of an impact on "front-of-house" activities, that have direct interaction with the patient, typically carried out by physicians, therapists and nurses **(Figure 3)**. As the role of nursing is predominantly patient-facing, their roles will not be significantly changed by the integration of AI in the clinic.

A. Impact on Radiation Oncologists

As AI-based segmentation algorithms begin to replace the manual work of radiation oncologists, there will be a shift in focus on quality control of AI output and provide more time for high-value, "front-of-house" activities of human interaction, such as patient counseling, education, support and clinical management. Implementation of AI solutions will likely increase standardization of tumor segmentations and reduce unwarranted variation, particularly in under-resourced health care environments, which may translate into improved clinical outcomes and quality of care.

Training will need to evolve from current residency training models that focus on memorization of clinical facts and lengthy apprenticeships in order to gain expertise to perform manual segmentation and evaluate plans. Instead, we predict that future training will focus on a deeper understanding of how to integrate and interpret information from large data-sets to support clinical decision making.

B. Impact on Medical Physicists

AI has the potential to reduce the frequency and/or breadth of routine QA tasks of medical physicists by analyzing patterns and trends to predict when a technology may need to be serviced. This would cause a shift in the focus of physicists towards proactive prevention of non-routine, high-risk problems and implementation of new technologies that require human creativity and intuition. As the field of RT moves towards greater complexity treatments, the role of the physicist will continue to be key to ensuring the accuracy, precision and clinical release of technologies involved, including AI. Thought leaders have called for transitioning physicists from "back-of-house" work into a more clinical, patient-facing role as a means of improving the quality of information provided to patients, as well as enhancing their experience and satisfaction[131–133]. If realized, and with appropriate training[132], our perspective is that the physicist's role will be further strengthened, even with automation of their technical tasks.

C. Impact on Medical Dosimetrists/Treatment Planners

The medical dosimetrist currently performs many of the manual treatment planning tasks, which are most likely to be disrupted by AI approaches. Studies have shown that variation in plan quality is generally attributed to the overall "planner skill" as opposed to other parameters including experience, certification, and education[134]. This suggests the potential benefits of automating dosimetrists' tasks, especially to reduce variability of delivered care. The potential for auto-generated treatment plans to reduce the workload for medical dosimetrists has been suggested to be reliant on the clinical accuracy of the plans generated[135]. Further study is required to provide the confidence for a shift towards complete automation, yet early studies have demonstrated promising potential. In the short term, the dosimetrist's scope is expected to focus on more high risk and complex cases that are challenging for current AI approaches. AI will likely disrupt this profession substantially in the long term, due to automation. According to the 2017 American Association of Medical Dosimetry salary survey, 45% of respondents felt they were understaffed[136] – automation may have a place for reducing the dosimetrists' workload to reach appropriate staffing levels, or lead to significant reductions in the number of dosimetrists.

D. Impact on Radiation Therapists

The radiation therapists serve as the final gatekeeper of therapeutic delivery to ensure safe treatments and avoid treatment misadministration. AI could provide software tools to help the therapists ensure accurate and safe treatments, and increase efficiency and patient access, however, we believe that the radiation therapists will continue to serve an important role in being present to monitor the performance of these automated systems and the patient.

Future Perspectives

Beyond gains in accuracy, reproducibility and consistency, partnering human intuition and the capacity of AI to handle large data sets has the potential to drastically improve

efficiency and throughput in RT. This has recently become of prime importance in an era of cost reduction together with the shift from fee for service to value-based care[137].

The global health landscape also stands to benefit from AI interventions[138]. Over half of cancer patients live in low- and middle-income countries[139]. Workforce and equipment shortages in these resource-constrained settings have left over 50% of patients expected to benefit from RT without access to treatment, and up to 90% in some low-income countries[140]. Software AI applications promise to alleviate some of these shortages by providing specialized expert knowledge across disease sites and treatment modalities. Addressing hardware equipment shortages with AI, however, remains unclear, although AI may help support existing equipment upkeep by analyzing machine QA reports[82].

Ultimately, while the impact of AI will undoubtedly change the composition and skill set of the radiation oncology workforce, these changes will largely be for the positive and allow the field to continue to bend the cost curve through greater efficiency while improving the quality of care.

AI method	Description	Select applications in radiation oncology	Select examples
XGBoost	A prediction modeling technique consisting of an ensemble of weaker prediction models, usually decision-trees	Outcome prediction from structured data e.g. tabular data, comorbidities, dosimetric indices, age etc., as well as radiomic features extracted from images[141]	Prediction of radiation-related fibrosis of neck muscles in MRI for nasopharyngeal carcinoma patients[142]
		Artifact suppression in images e.g. motion management	Predict tumor motion ranges in 4-dimensional CT for lung radiotherapy patients[143]
Neural networks	Algorithms - loosely modeled after the human brain - comprising layers which in turn are composed of nodes that activate based on input	Radiation dose quality assurance	Pretreatment dose verification in prostate and nasopharynx radiotherapy patients[144]
Convolutional neural networks (CNN)	Neural networks that are composed of convolutional layers (for perception), followed by fully connected layers (for cognition)	Outcome prediction from unstructured data e.g. images	Rectum toxicity prediction in cervical cancer radiotherapy[145]
		Patient-specific quality assurance (QA) measurement	QA of dose distribution in prostate radiation patients[146]
Fully convolutional neural networks (FCN)	Neural networks that are composed entirely of convolutional layers. Images are encoded then decoded, thus producing a probability map per voxel pointing to a prediction class	Image segmentation in unstructured imaging data	Organ-at-risk segmentation in CT for head & neck radiotherapy patients[69]
		Radiation dose distribution prediction	3D dose distribution prediction of prostate stereotactic body radiation therapy[147], Dose distribution prediction of nasopharynx cancer[148]
Variational auto-encoders (VAE)	Neural networks that perform dimensionality reduction on input data converting it into low-dimensional latent vectors	Outcome prediction from unstructured data e.g. images	Prediction of lung radiation pneumonitis[149], predicting intrahepatic failure and overall survival in liver radiotherapy patients[150]
Generative adversarial networks (GAN)	Neural networks comprised of generator and discriminator components that participate together in a zero-sum game: The generator attempts to generate synthetic samples that match the input data distribution, while the discriminator attempts to discern synthetic from real data	Generation of synthetic CT images	Synthetic CT images for accurate MR-based dose calculations in the pelvis[33]
		Radiation dose distribution prediction	Predicting desirable 3D dose distributions in oropharyngeal cancer patients[151]
Deep Q-networks - Reinforcement learning (RL)	RL involves training an agent to interact with its environment by performing "actions" and arriving at "states". Certain actions lead to rewards which could be positive and negative	Radiation dose adaptation	Automated radiation adaptation protocols for non-small cell lung cancer patients[78]

Table 1. A non-exhaustive list of most recent AI methods and their applications in radiation oncology.

Author Contributions

The manuscript outline and design was developed by EH, AH, RM and HA. The manuscript was written by EH, AH, RM, HA, DB, CG, SP, and edited by DH and BK. Figures were designed by AH, EH and RM, and created by AH.

Acknowledgements

The authors acknowledge financial support from the US NIH (grants U24CA194354, U01CA190234, U01CA209414 and R35CA220523).

References

1. Pan, H. Y. *et al.* Supply and Demand for Radiation Oncology in the United States: Updated Projections for 2015 to 2025. *International Journal of Radiation Oncology*Biology*Physics* **96**, 493–500 (2016).
2. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–86 (2015).
3. Miller, K. D. *et al.* Cancer treatment and survivorship statistics, 2016. *CA Cancer J. Clin.* **66**, 271–289 (2016).
4. Atun, R. *et al.* Expanding global access to radiotherapy. *Lancet Oncol.* **16**, 1153–1186 (2015).
5. Grover, S. *et al.* A Systematic Review of Radiotherapy Capacity in Low- and Middle-Income Countries. *Front. Oncol.* **4**, (2015).
6. Elmore, S. N. C., Ben Prajogi, G., Rubio, J. A. P. & Zubizarreta, E. The global radiation oncology workforce in 2030: Estimating physician training needs and proposing solutions to scale up capacity in low- and middle-income countries. *Advances in Radiation Oncology* (2019).
7. Peters, L. J. *et al.* Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. *J. Clin. Oncol.* **28**, 2996–3001 (2010).
8. Brade, A. M. *et al.* Radiation Therapy Quality Assurance (RTQA) of Concurrent Chemoradiation Therapy for Locally Advanced Non-Small Cell Lung Cancer in the PROCLAIM Phase 3 Trial. *Int. J. Radiat. Oncol. Biol. Phys.* **101**, 927–934 (2018).
9. Kalet, I. J. & Paluszynski, W. Knowledge-based computer systems for radiotherapy planning. *Am. J. Clin. Oncol.* **13**, 344–351 (1990).
10. Laramore, G. E. *et al.* Applications of data bases and AI/expert systems in radiation therapy. *Am. J. Clin. Oncol.* **11**, 387–393 (1988).
11. Sanders, G. D. & Lyons, E. A. The potential use of expert systems to enable physicians to order more cost-effective diagnostic imaging examinations. *J. Digit. Imaging* **4**, 112–122 (1991).
12. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* (2018) doi:10.1038/s41568-018-0016-5.
13. Dreyfus, S. The numerical solution of variational problems. *J. Math. Anal. Appl.* **5**, 30–45 (1962).
14. Fukushima, K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980).
15. LeCun, Y., Haffner, P., Bottou, L. & Bengio, Y. Object Recognition with Gradient-Based Learning. in *Shape, Contour and Grouping in Computer Vision* (eds. Forsyth, D. A., Mundy, J. L., di Gesú, V. & Cipolla, R.) 319–345 (Springer Berlin Heidelberg, 1999).
16. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
17. Ngiam, J. *et al.* Multimodal deep learning. in *Proceedings of the 28th international conference on machine learning (ICML-11)* 689–696 (2011).

18. Feng, M., Valdes, G., Dixit, N. & Solberg, T. D. Machine Learning in Radiation Oncology: Opportunities, Requirements, and Needs. *Front. Oncol.* **8**, 110 (2018).

19. Kann, B. H. *et al.* Pretreatment Identification of Head and Neck Cancer Nodal Metastasis and Extranodal Extension Using Deep Learning Neural Networks. *Sci. Rep.* **8**, 14036 (2018).

20. Savova, G. K. *et al.* DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. *Cancer Res.* **77**, e115–e118 (2017).

21. Hong, J. C., Niedzwiecki, D., Palta, M. & Tenenbaum, J. D. Predicting Emergency Visits and Hospital Admissions During Radiation and Chemoradiation: An Internally Validated Pretreatment Machine Learning Algorithm. *JCO Clinical Cancer Informatics* 1–11 (2018).

22. Oberije, C. *et al.* A Validated Prediction Model for Overall Survival From Stage III Non-Small Cell Lung Cancer: Toward Survival Prediction for Individual Patients. *Int. J. Radiat. Oncol. Biol. Phys.* **92**, 935–944 (2015).

23. Jochems, A. *et al.* Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries. *Int. J. Radiat. Oncol. Biol. Phys.* **99**, 344–352 (2017).

24. Deist, T. M. *et al.* Expert knowledge and data-driven Bayesian Networks to predict post-RT dyspnea and 2-year survival. *Radiother. Oncol.* **118**, S29–S30 (2016).

25. M., D. T. *et al.* Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers. *Med. Phys.* **0**,.

26. Gilmer, V., Timothy, D. S., Marina, H., Lyle, U. & Charles, B. S., II. Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy. *Physics in Medicine & Biology* **61**, 6105 (2016).

27. Lou, B. *et al.* An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction. *The Lancet Digital Health* **1**, e136–e147 (2019).

28. Campbell, W. G. *et al.* Neural network dose models for knowledge-based planning in pancreatic SBRT. *Med. Phys.* **44**, 6148–6158 (2017).

29. Nguyen, D. *et al.* A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci. Rep.* **9**, 1076 (2019).

30. Häring, M., Großhans, J., Wolf, F. & Eule, S. Automated Segmentation of Epithelial Tissue Using Cycle-Consistent Generative Adversarial Networks. doi:10.1101/311373.

31. Dinkla, A. M. *et al.* MR-Only Brain Radiation Therapy: Dosimetric Evaluation of Synthetic CTs Generated by a Dilated Convolutional Neural Network. *Int. J. Radiat. Oncol. Biol. Phys.* **102**, 801–812 (2018).

32. Han, X. MR-based synthetic CT generation using a deep convolutional neural network method. *Med. Phys.* **44**, 1408–1419 (2017).

33. Maspero, M. *et al.* Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. *Phys. Med. Biol.* **63**, 185001 (2018).

34. Chandarana, H., Wang, H., Tijssen, R. H. N. & Das, I. J. Emerging role of MRI in radiation therapy. *J. Magn. Reson. Imaging* **48**, 1468–1478 (2018).

35. Rai, R. *et al.* The integration of MRI in radiation therapy: collaboration of radiographers and radiation therapists. *J Med Radiat Sci* **64**, 61–68 (2017).

36. Kerkmeijer, L. G. W. *et al.* The MRI-Linear Accelerator Consortium: Evidence-Based Clinical Introduction of an Innovation in Radiation Oncology Connecting Researchers, Methodology, Data Collection, Quality Assurance, and Technical Development. *Front. Oncol.* **6**, 215 (2016).

37. Hyun, C. M., Kim, H. P., Lee, S. M., Lee, S. & Seo, J. K. Deep learning for undersampled MRI reconstruction. *Phys. Med. Biol.* **63**, 135007 (2018).

38. Wang, S. *et al.* Accelerating magnetic resonance imaging via deep learning. in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (2016). doi:10.1109/isbi.2016.7493320.

39. Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R. & Rosen, M. S. Image reconstruction by domain-transform manifold learning. *Nature* **555**, 487–492 (2018).

40. Schlemper, J., Caballero, J., Hajnal, J. V., Price, A. N. & Rueckert, D. A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction. *IEEE Trans. Med. Imaging* **37**, 491–503 (2018).

41. Fallone, B. G. The Rotating Biplanar Linac–Magnetic Resonance Imaging System. *Semin. Radiat. Oncol.* **24**, 200–202 (2014).

42. Mutic, S. & Dempsey, J. F. The ViewRay System: Magnetic Resonance–Guided and Controlled Radiotherapy. *Semin. Radiat. Oncol.* **24**, 196–199 (2014).

43. Raaymakers, B. W. *et al.* Integrating a 1.5 T MRI scanner with a 6 MV accelerator: proof of concept. *Phys. Med. Biol.* **54**, N229–N237 (2009).

44. Bahrami, K., Shi, F., Rekik, I. & Shen, D. Convolutional Neural Network for Reconstruction of 7T-like Images from 3T MRI Using Appearance and Anatomical Features. in *Lecture Notes in Computer Science* 39–47 (2016).

45. de Tournemire P. Grbic S. Kamen A. Mansi T. and Comaniciu D., L. R. M. S. An Artificial Agent for Robust Image Registration. in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* 4168–4175.

46. Wu, G., Kim, M., Wang, Q., Munsell, B. C. & Shen, D. Scalable High-Performance Image Registration Framework by Unsupervised Deep Feature Representations Learning. *IEEE Trans. Biomed. Eng.* **63**, 1505–1516 (2016).

47. Miao, S., Piat, S., Fischer, P., Tuysuzoglu, A., Mewes, P., Mansi, T. and Liao, R. Dilated FCN for multi-agent 2D/3D medical image registration. in (2017).

48. Hou, B. *et al.* Predicting Slice-to-Volume Transformation in Presence of Arbitrary Subject Motion. in *Lecture Notes in Computer Science* 296–304 (2017).

49. Yang, X., Kwitt, R., Styner, M. & Niethammer, M. Fast predictive multimodal image registration. in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (2017). doi:10.1109/isbi.2017.7950652.

50. Miao, S., Jane Wang, Z., Zheng, Y. & Liao, R. Real-time 2D/3D registration via CNN regression. in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (2016). doi:10.1109/isbi.2016.7493536.

51. Kearney, V., Haaf, S., Sudhyadhom, A., Valdes, G. & Solberg, T. D. An unsupervised convolutional neural network-based algorithm for deformable image registration. *Phys. Med. Biol.* **63**, 185017 (2018).

52. Ma, K. *et al.* Multimodal Image Registration with Deep Context Reinforcement Learning. in *Lecture Notes in Computer Science* 240–248 (2017).

53. Suk, H.-I., Lee, S.-W. & Shen, D. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage* **101**, 569–582 (2014).

54. Van de Steene, J. *et al.* Definition of gross tumor volume in lung cancer: inter-observer variability. *Radiother. Oncol.* **62**, 37–49 (2002).

55. Cui, Y. *et al.* Contouring variations and the role of atlas in non-small cell lung cancer radiation therapy: Analysis of a multi-institutional preclinical trial planning study. *Pract. Radiat. Oncol.* **5**, e67–75 (2015).

56. Wuthrick, E. J. *et al.* Institutional Clinical Trial Accrual Volume and Survival of Patients With Head and Neck Cancer. *J. Clin. Oncol.* **33**, 156–164 (2015).

57. Ohri, N. *et al.* Radiotherapy protocol deviations and clinical outcomes: A meta-analysis of cooperative group clinical trials. *Journal of Clinical Oncology* vol. 30 181–181 (2012).

58. Delpon, G. *et al.* Comparison of Automated Atlas-Based Segmentation Software for Postoperative Prostate Cancer Radiotherapy. *Front. Oncol.* **6**, 178 (2016).

59. Kim, Y. *et al.* Impact of Contouring Accuracy on Expected Tumor Control Probability for Head and Neck Cancer: Semiautomated Segmentation Versus Manual Contouring. *Int. J. Radiat. Oncol. Biol. Phys.* **96**, E545 (2016).

60. Men, K. *et al.* Deep Deconvolutional Neural Network for Target Segmentation of Nasopharyngeal Cancer in Planning Computed Tomography Images. *Front. Oncol.* **7**, (2017).

61. Mak, R. H. *et al.* Use of Crowd Innovation to Develop an Artificial Intelligence–Based Solution for Radiation Therapy Targeting. *JAMA Oncology* vol. 5 654 (2019).

62. Cardenas, C. E. *et al.* Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function. *Int. J. Radiat. Oncol. Biol. Phys.* **101**, 468–478 (2018).

63. Ibragimov, B. & Xing, L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med. Phys.* **44**, 547–557 (2017).

64. Lustberg, T. *et al.* Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother. Oncol.* **126**, 312–317 (2018).

65. Jackson, P. *et al.* Deep Learning Renal Segmentation for Fully Automated Radiation Dose Estimation in Unsealed Source Therapy. *Front. Oncol.* **8**, 215 (2018).

66. Peijun, H., Fa, W., Jialin, P., Ping, L. & Dexing, K. Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution. *Physics in Medicine & Biology* **61**, 8676 (2016).

67. Ibragimov, B., Toesca, D., Chang, D., Koong, A. & Xing, L. Combining deep learning with anatomical analysis for segmentation of the portal vein for liver SBRT planning. *Phys. Med. Biol.* **62**, 8943–8958 (2017).

68. Morris, E. D. *et al.* Cardiac Substructure Segmentation with Deep Learning for Improved Cardiac Sparing. *Med. Phys.* (2019) doi:10.1002/mp.13940.

69. Nikolov, S. *et al.* Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv [cs.CV]* (2018).

70. Zhang, J., Ates, O. & Li, A. Implementation of a Machine Learning–Based Automatic Contour Quality Assurance Tool for Online Adaptive Radiation Therapy of Prostate Cancer. *International Journal of Radiation Oncology*Biology*Physics* **96**, E668 (2016).

71. Berry, S. L., Boczkowski, A., Ma, R., Mechalakos, J. & Hunt, M. Interobserver variability in radiation therapy plan output: Results of a single-institution study. *Pract. Radiat. Oncol.* **6**, 442–449 (2016).

72. Appenzoller, L. M., Michalski, J. M., Thorstad, W. L., Mutic, S. & Moore, K. L. Predicting dose-volume histograms for organs-at-risk in IMRT planning. *Med. Phys.* **39**, 7446–7461 (2012).

73. Hussein, M., Heijmen, B. J. M., Verellen, D. & Nisbet, A. Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations. *Br. J. Radiol.* 20180270 (2018).

74. Xing, Y., Nguyen, D., Lu, W., Yang, M. & Jiang, S. Technical Note: A feasibility study on deep learning-based radiotherapy dose calculation. *Med. Phys.* (2019) doi:10.1002/mp.13953.

75. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).

76. Chen, J. X. The Evolution of Computing: AlphaGo. *Computing in Science & Engineering* vol. 18 4–7 (2016).

77. Shen, C. *et al.* Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer. *Phys. Med. Biol.* **64**, 115013 (2019).

78. Tseng, H.-H. *et al.* Deep reinforcement learning for automated radiation adaptation in lung cancer. *Med. Phys.* **44**, 6690–6705 (2017).

79. Valdes, G. *et al.* A mathematical framework for virtual IMRT QA using machine learning. *Med. Phys.* **43**, 4323 (2016).

80. Valdes, G. *et al.* IMRT QA using machine learning: A multi-institutional validation. *J. Appl. Clin. Med. Phys.* **18**, 279–284 (2017).

81. Carlson, J. N. K. *et al.* A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. *Phys. Med. Biol.* **61**, 2514–2531 (2016).

82. Li, Q. & Chan, M. F. Predictive time-series modeling using artificial neural networks for Linac beam symmetry: an empirical study. *Ann. N. Y. Acad. Sci.* **1387**, 84–94 (2017).

83. Valdes, G. *et al.* Use of TrueBeam developer mode for imaging QA. *J. Appl. Clin. Med. Phys.* **16**, 322–333 (2015).

84. Paul, C. *et al.* Cancer patients' concerns regarding access to cancer care: perceived impact of waiting times along the diagnosis and treatment journey. *Eur. J. Cancer Care* **21**, 321–329 (2012).

85. A. Joseph, T. Hijal, J. Kildea, L. Hendren and D. Herrera. Predicting Waiting Times in Radiation Oncology Using Machine Learning. in *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on* 1024–1029 (McGill Univ. Health Centre, Montreal, QC, Canada, 2018).

86. Kida, S. *et al.* Cone Beam Computed Tomography Image Quality Improvement Using a Deep Convolutional Neural Network. *Cureus* **10**, e2548 (2018).

87. Isaksson, M., Jalden, J. & Murphy, M. J. On using an adaptive neural network to predict lung tumor motion during respiration for radiotherapy applications. *Med. Phys.* **32**, 3801–3809 (2005).

88. Kakar, M., Nyström, H., Aarup, L. R., Nøttrup, T. J. & Olsen, D. R. Respiratory motion prediction by using the adaptive neuro fuzzy inference system (ANFIS). *Phys. Med. Biol.* **50**, 4721–4728 (2005).

89. Murphy, M. J. & Pokhrel, D. Optimization of an adaptive neural network to predict breathing. *Med. Phys.* **36**, 40–47 (2009).

90. Guidi, G. *et al.* A support vector machine tool for adaptive tomotherapy treatments: Prediction of head and neck patients criticalities. *Phys. Med.* **31**, 442–451 (2015).

91. Guidi, G. *et al.* A machine learning tool for re-planning and adaptive RT: A multicenter cohort investigation. *Phys. Med.* **32**, 1659–1666 (2016).

92. Varfalvy, N., Piron, O., Cyr, M. F., Dagnault, A. & Archambault, L. Classification of changes occurring in lung patient during radiotherapy using relative γ analysis and hidden Markov models. *Med. Phys.* **44**, 5043–5050 (2017).

93. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).

94. Xu, Y. *et al.* Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clin. Cancer Res.* **25**, 3266–3275 (2019).

95. Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **15**, e1002711 (2018).

96. Cha, K. H. *et al.* Bladder Cancer Treatment Response Assessment in CT using Radiomics with Deep-Learning. *Sci. Rep.* **7**, 8738 (2017).

97. Chen, X. *et al.* Assessment of treatment response during chemoradiation therapy for pancreatic cancer based on quantitative radiomic analysis of daily CTs: An exploratory study. *PLoS One* **12**, e0178961 (2017).

98. Horvat, N. *et al.* MR Imaging of Rectal Cancer: Radiomics Analysis to Assess Treatment Response after Neoadjuvant Therapy. *Radiology* 172300 (2018).

99. Mattonen, S. A. *et al.* Detection of Local Cancer Recurrence After Stereotactic Ablative Radiation Therapy (SABR) for Lung Cancer: Physician Performance Versus Radiomic Assessment. *International Journal of Radiation Oncology*Biology*Physics* **96**, S48 (2016).

100. Lee, S. *et al.* Machine Learning on a Genome-wide Association Study to Predict Late Genitourinary Toxicity After Prostate Radiation Therapy. *International Journal of Radiation Oncology*Biology*Physics* **101**, 128–135 (2018).

101. Dean, J. *et al.* Incorporating spatial dose metrics in machine learning-based normal tissue complication probability (NTCP) models of severe acute dysphagia resulting from head and neck radiotherapy. *Clin Transl Radiat Oncol* **8**, 27–39 (2018).

102. Gabryś, H. S., 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.* **8**, 35 (2018).

103. Dean, J. A. *et al.* Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy. *Radiother. Oncol.* **120**, 21–27 (2016).

104. Cunliffe, A. *et al.* Lung texture in serial thoracic computed tomography scans: correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development. *Int. J. Radiat. Oncol. Biol. Phys.* **91**, 1048–1056 (2015).

105. Chen, S., Zhou, S., Yin, F.-F., Marks, L. B. & Das, S. K. Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis. *Med. Phys.* **34**, 3808–3814 (2007).

106. Moran, A., Daly, M. E., Yip, S. S. F. & Yamamoto, T. Radiomics-based Assessment of Radiation-induced Lung Injury After Stereotactic Body Radiotherapy. *Clin. Lung Cancer* **18**, e425–e431 (2017).

107. Luna, J. M. *et al.* Novel Use of Machine Learning for Predicting Radiation Esophagitis in Locally Advanced Stage II-III Non–small Cell Lung Cancer. *International Journal of Radiation Oncology*Biology*Physics* **99**, E476–E477 (2017).

108. Zhen, X. *et al.* Deep Convolutional Neural Networks With Transfer Learning for Rectum Toxicity Prediction in Combined Brachytherapy and External Beam Radiation Therapy for Cervical Cancer. *International Journal of Radiation Oncology*Biology*Physics* **99**, S168 (2017).

109. Liu, Z. *et al.* Radiomics analysis allows for precise prediction of epilepsy in patients with low-grade gliomas. *NeuroImage: Clinical* **19**, 271–278 (2018).

110. Wright, J. L. *et al.* Standardizing Normal Tissue Contouring for Radiation Therapy Treatment Planning: An ASTRO Consensus Paper. *Pract. Radiat. Oncol.* **9**, 65–72 (2019).

111. Covington, E. L. *et al.* Improving treatment plan evaluation with automation. *J. Appl. Clin. Med. Phys.* **17**, 16–31 (2016).

112. Evans, S. B. *et al.* Standardizing dose prescriptions: An ASTRO white paper. *Pract. Radiat. Oncol.* **6**, e369–e381 (2016).

113. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).

114. Mayo, C. S. *et al.* American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology. *Int. J. Radiat. Oncol. Biol. Phys.* **100**, 1057–1066 (2018).

115. Hayman, J. A. *et al.* Minimum Data Elements for Radiation Oncology: An ASTRO Consensus Paper. *Pract. Radiat. Oncol.* (2019) doi:10.1016/j.prro.2019.07.017.

116. Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y. & Park, S. H. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. *Korean Journal of Radiology* vol. 20 405 (2019).

117. Allen, B., Jr *et al.* A Road Map for Translational Research on Artificial Intelligence in Medical Imaging: From the 2018 National Institutes of Health/RSNA/ACR/The Academy Workshop. *J. Am. Coll. Radiol.* **16**, 1179–1189 (2019).

118. Gilpin, L. H. *et al.* Explaining Explanations: An Overview of Interpretability of Machine Learning. in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* 80–89 (2018).

119. Goodman, B. & Flaxman, S. European Union regulations on algorithmic decision-making and a 'right to explanation'. *AI Magazine* **38**, 50–57 (2017).

120. Kaminski, M. E. The right to explanation, explained. *Berkeley Tech. LJ* **34**, 189 (2019).

121. Harned, Z., Lungren, M. P. & Rajpurkar, P. Machine Vision, Medical AI, and Malpractice. (2019).

122. Buolamwini, J. & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (eds. Friedler, S. A. & Wilson, C.) vol. 81 77–91 (PMLR, 2018).

123. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. Machine bias. *ProPublica, May* **23**, 2016 (2016).

124. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).

125. Char, D. S., Shah, N. H. & Magnus, D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *N. Engl. J. Med.* **378**, 981–983 (2018).

126. [No title]. http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf.

127. Food & Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback. (2019).

128. [No title]. https://www.fda.gov/media/109618/download.

129. Hwang, T. J., Kesselheim, A. S. & Vokinger, K. N. Lifecycle Regulation of Artificial Intelligence- and Machine Learning-Based Software Devices in Medicine. *JAMA* (2019) doi:10.1001/jama.2019.16842.

130. Bitterman, D. S. *et al.* Master protocol trial design for efficient and rational evaluation of novel therapeutic oncology devices. *J. Natl. Cancer Inst.* (2019) doi:10.1093/jnci/djz167.

131. Schuller, B. W., Hendrickson, K. R. G. & Rong, Y. Medical physicists should meet with patients as part of the initial consult. *J. Appl. Clin. Med. Phys.* **19**, 6–9 (2018).

132. Brown, D. W. *et al.* A program to train medical physicists for direct patient care responsibilities. *J. Appl. Clin. Med. Phys.* **19**, 332–335 (2018).

133. Atwood, T. F. *et al.* Establishing a New Clinical Role for Medical Physicists: A Prospective Phase II Trial. *Int. J. Radiat. Oncol. Biol. Phys.* **102**, 635–641 (2018).

134. Nelms, B. E. *et al.* Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems. *Practical Radiation Oncology* vol. 2 296–305 (2012).

135. Adams, R. D. The future of medical dosimetry. *Med. Dosim.* **40**, 159–165 (2015).

136. American Association of Medical Dosimetrists. *2017 Salary Survey of Currently Active Medical Dosimetrists*. (2018).

137. Radiation Oncology Model | Center for Medicare & Medicaid Innovation. https://innovation.cms.gov/initiatives/radiation-oncology-model/.

138. Hosny, A. & Hugo J W. Artificial intelligence for global health. *Science* **366**, 955–956 (2019).

139. Barton, M. B., Frommer, M. & Shafiq, J. Role of radiotherapy in cancer control in low-income and middle-income countries. *Lancet Oncol.* **7**, 584–595 (2006).

140. Zubizarreta, E. H., Fidarova, E., Healy, B. & Rosenblatt, E. Need for radiotherapy in low and middle income countries--the silent crisis continues. *Clin. Oncol.* **27**, 107–114 (2015).

141. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).

142. Wang, J. *et al.* A Predictive model of radiation-related fibrosis based on radiomic features of Magnetic Resonance Imaging. *Int. J. Radiat. Oncol. Biol. Phys.* **105**, E599 (2019).

143. Lin, H. *et al.* A Super-Learner Model for Tumor Motion Prediction and Management in Radiation Therapy: Development and Feasibility Evaluation. *Sci. Rep.* **9**, 14868 (2019).

144. Mahdavi, S. R. *et al.* Use of artificial neural network for pretreatment verification of intensity modulation radiation therapy fields. *Br. J. Radiol.* **92**, 20190355 (2019).

145. Zhen, X. *et al.* Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. *Phys. Med. Biol.* **62**, 8246–8263 (2017).

146. Tomori, S. *et al.* A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance. *Med. Phys.* **45**, 4055–4065 (2018).

147. Kearney, V., Chan, J. W., Haaf, S., Descovich, M. & Solberg, T. D. DoseNet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks. *Phys. Med. Biol.* **63**, 235022 (2018).

148. Chen, X., Men, K., Li, Y., Yi, J. & Dai, J. A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning. *Med. Phys.* **46**, 56–64 (2019).

149. Cui, S., Luo, Y., Tseng, H., Ten Haken, R. K. & El Naqa, I. Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. *Med. Phys.* **46**, 2497–2511 (2019).

150. Wei, L. *et al.* Variational Autoencoder Graph-based Radiomics Outcome Modeling of Intrahepatic Progression Risk and Overall Survival for HCC post-SBRT Patients. *Int. J. Radiat. Oncol. Biol. Phys.* **105**, S83–S84 (2019).

151. Mahmood, R., Babier, A., McNiven, A., Diamant, A. & Chan, T. C. Y. Automated Treatment Planning in Radiation Therapy using Generative Adversarial Networks. *arXiv [cs.LG]* (2018).

II

PART II

Prognostic and Therapeutic Deep Learning Applications

5

Chapter 5

Deep Learning for Lung Cancer Prognostication: a Retrospective Multi-Cohort Radiomics Study

A Hosny, C Parmar, T Coroller, P Grossmann, R Zeleznik, A Kumar, J Bussink, RJ Gillies, RH Mak & HJWL Aerts

PLOS Medicine 2018

Abstract

Background: Non-small-cell lung cancer (NSCLC) patients often demonstrate varying clinical courses and outcomes, even within the same tumor stage. This study explores deep learning applications in medical imaging allowing for the automated quantification of radiographic characteristics and potentially improving patient stratification.

Methods and findings: We performed an integrative analysis on 7 independent datasets across 5 institutions totaling 1,194 NSCLC patients (age median = 68.3 years [range 32.5–93.3], survival median = 1.7 years [range 0.0–11.7]). Using external validation in computed tomography (CT) data, we identified prognostic signatures using a 3D convolutional neural network (CNN) for patients treated with radiotherapy (n = 771, age median = 68.0 years [range 32.5–93.3], survival median = 1.3 years [range 0.0–11.7]). We then employed a transfer learning approach to achieve the same for surgery patients (n = 391, age median = 69.1 years [range 37.2–88.0], survival median = 3.1 years [range 0.0–8.8]). We found that the CNN predictions were significantly associated with 2-year overall survival from the start of respective treatment for radiotherapy (area under the receiver operating characteristic curve [AUC] = 0.70 [95% CI 0.63–0.78], $p < 0.001$) and surgery (AUC = 0.71 [95% CI 0.60–0.82], $p < 0.001$) patients. The CNN was also able to significantly stratify patients into low and high mortality risk groups in both the radiotherapy ($p < 0.001$) and surgery ($p = 0.03$) datasets. Additionally, the CNN was found to significantly outperform random forest models built on clinical parameters—including age, sex, and tumor node metastasis stage—as well as demonstrate high robustness against test–retest (intraclass correlation coefficient = 0.91) and inter-reader (Spearman's rank-order correlation = 0.88) variations. To gain a better understanding of the characteristics captured by the CNN, we identified regions with the most contribution towards predictions and highlighted the importance of tumor-surrounding tissue in patient stratification. We also present preliminary findings on the biological basis of the captured phenotypes as being linked to cell cycle and transcriptional processes. Limitations include the retrospective nature of this study as well as the opaque black box nature of deep learning networks.

Conclusions: Our results provide evidence that deep learning networks may be used for mortality risk stratification based on standard-of-care CT images from NSCLC patients. This evidence motivates future research into better deciphering the clinical and biological basis of deep learning networks as well as validation in prospective data.

Author summary

Why was this study done?

- Cancer is one of the leading causes of death worldwide, with lung cancer being the second most commonly diagnosed cancer in both men and women in the USA.

- Prognosis in lung cancer patients is primarily determined through tumor staging, which in turn is based on a relatively coarse and discrete stratification.

- Radiographic medical images offer patient- and tumor-specific information that could be used to complement clinical prognostic evaluation efforts.

- Recent advances in radiomics through applications of artificial intelligence, computer vision, and deep learning allow for the extraction and mining of numerous quantitative features from radiographic images.

What did the researchers do and find?

- We designed an analysis setup comprising seven independent datasets across five institutions totaling 1194 NSCLC patients imaged with computed tomography and treated with either radiotherapy or surgery.

- We evaluated the prognostic signature of quantitative imaging features extracted through deep learning networks, and assessed its ability to stratify patients into low and high mortality risk groups as per a two-year overall survival cut off.

- In patients treated with surgery, deep learning networks significantly outperformed models based on predefined tumor features as well as volume and maximum diameter.

- In addition to highlighting image regions with prognostic influence, we evaluated the deep learning features for robustness against physiological imaging artifacts and input variability, as well as correlated them with molecular information through gene expression data.

What do these findings mean?

- We found that deep learning features significantly outperform existing prognostication methods in surgery patients, hinting at their utility in patient stratification and potentially sparing low mortality risk groups from adjuvant chemotherapy.

- We demonstrated that areas within and beyond the tumor - especially the tumor-stroma interfaces - had the largest contributions to the prognostic signature, highlighting the importance of tumor-surrounding tissue in patient stratification.

- Preliminary genomic associations in this study suggest correlations between the deep learning feature representations and cell cycle and transcriptional processes.

- Despite their obscure inner workings and lack of a strong theoretical backing, deep learning networks demonstrate a prognostic signal and robustness against specific noise artifacts. This motivates further prospective studies validating their utility in patient stratification and the development of personalized cancer treatment plans.

Introduction

Cancer's ever evolving nature and interaction with its surroundings continue to challenge patients, clinicians, and researchers alike. One of its deadliest forms appears in the lungs, leading to the most cancer-related mortalities worldwide[1]. Lung cancer is also the second most commonly diagnosed cancer in both men and women[2] with non-small cell lung cancer (NSCLC) comprising 85% of cases[3]. The ability to accurately categorize NSCLC patients into groups structured around clinical factors represents a crucial step in cancer care. This stratification allows for evaluating tumor progression, establishing prognosis, providing standard terminologies for effective clinical communication, and most importantly identifying appropriate treatment plans from chemotherapy and surgery to radiation and targeted therapy. In addition to clinical factors including performance status, and to a lesser extent, age and gender[4], tumor stage — as evaluated through the predominant tumor node metastasis (TNM) staging manual — is often regarded as a universal benchmark for performing such classification[5].

The TNM staging manual represents a body of knowledge combining evidence-based findings from clinical studies with empirical knowledge from site-specific experts[6]. However, we find that patients within the same stage can exhibit wide variations in their response to treatment[7]. This owes, in part, to the inevitable gap that exists between yesterday's statistics and today's more advanced treatment options, as well as the practical challenges of stratifying patients into groups that fit historical data, while balancing the ability of clinicians to identify the stratification features and apply the stratification algorithm at the point of care[8]. The limitations in our clinical gold standards, combined with our improved understanding of intra-tumor heterogeneity[9], both signal the need for developing personalized biomarkers that can operate at the individual patient- as opposed to the population-level — eventually leading to more robust patient stratification and building a foundation for precision oncology practices.

The aforementioned clinician-driven stratification algorithms used in NSCLC staging rely on high-level semantic features describing tumor extent, location, and metastatic status. These are often inferred from standard medical images of the upper abdomen and thorax. These non-invasive images, however, offer information that goes beyond that captured through routine radiographic evaluation. Hardware advances in high-resolution image acquisition equipment and computational processing power, coupled with novel artificial intelligence (AI) algorithms and large amounts of data, have all contributed to a proliferation of AI applications in radiology, medicine, and beyond. These have enabled the high-throughput extraction, and subsequent processing, of high-dimensional quantitative features from images. More specifically, this dialogue between AI and medical imaging has been recently manifested in radiomics.

Radiomics is a data-centric field involving the extraction and mining of quantitative features as a means to quantify the solid tumor radiographic phenotype[10]. It hypothesizes that radiographic phenotypes represent underlying pathophysiologies and are thus capable of discriminating between disease forms for predicting prognosis and therapeutic response[11]. Radiomics research has primarily relied on explicitly programmed algorithms

that extract engineered (hand-crafted) imaging features. Such features commonly represent tumor shape, voxel intensity information (statistics), and patterns (textures). More specifically within oncology, Radiomics has demonstrated success in stratifying tumor histology[12], tumor grades[13], and clinical outcomes[10]. Additionally, associations with underlying gene expression patterns have also been reported[14]. Given these associations, radiomic features have been used to build prognostic and predictive models making use of statistical machine learning algorithms coupled with feature selection strategies[15]. More recent work, however, has shifted towards deep learning as the *de facto* machine learning approach[16].

Deep learning has shown great promise in areas that rely on imaging data including radiology[17], pathology[18], dermatology[19], and ophthalmology[20] to name a few. In lieu of the often subjective visual assessment of images by trained clinicians, deep learning automatically identifies complex patterns in data and hence provides evaluations in a quantitative manner. Compared to feature engineering approaches, crafting and selecting the most robust features is inherent to deep learning networks and thus they require little to no human input. Deep learning methods have outperformed their engineered feature counterparts in many tasks including mammographic lesion detection[21], mortality prediction[22], and multimodal image registration[23].

Convolutional neural networks (CNN) are a class of deep learning models that combines imaging filters with artificial neural networks through a series of successive linear and nonlinear layers. CNN layers learn increasingly higher level features from images, eventually making predictions, essentially mapping image inputs to desired outputs. CNN's have demonstrated great potential in various classification[24], detection[25], segmentation[26], registration[27], and reconstruction[28] tasks - learning from photographic, pathology, and radiographic images[17]. Other efforts use pretrained networks on images from other domains, an approach known as transfer learning[29], as a workaround when sample size is perceived to be insufficient. In some instances, classifiers are built using a combination of deep learning and engineered features[30]. However, and with a few exceptions, most studies lack generalization power due to insufficient data - usually under 100 patients. With limited data and to avoid overfitting, most efforts have been confined to solving 2D problems or alternatively a 3D problem space is often treated as a composition of 2D orthogonal planes[31], with a few recent studies capitalizing on information within the entire 3D tumor volume[32]. No studies to date have explored medical-to-medical transfer learning, with learned representations usually being transferred from general imagery. Only a few studies have assessed the stability of deep learning features extracted from medical images, with most solely relying on the presumed robustness of CNN's in other application areas.

In this study, we investigated the ability of deep learning networks, 3D CNN's in particular, to quantify radiographic tumor characteristics and predict overall survival likelihood. We designed a rigorous analytical setup **(Figure 1)**, with seven large and independent datasets of 1194 NSCLC patients imaged with computed tomography (CT) across five institutions, to discover and validate the prognostic power of CNN's in patients treated with radiotherapy and surgery. The prognosis is formulated as a

binary two-year overall survival classification problem. We benchmarked the CNN's performance against models built on clinical parameters and engineered features, as well as demonstrated its stability in both test-retest and inter-reader variability scenarios. To gain a better understanding of the characteristics captured by CNN's, we maped salient regions in images as per their contributions to predictions, both within and beyond the tumor. Additionally, we aimed at assessing the driving biological pathways as a means to explore the biological basis of the captured phenotypes. Our results highlight the improved performance of deep learning networks over their engineered counterparts, their robustness against specific types of input variability, their perceived biological basis, and their ultimate potential in improving patient stratification.

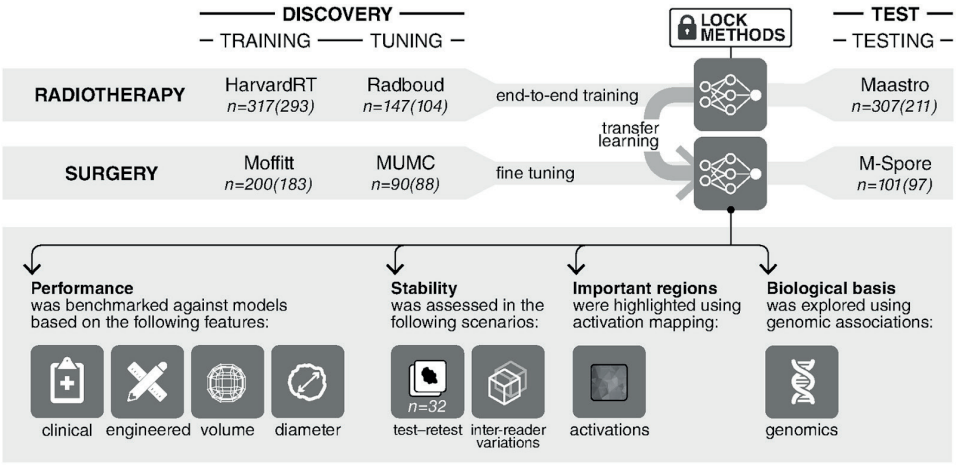

Figure 1. General design of the analytical setup. A 3D convolutional neural network is trained end-to-end on the radiotherapy dataset group. This is followed by a transfer learning approach where the same network is fine-tuned on the surgery dataset group. The training, tuning, and testing of these networks are all carried out on independent datasets as illustrated. Four further experiments are carried out on the networks in order to benchmark their performance against random forest models, assess their stability, identify regions in images responsible for predictions, and finally explore their biological basis. Number of patients outside parentheses refer to patients with survival follow-up per dataset. Numbers within parentheses refer to patients with 2 year overall survival follow-up only. Refer to **Methods** for patient censoring information and **S1 Table** for further dataset breakdown and information.

Methods

Datasets. We utilized seven independent datasets in this study **(S1 Table; S1 Text)** - divided into radiotherapy and surgery dataset groups, in addition to a stability assessment dataset. They come from a combination of European and US institutions as well as open-access online repositories.

Radiotherapy dataset group

- **HarvardRT** (training) consists of 317 NSCLC stages I–IIIb patients imaged with CT, with or without intravenous contrast, and treated with radiation therapy at the Dana-Farber Cancer Institute and Brigham and Women's Hospital, Boston, USA. Images were acquired between 2001 and 2015.

- **Radboud** (tuning) consists of 147 NSCLC stages I–IIIb patients imaged with contrast enhanced CT and treated with radiation therapy at Radboud University Nijmegen Medical Centre, The Netherlands. Images were acquired between February 2004 and October 2011.

- **Maastro** (testing) consists of 307 NSCLC stages I–IIIb patients, imaged with CT, with or without intravenous contrast, and treated with radiation therapy at MAASTRO Clinic, The Netherlands. Images were acquired between 2004 and 2010. This dataset is available at https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics

Surgical dataset group

- **Moffitt** (training) consists of 200 NSCLC stages I–IIIb patients imaged primarily (89%) with contrast-enhanced CT and treated with surgical dissection at the Thoracic Oncology Program at the H. Lee Moffitt Cancer Center, Tampa, Florida, USA. Images were acquired between 2006 and 2009.

- **MUMC** (tuning) consists of 90 NSCLC stages I–IIIb patients, imaged with CT, with or without intravenous contrast, and treated with surgical dissection at MAASTRO Clinic, The Netherlands. Images were acquired between 2004 and 2010. This dataset is available at https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics-Genomics

- **M-SPORE** (testing) consists of 101 NSCLC stages I–IIIb patients imaged with contrast-enhanced CT and treated with surgical dissection at the Thoracic Oncology Program at the H. Lee Moffitt Cancer Center, Tampa, Florida, USA. Images were acquired between 2006 and 2009.

Stability test dataset

- **RIDER** consists of 32 patients with NSCLC, each of whom underwent two CT scans of the chest within 15 minutes[33]. Images were acquired between January 2007 and September 2007. This dataset is available at https://wiki.cancerimagingarchive.net/display/Public/RIDER+Collections;jsessionid=C78203F71E49C7EA3A43E0D213CE5555

Overall survival times were calculated from the start of respective treatment for the radiotherapy and surgery datasets. These continuous survival times were dichotomized using a two-year cutoff. Datasets were then right-censored; alive patients at a last known follow-up of less than two years were excluded. This setup allows for a binary two-year survival endpoint of 0 for deceased patients and 1 for alive patients - relative to the two-year cutoff. To ensure non-bias dataset assignments of training, tuning, and testing, datasets with the most and least patients were assigned as training and tuning respectively. The remaining dataset was locked for testing. This assignment system was applied to both the radiotherapy and surgery dataset groups. Initial experiments were done on the radiotherapy datasets as they contained the most data, followed by transfer learning and fine tuning on the surgery datasets. This design also allows for averting noise as a result of large variability in tumor sizes between the two dataset groups, with the surgery group comprising consistently smaller tumors on average. All patients were utilized as per the survival data available without introducing artificial temporal cutoffs.

Data preprocessing. Tumors were manually contoured and approved by an expert reader (**S1 Text**). With slice thickness exceeding in-plane resolution, all datasets were resampled into isotropic voxels of unit dimension to ensure comparability where 1 voxel corresponds to $1mm^3$. This is achieved using linear and nearest neighbor interpolations for the image and annotations respectively. If multiple disconnected annotation masks were found, the largest by volume was chosen.

Data preprocessing for deep learning. Given full 3D tumor segmentations, both the center of mass (COM) and bounding box of the tumor annotations were calculated. 3D isotropic patches of size 50x50x50 were extracted around each COM capturing around 60% of the tumor bounding boxes' dimensions in the radiotherapy training dataset **(S1 Figure)**. The patches were then normalized to a 0-1 range using lower and upper Hounsfield units bounds of -1024 and 3071 respectively. An augmentation factor of 32k was applied to the patches yielding a training size of ~9.4M and ~5.9M for the radiotherapy and surgery datasets respectively. This included random translations ±10 pixels in all 3 axes, random rotation at 90° intervals along the longitudinal axes only, and random flipping along all 3 axes. Augmentation was done in real-time during training. No tuning- or testing-time augmentation has been applied.

Deep learning. We employed a 3D convolutional neural network (CNN) architecture **(Figure 2)**. The network comprises a total of 4 3D convolutional layers of 64, 128, 256, and 512 filters with kernel sizes of 5x5x5, 3x3x3, 3x3x3, and 3x3x3 respectively. 2 max pooling layers of kernel size 3x3x3 were applied after the 2nd and 4th convolutional

layers. A series of four fully-connected layers - with sizes 13824, 512, 256, and 2 - provide high level reasoning before the prediction probabilities were calculated in the final softmax classifier layer. Training details as follows: We used the gradient-based stochastic optimizer Adam[34] with a global learning rate of $1x10^{-03}$ without decay, a batch size of 16, dropout[35] of 25% and 50% on the convolutional and fully connected layers respectively, and a L2 regularization[36] penalty term of $1x10^{-05}$. To avoid the internal covariance shift problem[37], batch normalization was applied across all layers with the input layer as an exception. Leaky rectified linear units ReLU[38] with alpha=0.1 was the activation function of choice across the entire network prior to the final softmax activation. In training the CNN within the radiotherapy dataset, we used a random grid search exploring different hyper parameters including input patch size, batch size, learning rate, regularization term, and convolution kernel sizes. As for the general architecture, we started with a shallow network where underfitting occurs and incrementally added layers. The model was optimized on the tuning dataset using early stopping[39]. With a 1k epoch limit, the model with the best performance on the tuning dataset was chosen. In applying transfer learning on the surgery training dataset, the number of final layers to fine-tune was explored. The optimal setting included fine-tuning the final classification layer only, while keeping earlier layers fixed. With much fewer parameters to train, the learning rate and batch size were increased to $1x10^{-02}$ and 24 respectively. Google's deep learning framework TensorFlow[40] was used to train, tune, and test the CNN.

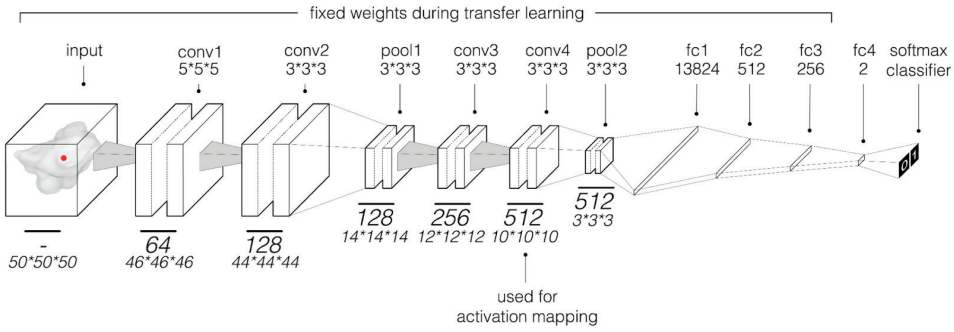

Figure 2. Illustration of the convolutional neural network. This network was used to predict overall two-year survival of NSCLC patients. The final classifier layer outputs normalized probabilities for both classes (0=deceased and 1=alive). Only the weights of the final fully connected layer were fine-tuned during transfer learning. The final convolutional layer (conv4) was used for activation mapping.

Data preprocessing for engineered feature extraction. Image intensity was binned by 25 HU to increase pattern sensitivity. Preprocessing filters were applied prior to feature extraction in order to reveal underlying information. Those included Laplacian of Gaussian, Wavelet, Square, Exponential, Square, Square Root, and Logarithm filters.

Engineered feature extraction and selection. Engineered features were computed using PyRadiomics[41], an open-source radiomics package. Feature stability was quantified using Intraclass Correlation (ICC) using the irr package[42], and the test-retest RIDER

dataset[33,43]. Features with an ICC>0.8 were regarded as highly robust and selected for the study. Supervised selection was done using the mRMR method (minimum redundancy maximum relevance) with the mRMRe package[44]. The mRMR was applied on the tuning datasets to select the top 40 engineered features with the highest mRMR ranks. Those features were then used for the final model on the training and testing datasets.

Machine learning on clinical parameters and engineered features. A random forest classifier was built using clinical parameters and engineered features. The tuning process involved a nested cross-validation technique (5 k-folds, 5 times) using the caret package[45] on the training dataset to select the best parameters such as the number of variables randomly sampled. The predictive power was measured on the testing dataset using the Area Under receiver operating characteristic Curve (AUC). Significance over random permutation was done using two-sided Wilcoxon rank-sum test between the score of the two classes.

Benchmarking. Benchmarking of deep learning networks against other models was done using a permutation test. AUC difference is defined as a Δ. For N permutations (N=1000 in our case), new models were built after randomly permuting class labels and new AUC's were computed from their respective scores. The new difference Δi was then converted to 0 if below Δ or 1 if above. Finally, the p-value was defined as:

$$p = \frac{1}{N}\sum_{i}^{N} \Delta i;$$

$$where\ \Delta i = 0\ if\ \Delta i < \Delta\ ,\ \Delta i = 1\ if\ \Delta i > \Delta$$

If the AUC difference between those two random models was higher than the true value, then the true class label was randomly permuted. A new model was then built and its score distribution was compared to the true distribution. Finally, a meta p-value was computed to compare the trend between the radiotherapy and surgery datasets (e.g. deep learning vs random forest models across two datasets) using the survcomp package[46].

Activation mapping. To generate activation maps, we used a gradient-weighted activation mapping method[47,48] to map important regions in an input image with respect to predictions made. The final convolutional layer **(Conv4 in Figure 2)** was set as the penultimate layer where the activation heatmaps (gradients) were generated during backpropagation. The heat maps were then thresholded at 0, normalized and enlarged to match the input image size. The heatmaps indicate regions in the input image contributing the most impact on the final prediction layer.

Masking experiment. Ground truth tumor annotations were used to delineate tumor areas and all voxels beyond the annotations were given the value of air (-1000 HU). The deep learning network was retained with the masked data while keeping all hyper parameters locked.

Genomic studies. We performed a pre-ranked Gene Set Enrichment Analysis (GSEA) as in previously published studies[14,49,50]. Briefly, more than 60,000 probes measured global gene expression on custom Affymetrix 2.0 microarray chipsets (HuRSTA_2a520709.CDF, GEO accession number GPL15048). Measured expression was normalized according to the robust multi-array average method[51]. These values were correlated with the network predictions to create a rank of all genes using Spearman rank correlation coefficient. This gene rank was input to a pre-ranked version of GSEA[52]. GSEA calculates scores that quantify the association of a given rank of genes with a pre-defined list of gene sets representing biological pathways. In such manner, GSEA allows for understanding what biological types of pathways the rank of genes corresponds to. As gene sets, we tested expert-curated pathways from the C2 Reactome collection version 6 available at MSigDB[53] using the GSEA version 3 with 1,000 permutations. Gene sets were restricted to sizes between 5 and 500, resulting in 669 tested gene sets. Expression data are publically available here https://elifesciences.org/articles/23421 & https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58661. We used GSEA's Normalized Enrichment Scores (NES) to quantify the association of the rank of genes with pathways and validated the NES with the false-discovery-rate (FDR) as per[54] to correct for multiple hypothesis testing.

Results

Tumor characterization using 3D deep learning networks. In assessing the ability of deep learning networks to quantify radiographic characteristics of tumors, we performed an integrative analysis on seven independent datasets totaling 1194 patients **(Figure 1; S1 Table)**. We identified and independently validated prognostic signatures using CNN's for patients treated with radiotherapy (n=771, including 608 with two-year survival follow-up). We then employed a transfer learning approach to achieve the same for surgery patients (n=391, including 368 with two-year survival follow-up). The architecture of the network **(Figure 2)** was designed to receive 3D input cubes surrounding the center of the primary tumor - based on clinician-located seed points. The network was trained to predict overall survival likelihood, two years after the start of the respective treatments.

Starting with the radiotherapy patients, the analysis was split into a discovery phase and an independent test phase **(Figure 1; S1 Table)**. Within the discovery phase, a 3D CNN was trained on the HarvardRT dataset (age median=69.6 (32.52 - 93.3), male/female=140/153, survival median=2.18 (0.0 - 11.68), 2-yr survival deceased/alive=134/159) using augmentation, while the independent Radboud dataset (age median=65.9 (44.38 - 85.93), male/female=n/a, survival median=0.9 (0.1 - 8.18), 2-yr survival deceased/alive=76/28) was used to iteratively tune and optimize the CNN's hyper parameters as well as the tumor 3D input patch sizes **(S1 Figure; Methods)** until the best prediction score was achieved. Beyond this discovery phase, the prognostic CNN was locked and tested on the independent Maastro dataset (age median=69 (34.0 - 91.7), male/female=142/69, survival median=1.04 (0.03 - 5.8), 2-yr survival deceased/alive=151/60). The CNN network showed a significant prognostic power in predicting

two-year survival (AUC=0.70 (0.63 - 0.78), p=1.13x10^{-07}) **(Figure 3A)**. Kaplan-Meier curve analysis was performed to evaluate the CNN's performance in stratifying low and high mortality risk groups. A significant survival difference (p=0.0001) was observed between the two groups on the independent Maastro dataset **(Figure 3B)**.

In order to develop a prognostic deep learning network for surgical patients, we employed a transfer learning approach **(Figure 1; S1 Table)**. The final prediction layers of the radiotherapy-trained CNN were fine-tuned on the Moffitt dataset (age median=n/a, male/female=83/100, survival median=2.83 (0.0 - 6.33), 2-yr survival deceased/alive=50/133) using augmentation **(Figure 2; Methods)**. The independent MUMC dataset (age median=68 (37.2 - 83.33), male/female=61/27, survival median=3.26 (0.24 - 8.78), 2-yr survival deceased/alive=24/64) was used to iteratively tune and optimize the CNN's hyper parameters as well as identify the optimum layers for fine-tuning. The CNN was then locked and tested on the independent test dataset M-SPORE (age median=70 (46.0 - 88.0), male/female=44/53, survival median=4.5 (0.33 - 7.83), 2-yr survival deceased/alive=17/80), where it demonstrated a significant prognostic performance (AUC=0.71 (0.60 - 0.82), p=3.02x10^{-04}) **(Figure 3C)**. Kaplan-Meier curve analysis showed significant survival difference (p=0.03) between low and high mortality risk groups within the M-SPORE test dataset **(Figure 3D)**.

Benchmarking against clinical parameters and engineered imaging features. The deep learning networks were benchmarked against random forest models based on clinical information (age, gender, and TNM stage). These clinical models achieved a performance of (AUC=0.55 (0.47 - 0.64), p=0.21) and (AUC=0.58 (0.39 - 0.77), p=0.4) for the radiotherapy and surgery datasets respectively. Additionally, univariate analysis suggested that these demographic and clinical variables did not have a significant association with survival **(S2 Table)**. Deep learning performed significantly better for both treatment types **(S2 Figure)**.

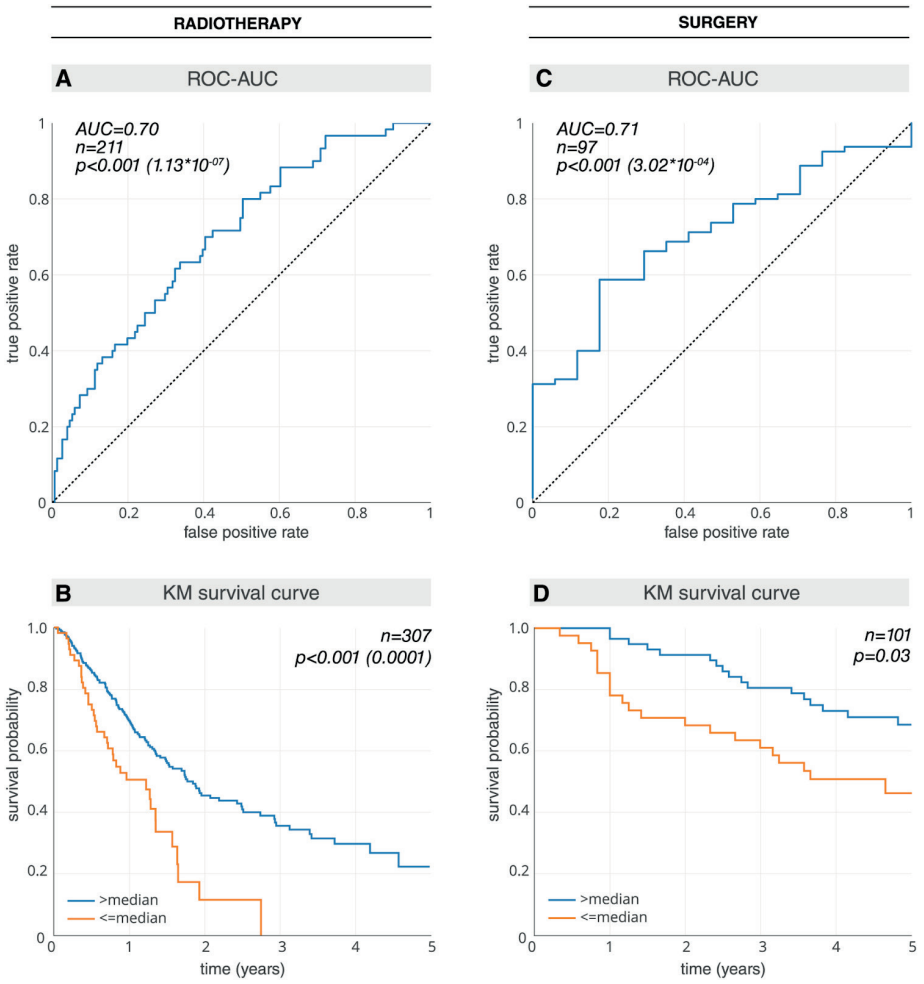

Figure 3. Prognostic power (AUC) and Kaplan-Meier (KM) curves of deep learning features for both the radiotherapy and surgical networks. **(A)** AUC plot for the radiotherapy test dataset Maastro (n=211). **(B)** KM plot for the Maastro dataset (n=307). Patients that have been previously excluded for lack of 2 year survival follow-up have been reincluded **(S1 Table)**. To ensure an independent evaluation, the median split is calculated on the radiotherapy tuning dataset Radboud (n=147) and locked for evaluation on the radiotherapy test dataset Maastro. **(C)** AUC plot for the surgery test dataset M-SPORE (n=97). **(D)** KM plot for the M-SPORE dataset (n=101). The median split is calculated on the surgery tuning dataset MUMC (n=90) and locked for evaluation on the surgery test dataset M-SPORE.

The deep learning networks were also compared to random forest models based on engineered features describing tumor shape, texture, and histogram. The engineered feature models demonstrated a prognostic performance of (AUC=0.66 (0.58 - 0.75), $p=1.91x10^{-04}$) and (AUC=0.58 (0.44 - 0.75), p=0.275) for the radiotherapy and surgery datasets respectively **(S2 Figure)**. Although the deep learning networks demonstrated improved performance over the engineered models for both patient groups, this difference

was not significant for radiotherapy patients (p=0.132; permutation test, N=1000), but was significant for surgery patients (p=0.035; permutation test, N=1000). These results were confirmed with a meta p-value test (p=0.06).

Finally, the deep learning networks were compared to imaging parameters commonly used in clinical practice, namely tumor volume and maximum diameter. We found that tumor volume achieved a performance of (AUC=0.64 (0.56 - 0.73), $p=6.18x10^{-04}$) and (AUC=0.51(0.37 - 0.66), p=0.85) for the radiotherapy and surgery datasets respectively. The deep learning networks were borderline non-significantly better on the radiotherapy dataset (p=0.056), and significantly better for the surgery datasets (p=0.004), as confirmed with a meta p-value test ($p=7.60x10^{-05}$). Similar results were found for maximum diameter **(S2 Figure)**.

Stability of deep learning networks. To evaluate the stability of the deep learning networks, we tested robustness against test-retest scenarios as well as variations in input seed annotations. We used the publicly available test-retest RIDER dataset comprising 32 patients with lung cancer, each of whom underwent two chest CT scans within 15 minutes by using the same imaging protocol and in a similar position[33]. Using this dataset, we evaluated the stability of network predictions between the test and retest scans. A high stability was demonstrated through the intraclass correlation coefficient (ICC) between both predictions (ICC=0.91).

To assess stability against variations in input data, we randomly relocated the input seed points in 3D-space around the center of the tumor **(S3 Figure)**. This randomly shifts the network inputs during testing and can be regarded as simulating multiple human readers annotating the tumor's center with the inevitable variability among them. The network outputs show high correlation (Spearman's Rank-Order Correlation=0.88). We also observed a high stability in prognostic predictions (AUC, μ=0.68, σ=0.014) **(S3 Figure)**.

Activation mapping of deep learning networks. To gain an understanding of regions within the CT images responsible for network predictions, we mapped the network's activation maps over the final convolutional layer **(Figure 4)**. The magnitudes of gradients flowing through this layer are used to decide on the "importance" of each node or voxel relative to the final prediction layer. This allowed us to highlight the most relevant regions with the most impact on predictions, both within and beyond the tumor. We observed that the network tended to fixate on the interface between the tumor and stroma (parenchyma or pleura). Most contributions to predictions came in the form of large uninterrupted areas of relatively higher CT density - spanning regions within and beyond the tumor. Areas with lower CT density, however, contributed the least to predictions. Examples of these include lobe areas with infrequent vessels or jagged interfaces between low and high CT density areas. We also observed that normal tissue, such as high density bone tissue, was disregarded - as it is likely to exist in most images and is thus non-informative. This visual mapping demonstrates that tissue within and beyond the tumor were both crucial for characterization and eventual prediction. In order to further validate these findings, we re-trained the deep learning network with

masked images - essentially discarding data beyond the tumor. A drop in prognostic power was observed (from AUC=0.70 to 0.63) **(S4 Figure)**, hinting at the existence of discriminative texture features in tumor-surrounding regions.

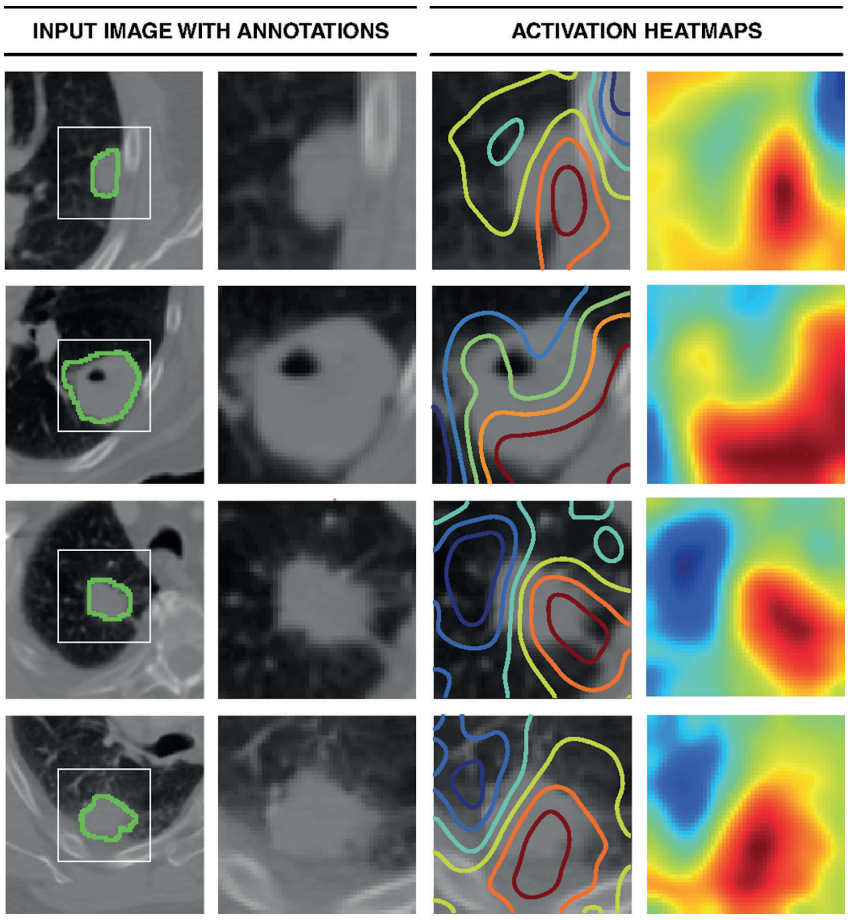

Figure 4. Activation mapping. Visual highlights of the most 'important' regions within the input image - those with the most contributions to maximizing the outputs of the final prediction layer. The rows represent four randomly selected samples. From the left, the first column represents the central axial slice of the network input (150x150mm) with tumor annotations. In the second column, a 50x50mm patch is cropped around the tumor. In the third column, activation contours are overlaid with blue and red showing the lowest and highest contributions (gradients) respectively. Column four represents the activation heatmaps for a better visual reference. While the heatmaps are 3 dimensional, only the central axial slice is shown. Therefore, the entire color spectrum might not be fully visualized.

Biological basis of deep learning networks. We also explored the biological basis of the radiographic phenotypes quantified by deep learning networks through investigating imaging and gene-expression assays on the surgery training dataset Moffitt (n=200). We

linked the CNN's predictions to global gene expression patterns using a pre-ranked gene set enrichment analysis (GSEA). Notably, the majority of the most significantly enriched pathways (false-discovery-rate, FDR ≤ 10^{-3}) are directly linked to cell cycle and transcriptional processes **(Figure 5; S1 File)**. For example, meiotic synapsis, telomere packaging, and various cell cycle stages such as G1 and S were among the top associations. Notably, these enrichments were highly negative - thus suggesting that the network predictions show inverse correlation to a proliferating phenotype. These results were consistent when reproduced on the surgery tuning dataset MUMC (n=90) **(S5 Figure; S2 File)**, where cell cycle and proliferation pathways, as well as various transcriptional processes were observed among the most significant associations.

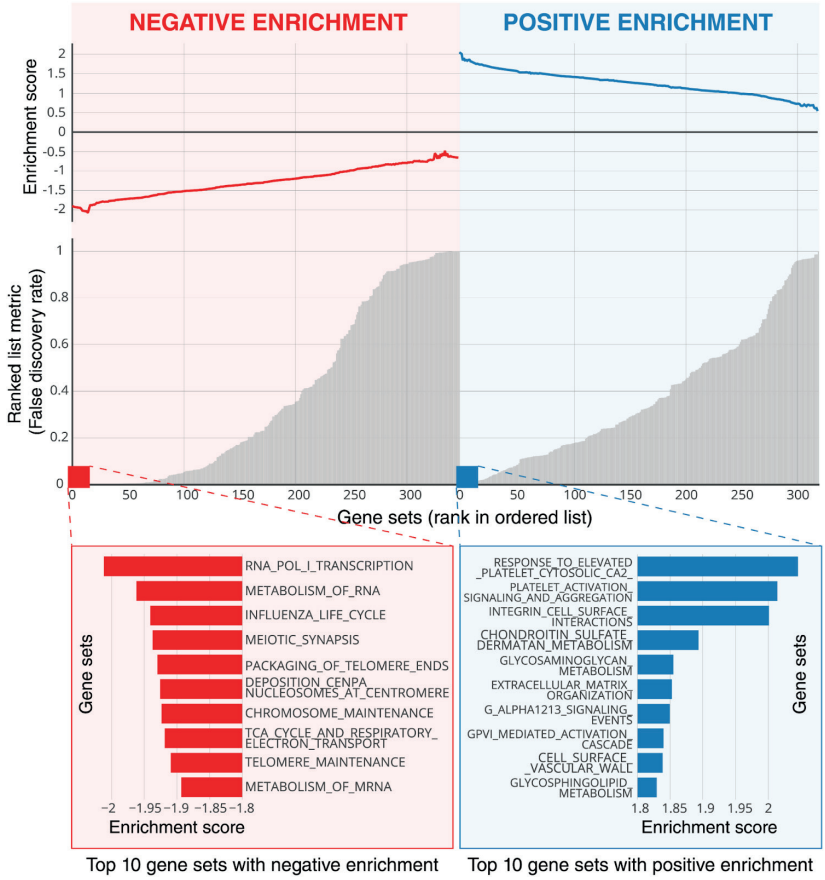

Figure 5. Global gene set expression patterns - Moffitt dataset. The deep learning network predictions on the surgery training dataset Moffitt were linked to global gene expression patterns using a pre-ranked gene set enrichment analysis (GSEA). Negative and positive enrichments are shown in red and blue respectively. The top ten enrichments in each category are highlighted. See **(S1 File)** for full ranking and associated enrichment scores.

Discussion

In this study, we assessed the utility of deep learning networks in predicting two-year overall survival of NSCLC patients from CT data. We trained a 3D CNN end-to-end on patients treated with radiotherapy, and employed a transfer learning approach for those treated with surgery. We demonstrated CNN's ability in significantly stratifying patients into low and high mortality risk groups, as well as their stability in test-retest and inter-reader variability scenarios. In addition to benchmarking against feature engineering methods, we also highlighted regions with the largest contributions to the captured prognostic signatures, both within and beyond the tumor volume. Finally, our preliminary genomic association studies suggested correlations between deep learning features and cell cycle and transcriptional processes.

This effort builds upon a body of deep learning applications in medical imaging that has emerged since the unprecedented superior performance of CNN's in recent image classification competitions[55]. Few deep learning studies to date have explored prognostication, with most addressing other tasks including segmentation, detection, and malignancy classification[17]. While feature definition is automated in these deep learning approaches, radiomics has primarily relied on the extraction, selection, and subsequent classification of predefined features using other machine learning methods including shallow neural networks, random forests, and support vector machines among others[15]. These methods have found applications in the prognostication of nasopharyngeal carcinoma in MRI[56], pulmonary adenocarcinoma in CT[57], and early-stage NSCLC in PET/CT[58] to name a few. Consequently, in this study, we benchmarked the deep learning networks against random forest models built on engineered features, with performance being within previously observed ranges[10]. These models exhibited an inferior performance when compared to the deep learning networks; although this difference was only significant for surgery patients. These results may be attributed to the higher levels of abstraction inherent in deep learning features over their engineered counterparts. Additionally, and in terms of input formats, engineered features were extracted exclusively from within tumor annotations. Deep learning inputs, however, were comprised of 3D cubes allowing the network to consider tumor-surrounding tissue. This effect is magnified in the smaller tumors treated with surgery relative to their larger radiotherapy counterparts, potentially explaining the significance of the surgery results. Surgery patients are often excluded from engineered radiomics studies[59–61] where no prognostic signal has been detected, citing as reasoning the lack of rationale in predicting a tumor response based on its phenotype if it is resected. This hints at the potential utility of deep learning networks in stratifying this specific patient group.

We also explored models built on a set of clinical features, including age, gender, and TNM stage. These models performed poorly in both the radiotherapy and surgery datasets, potentially attributed to the limited features available and common to all six datasets. Imaging features commonly used in the clinic, namely tumor volume and max diameter, performed relatively well on the radiotherapy datasets, but rather poorly on the surgery datasets which had previously been demonstrated[62]. Both models were outperformed by deep learning networks; although, this difference was only significant

for the surgery datasets. Further studies are needed to investigate the prognostic relationship between these features and deep learning features for radiotherapy patients, especially given the well established relationship between tumor volume and survival in this group[63]. These results also hints at the prognostic superiority of deep learning features for surgery patients.

Our efforts in identifying salient regions within images through activation mapping hint at the significance of tumor-surrounding tissue in patient stratification. This aligns with efforts that showcase the prognostic value of tumor location[64] as well as the importance of understanding the interactions between tumors and their surroundings as a means for effective cancer prevention and care[65].

Finally, our preliminary genomic association study showcases correlations between the deep learning network predictions and cell cycling, transcriptional, and other DNA-replication processes, such as DNA repair or damage response. This suggests that deep learning features may be driven by underlying molecular processes mostly related to proliferation of cells and hence progression of tumors. Moreover, nearly all significantly enriched biological processes had a negative enrichment score, indicating an inverse relationship to the survival predictions. This suggests that the gene expression present in cell proliferating pathways tend to be downregulated with higher network scores indicating a higher survival probability. As associations between engineered imaging features and biological pathways have already been established[14,66], our study extends these associations to deep learning.

Strengths of this study include the relatively large - in cancer imaging terms - set of 1194 NSCLC patients with training, tuning, and testing on independent datasets. The datasets were heterogeneous in terms of imaging acquisition parameters, clinical stage, and management, thus reflecting clinical reality. This suggests that deep learning methods may eventually be sufficiently robust and generalizable for practical application to clinical care. In addition to being a non-invasive and cost effective routine medical test[67,68], CT imaging provides a relatively stable radiodensity metric standardized across equipment vendors and imaging protocols compared to other imaging modalities (e.g. MRI or PET). In comparison to engineered radiomic methods that require slice-by-slice tumor annotations, a time consuming and expensive process that is highly prone to inter-reader variability, our approach may yield higher throughput as it only requires a single-click seed point placement roughly within the center of the tumor volume. The two-year survival endpoint utilized here is a relevant survival cutoff for NSCLC patients and one that has been previously used in prognostication efforts[69–71]. Our study hints at the utility of transfer learning within medical imaging and across treatment types, a finding that is also strengthened through benchmarking against end-to-end training of the surgery test dataset **(S6 Figure)**.

Several limitations should also be noted. By design, the retrospective nature of this study hindered the ability to gauge how and where such a tool can potentially be integrated into the clinical workflow. Consequently, the prognostic knowledge distilled into the deep learning networks is based on earlier treatment options and protocols, and may not

be adequately positioned to infer a prognostic signature of a patient treated with more modern means. The opaqueness of deep learning networks is another limitation. Feature definition, extraction, and selection in these methods - a major source of variability in engineered radiomics[15] - are all automated and occur implicitly. This comes at an expensive cost: interpretability. Consequently, these black box-like networks are very difficult to debug, isolate the reason behind certain outcomes, and predict when and where failures would happen. Without a strong theoretical backing[72], deep learning features are nameless and the imaging characteristics they measure are highly obscure. This ambiguity is in sharp contrast to the expert-based well-defined engineered features, and is often exacerbated in prognostication problems where the only means to validation is long term mortality follow up through prospective studies. Additionally, a better understanding of the network hyperparameter space is needed, potentially provided by using multiple tuning datasets within the discovery phase and prior to the final test phase. Another limitation lies in the input data space. Despite the aforementioned dataset heterogeneity, CT stability, as well as the test/retest and inter-reader variability studies performed herein, the networks' sensitivity to other variations in clinical parameters and image acquisition parameters including tube current, noise index levels, and reconstruction-specific parameters among others has not been explored. Finally, as survival times used in this study are overall as opposed to being cancer-specific, they may be influenced by external factors and introduce uncertainty into the problem.

Given the fixed input size of the deep learning networks used in this study, research implications include exploring classification network architectures that accept inputs of simultaneous multi-scale resolutions[73] or variable sizes[74] - an approach common to fully convolutional networks used in image segmentation. This can potentially allow for combining the large tumors in radiotherapy patients with their relatively smaller counterparts in surgery patients into one prognostic network whilst maintaining robustness against such variation. In terms of interpretability, training neural networks with disentangled hidden layer representations is an active area of research[75]. While our activation mapping studies offer a qualitative measure of network attention, a more quantitative visualization and diagnosis of network representations is needed, especially with applications in the medical space. Additionally, a safeguard against neural networks' blind-spots is required in addressing our weak understanding of their susceptibility to adversarial attacks[76], and more specifically the sensitivity of medical images to certain reported counter-intuitive properties of CNN's[77]. Finally, recent advances in Imaging-Genomics[78] motivate further explorations beyond our preliminary GSEA study. When rigorously evaluated in future prospective studies, deep learning-based prognostic signatures could highlight the specific biological states of tumorigenesis exhibited by a given patient, and thus enable more targeted therapy applications that exploit specific biological traits.

The development of prognostic biomarkers for NSCLC patients is an active area of research where tumor staging information is augmented with radiographic, genetic, molecular, and protein-based evidence[79,80]. The lack of a truly prognostic clinical gold standard hinders the ability to accurately benchmark these biomarkers and further stresses the need for prospective validation. While TNM staging is often utilized in the

clinic as the primary means for NSCLC prognostication and treatment selection, it is mainly intended as a discrete measure of tumor extent and a clinical communication tool, in addition to being simple and static by design. Conversely, quantitative imaging features inferred through deep learning are continuous, high-dimensional, and may be used to augment the higher-level coarser stratification provided by TNM staging. After considering the aforementioned limitations, a prognostic imaging tool may allow the transition to a finer classification enabling the identification of appropriate treatment plans on the individual patient level. One potential application for such transition may be in managing early stage NSCLC patients, for whom surgery represents a therapeutic mainstay albeit having high recurrence risks[7]. Adjuvant chemotherapy is often administered as a means of reducing these risks[81,82]. While T- and N-stage are known to be associated with recurrence in these patients[83], we find that patients with similar clinical characteristics can exhibit wide variations in incidence of recurrence[84] and survival[85]. A finer classification within the same stage may allow for identifying low and high mortality risk patients. Accordingly, low risk patients may be spared the adverse physical and mental effects as well as associated costs of adjuvant chemotherapy, and conversely more stringent post-treatment surveillance of those at high risk may be planned. Additionally, a more detailed stratification could potentially inform surgical approaches and techniques, empower high risk patients with the choice of adjuvant therapy modalities that best fit their desired lifestyles, as well as identify long term beneficiaries from such therapy[86].

Deep learning algorithms that learn from experience offer access to unprecedented states of intelligence that, in some cases, match human intelligence. Beyond imaging, deep learning's multimodal nature[87] promises the integration of multiple parallel streams of information spanning genomics, pathology, electronic health records, social media, and many others, into powerful integrated diagnostic systems[88]. Despite numerous roadblocks including the need for standardized data collection methods, evaluation criteria, prospective validation, and reporting protocols[89], the greatest anticipated clinical impact of these algorithms will be within precision medicine. This emerging approach allows for early diagnosis and customized patient-specific treatments thus delivering the appropriate medical care to the right patient at the right time[90]. While medical imaging has always provided an individual assessment of ailments, AI algorithms promise to accurately stratify patients based on imaging biomarkers and enable new research avenues for personalized healthcare.

Acknowledgements

Authors acknowledge financial support from the National Institute of Health (NIH-USA U24CA194354, and NIH-USA U01CA190234); https://grants.nih.gov/funding/index.htm. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supporting Information

https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002711#sec030

References

1. Torre, L. A. *et al.* Global cancer statistics, 2012. *CA Cancer J. Clin.* **65**, 87–108 (2015).
2. American Cancer Society: Cancer Facts and Figures 2017. Atlanta, Ga: American Cancer Society, 2017. *Am. Cancer Soc. 2014* https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2016/cancer-facts-and-figures-2016.pdf.
3. Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E. & Adjei, A. A. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin. Proc.* **83**, 584–594 (2008).
4. Sculier, J.-P. *et al.* The impact of additional prognostic factors on survival and their relationship with the anatomical extent of disease expressed by the 6th Edition of the TNM Classification of Malignant Tumors and the proposals for the 7th Edition. *J. Thorac. Oncol.* **3**, 457–466 (2008).
5. Amin, M. B. *et al.* The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more 'personalized' approach to cancer staging. *CA Cancer J. Clin.* **67**, 93–99 (2017).
6. Gospodarowicz, M. K. *et al.* The process for continuous improvement of the TNM classification. *Cancer* **100**, 1–5 (2004).
7. Uramoto, H. & Tanaka, F. Recurrence after surgery in patients with NSCLC. *Transl Lung Cancer Res* **3**, 242–249 (2014).
8. Mirsadraee, S., Oswal, D., Alizadeh, Y., Caulo, A. & van Beek, E., Jr. The 7th lung cancer TNM classification and staging system: Review of the changes and implications. *World J. Radiol.* **4**, 128–134 (2012).
9. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).
10. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).
11. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446 (2012).
12. Ganeshan, B. *et al.* Non–Small Cell Lung Cancer: Histopathologic Correlates for Texture Parameters at CT. *Radiology* **266**, 326–336 (2013).
13. Ganeshan, B., Abaleke, S., Young, R. C. D., Chatwin, C. R. & Miles, K. A. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. *Cancer Imaging* **10**, 137–143 (2010).
14. Grossmann, P. *et al.* Defining the biological basis of radiomic phenotypes in lung cancer. *Elife* **6**, (2017).
15. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci. Rep.* **5**, 13087 (2015).
16. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* (2018) doi:10.1038/s41568-018-0016-5.

17. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).

18. Cruz-Roa, A. *et al.* Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci. Rep.* **7**, 46450 (2017).

19. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).

20. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* (2016) doi:10.1001/jama.2016.17216.

21. Kooi, T. *et al.* Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303–312 (2017).

22. Carneiro, G., Oakden-Rayner, L., Bradley, A. P., Nascimento, J. & Palmer, L. Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography. in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* 130–134 (2017).

23. Yang, X., Kwitt, R., Styner, M. & Niethammer, M. Quicksilver: Fast predictive image registration - A deep learning approach. *Neuroimage* **158**, 378–396 (2017).

24. Pan, Y. *et al.* Brain tumor grading based on Neural Networks and Convolutional Neural Networks. in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 699–702 (2015).

25. Forsberg, D., Sjöblom, E. & Sunshine, J. L. Detection and Labeling of Vertebrae in MR Images Using Deep Learning with Clinical Annotations as Training Data. *J. Digit. Imaging* (2017) doi:10.1007/s10278-017-9945-x.

26. Ghafoorian, M. *et al.* Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities. *Sci. Rep.* **7**, 5110 (2017).

27. Miao, S., Wang, Z. J. & Liao, R. A CNN Regression Approach for Real-Time 2D/3D Registration. *IEEE Trans. Med. Imaging* **35**, 1352–1363 (2016).

28. Hammernik, K., Würfl, T., Pock, T. & Maier, A. A Deep Learning Architecture for Limited-Angle Computed Tomography Reconstruction. in *Bildverarbeitung für die Medizin 2017* 92–97 (Springer Berlin Heidelberg, 2017).

29. Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).

30. Paul, R. *et al.* Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma. *Tomography* **2**, 388–395 (2016).

31. Zheng, Y., Liu, D., Georgescu, B., Nguyen, H. & Comaniciu, D. 3D Deep Learning for Efficient and Robust Landmark Detection in Volumetric Data. in *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015* 565–572 (Springer, Cham, 2015).

32. Milletari, F., Navab, N. & Ahmadi, S. A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. in *2016 Fourth International Conference on 3D Vision (3DV)* 565–571 (2016).

33. Zhao, B. *et al.* Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non–Small Cell Lung Cancer. *Radiology* **252**, 263–272 (2009).

34. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).

35. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

36. Ng, A. Y. Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance. in *Proceedings of the Twenty-first International Conference on Machine Learning* 78– (ACM, 2004).

37. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv [cs.LG]* (2015).

38. Maas, A. L., Hannun, A. Y. & Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. in *Proc. ICML* vol. 30 (2013).

39. Prechelt, L. Early Stopping - But When? in *Neural Networks: Tricks of the Trade* (eds. Orr, G. B. & Müller, K.-R.) 55–69 (Springer Berlin Heidelberg, 1998).

40. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv [cs.DC]* (2016).

41. van Griethuysen, J. J. M. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **77**, e104–e107 (2017).

42. Gamer, M., Lemon, J., Fellows, I. & Singh, P. irr: Various coefficients of interrater reliability and agreement. *R package version 0. 84* **137**, (2012).

43. Oxnard, G. R. *et al.* Variability of Lung Tumor Measurements on Repeat Computed Tomography Scans Taken Within 15 Minutes. *J. Clin. Oncol.* **29**, 3114–3119 (2011).

44. De Jay, N. *et al.* mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics* **29**, 2365–2368 (2013).

45. Kuhn, M. Building Predictive Models inRUsing thecaretPackage. *J. Stat. Softw.* **28**, (2008).

46. Schröder, M. S., Culhane, A. C., Quackenbush, J. & Haibe-Kains, B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics* **27**, 3206–3208 (2011).

47. Kotikalapudi, R. *keras-vis*. (Github).

48. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. in *2017 IEEE International Conference on Computer Vision (ICCV)* 618–626 (2017).

49. Grossmann, P., Gutman, D. A., Dunn, W. D., Jr, Holder, C. A. & Aerts, H. J. W. L. Imaging-genomics reveals driving pathways of MRI derived volumetric tumor phenotype features in Glioblastoma. *BMC Cancer* **16**, 611 (2016).

50. El-Hachem, N. *et al.* Characterization of Conserved Toxicogenomic Responses in Chemically Exposed Hepatocytes across Species and Platforms. *Environ. Health Perspect.* **124**, 313–320 (2016).

51. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).

52. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

53. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

54. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

55. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).

56. Zhang, B. *et al.* Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. *Cancer Lett.* **403**, 21–27 (2017).

57. Kim, H. *et al.* The prognostic value of CT radiomic features for patients with pulmonary adenocarcinoma treated with EGFR tyrosine kinase inhibitors. *PLoS One* **12**, e0187500 (2017).

58. Oikonomou, A. *et al.* Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy. *Sci. Rep.* **8**, 4003 (2018).

59. Coroller, T. P. *et al.* CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother. Oncol.* **114**, 345–350 (2015).

60. Huynh, E. *et al.* CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. *Radiother. Oncol.* **120**, 258–266 (2016).

61. Lao, J. *et al.* A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Sci. Rep.* **7**, 10353 (2017).

62. Coroller, T. P. *et al.* Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiother. Oncol.* **119**, 480–486 (2016).

63. Zhang, J. *et al.* Relationship between tumor size and survival in non-small cell lung cancer (NSCLC): An analysis of the Surveillance, Epidemiology, and End Results (SEER) registry. *J. Clin. Orthod.* **30**, 7047–7047 (2012).

64. Shien, K., Toyooka, S., Soh, J., Yamamoto, H. & Miyoshi, S. Is tumor location an independent prognostic factor in locally advanced non-small cell lung cancer treated with trimodality therapy? *Journal of thoracic disease* vol. 9 E489–E491 (2017).

65. Egeblad, M., Nakasone, E. S. & Werb, Z. Tumors as organs: complex tissues that interface with the entire organism. *Dev. Cell* **18**, 884–901 (2010).

66. Ahrendt, S. A. *et al.* p53 mutations and survival in stage I non-small-cell lung cancer: results of a prospective study. *J. Natl. Cancer Inst.* **95**, 961–970 (2003).

67. OECD. Computed tomography (CT) scanners. *Health equipment* (2015) doi:10.1787/bedece12-en.

68. Statistics / Health care use / Computed tomography (CT) exams. doi:10.1787/3c994537-en.

69. Oberije, C. *et al.* A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: a step toward individualized care and shared decision making. *Radiother. Oncol.* **112**, 37–43 (2014).

70. Hoang, T., Xu, R., Schiller, J. H., Bonomi, P. & Johnson, D. H. Clinical model to predict survival in chemonaive patients with advanced non-small-cell lung cancer treated with third-generation chemotherapy regimens based on eastern cooperative oncology group data. *J. Clin. Oncol.* **23**, 175–183 (2005).

71. Cistaro, A. *et al.* Prediction of 2 years-survival in patients with stage I and II non-small cell lung cancer utilizing 18F-FDG PET/CT SUV quantifica. *Radiol. Oncol.* **47**, 219–223 (2013).

72. Shwartz-Ziv, R. & Tishby, N. Opening the Black Box of Deep Neural Networks via Information. *arXiv [cs.LG]* (2017).

73. Ghafoorian, M. *et al.* Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. *Neuroimage Clin* **14**, 391–399 (2017).

74. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3431–3440 (2015).

75. Zhang, Q.-S. & Zhu, S.-C. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* **19**, 27–39 (2018).

76. Yuan, X., He, P., Zhu, Q. & Li, X. Adversarial Examples: Attacks and Defenses for Deep Learning. *arXiv [cs.LG]* (2017).

77. Finlayson, S. G., Chung, H. W., Kohane, I. S. & Beam, A. L. Adversarial Attacks Against Medical Deep Learning Systems. *arXiv [cs.CR]* (2018).

78. Bai, H. X. *et al.* Imaging genomics in cancer research: limitations and promises. *Br. J. Radiol.* **89**, 20151030 (2016).

79. Burotto, M., Thomas, A., Subramaniam, D., Giaccone, G. & Rajan, A. Biomarkers in Early-Stage Non–Small-Cell Lung Cancer: Current Concepts and Future Directions. *J. Thorac. Oncol.* **9**, 1609–1617 (2014).

80. Thakur, M. K. & Gadgeel, S. M. Predictive and Prognostic Biomarkers in Non-Small Cell Lung Cancer. *Semin. Respir. Crit. Care Med.* **37**, 760–770 (2016).

81. Zappa, C. & Mousa, S. A. Non-small cell lung cancer: current treatment and future advances. *Transl Lung Cancer Res* **5**, 288–300 (2016).

82. Non-Small Cell Lung Cancer Collaborative Group. Chemotherapy In Non-Small Cell Lung Cancer: A Meta-Analysis Using Updated Data On Individual Patients From 52 Randomised Clinical Trials. *BMJ: British Medical Journal* **311**, 899–909 (1995).

83. Wang, X. *et al.* Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital H&E images. *Sci. Rep.* **7**, 13543 (2017).

84. Pepek, J. M. *et al.* How well does the new lung cancer staging system predict for local/regional recurrence after surgery?: A comparison of the TNM 6 and 7 systems. *J. Thorac. Oncol.* **6**, 757–761 (2011).

85. Wu, C.-F. *et al.* Recurrence Risk Factors Analysis for Stage I Non-small Cell Lung Cancer. *Medicine* **94**, e1337 (2015).

86. Arriagada, R. *et al.* Long-term results of the international adjuvant lung cancer trial evaluating adjuvant Cisplatin-based chemotherapy in resected lung cancer. *J. Clin. Oncol.* **28**, 35–42 (2010).

87. Ngiam, J. *et al.* Multimodal deep learning. in *Proceedings of the 28th international conference on machine learning (ICML-11)* 689–696 (2011).

88. Lundström, C. F., Gilmore, H. L. & Ros, P. R. Integrated Diagnostics: The Computational Revolution Catalyzing Cross-disciplinary Practices in Radiology, Pathology, and Genomics. *Radiology* **285**, 12–15 (2017).

89. Lambin, P. *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **14**, 749–762 (2017).

90. European Society of Radiology. Medical imaging in personalised medicine: a white paper of the research committee of the European Society of Radiology (ESR). *Insights Imaging* **2**, 621–630 (2011).

6

Chapter 6

Deep Learning-Based Computed Tomography Radiomics for Non-Small Cell Lung Cancer Histopathologic Classification

TL Chaunzwa, *A Hosny*, Y Xu, A Shafer, N Diao, M Lanuti, DC Christiani, RH Mak & HJWL Aerts

Nature Scientific Reports 2021

Abstract

Tumor histology is an important predictor of therapeutic response and outcomes in lung cancer. Tissue sampling for pathologist review is the most reliable method for histology classification, however, recent advances in deep learning for medical image analysis allude to the utility of radiologic data in further describing disease characteristics and for risk stratification. In this study, we propose a radiomics approach to predicting non-small cell lung cancer (NSCLC) tumor histology from non-invasive standard-of-care computed tomography (CT) data. We trained and validated convolutional neural networks (CNNs) on a dataset comprising 311 early-stage NSCLC patients receiving surgical treatment at Massachusetts General Hospital (MGH), with a focus on the two most common histological types: adenocarcinoma (ADC) and Squamous Cell Carcinoma (SCC). The CNNs were able to predict tumor histology with an AUC of 0.71($p = 0.018$). We also found that using machine learning classifiers such as k-nearest neighbors (kNN) and support vector machine (SVM) on CNN-derived quantitative radiomics features yielded comparable discriminative performance, with AUC of up to 0.71 ($p = 0.017$). Our best performing CNN functioned as a robust probabilistic classifier in heterogeneous test sets, with qualitatively interpretable visual explanations to its predictions. Deep learning based radiomics can identify histological phenotypes in lung cancer. It has the potential to augment existing approaches and serve as a corrective aid for diagnosticians.

Introduction

Lung cancer is the leading cause of cancer-related death[1]. It is a heterogeneous disease with many clinically important subtypes[2]. Among these, histologic phenotype is a particularly important predictor of response to therapy and overall clinical outcome[1,2]. More than 80% of all primary lung cancers are classified as non-small cell lung cancer (NSCLC) where the major histological types include adenocarcinoma (ADC), and squamous cell carcinoma (SCC); deriving from small and large airway epithelia respectively[1,2]. In routine clinical practice, manual tissue assessment using conventional light microscopy is the gold standard and most widely used approach for histological categorization. However, this relies on complex invasive techniques, and biopsy may fail to capture the complete disease morphological and phenotypic profile due to inter- and intra-tumor heterogeneity[3,4]. Moreover, of every tissue block sent for diagnosis, only 1 or 2 slides are assessed[5], hindering the pathologist's ability to understand and capture the entire tumor environment[6].The manual interpretation of tissue samples also introduces diagnostic uncertainty due, in part, to the pathologist's decision tree, which relies on binary features that are susceptible to observer bias[4]. In many clinical settings, this has promoted routine adoption of molecular testing to distinguish between morphologically similar lesions[3,4]. In addition to being an expensive approach, the integration of diagnostic molecular pathology into the traditional pathology workflow remains challenging due to the lack of adequate training and expertise[7,8].

Given the complexity of lung cancer classification and the limitations of current practices, there is a need for innovative clinical data assessment tools to help better describe disease characteristics and ascertain treatment planning and prognosis. The automated interpretation of pathology slides through computer-assisted diagnosis (CADx) has the potential to reduce reader variability, and is an area of active research[9]. Despite the emergence of CADx-friendly ecosystems alongside advances in the digitization of 2-dimensional pathology slides as well as 3-dimensional microscopy imaging[9], the invasive nature of biopsies may expose patients to significant clinical complications, in addition to its limited cost effectiveness[10]. Existing approaches do not take full advantage of the vast amounts of other clinical information available in modern clinical practice, including radiographic imaging. Non-invasive histopathologic classification using routinely acquired radiographic images may serve as a viable alternative to routine interventional tissue sampling and could have significant implications for diagnostic and treatment decisions.

Radiomics has emerged as a tool for quantifying the solid tumor phenotype through the extraction and mining of quantitative radiographic features[11]. There is a growing body of evidence pointing at the prognostic value of such features[4,12,13] as well as their utility in stratifying patients by tumor grade[14]. While radiomics has primarily relied on the explicit extraction of engineered or hand-crafted imaging features[13,15], more recent studies have shifted towards deep learning - convolutional neural networks (CNNs) specifically - where representative features are learned automatically from data[16] . This has fostered the construction of advanced multi-parametric algorithms for cognitive decision-making in many clinical settings[10]. The combination of such powerful computer vision methods

with routine medical imaging promises to improve decision-support for the pathologist and oncologist at low cost[12].

In this study, we leverage recent advances in radiomics and deep learning to develop models for enhancing pathologist and clinician accuracy and productivity within the setting of early-stage NSCLC. Building on data collected through the comprehensive Boston Lung Cancer Survival (BLCS) cohort, we created deep learning models that can act as non-invasive pathological biomarkers for NSCLC. We trained a CNN to stratify patients into 2 groups based on lung cancer histology. We also found that the CNN-derived CT-radiomics features represented distinct biologic and diagnostic patterns in this cohort, and were associated with underlying tumor microanatomy. This preliminary work has the potential to enhance the human-based decision tree for NSCLC histologic classification, and non-invasive elucidation of tumor biology using radiographic data.

Materials and Methods

Table 1. Patient Characteristics and Follow-up Summary

Characteristic	Value (n=311)
Length of Follow-up, Median, yr	3.9
2-year survival, No. (%)	268 (86.2)
Stage, No. (%)	
I	186 (59.8)
II	125 (40.2)

Data retrieval and selection

Our model building and validation dataset consisted of a sample of 311 BLCS patients with early-stage NSCLC receiving care at Massachusetts General Hospital (MGH) between 1999-2011 **(Table 1)**. Most patients underwent primary surgery for their disease. Approval was obtained from the Partners Institutional Review Board (IRB# 1999P004935). Pre-resection computed tomography (CT) imaging data was obtained for the patient series. In addition, overall and progression free survival, cancer staging, and histopathologic data corresponding to these patients was documented. All patients had clinical Stage I or Stage II NSCLC. Clinical pathology reports read at MGH were used as ground truth. Patients were categorized into three groups; ADC, SCC and an "Other" category that comprised all other NSCLC histological subtypes, including large cell and mixed histology, bronchoalveolar carcinoma, carcinoid, and cases with more than one primary tumor (**Figure S1** in Supplementary Material). Because oncogenic driver mutation status is not routinely collected for early-stage NSCLC patients at this site, and EGFR/KRAS testing has only been offered since 2008, only a small subset of 18 (5.8%) patients also had this information available, and no further analysis using this information was pursued.

Data was partitioned randomly in order to pick test samples that are representative of the dataset as a whole, with no statistically significant difference in characteristics between model fine-tuning and test sets (**Table 2**).

Table 2. Training and validation dataset characteristics

Characteristic	Tuning set	Test set	p
Histology			
All adeno and SCC[a] (n= 223)	n= 172 (77%)	n= 51(23%)	
Adenocarcinoma	*n= 120 (70%)*	*n= 35 (69%)*	*p= 0.892*[b]
Squamous Cell Carcinoma	*n= 52 (30%)*	*n= 16 (31%)*	*p= 0.892*[b]
Stage			
I (n= 129)	n= 102 (59.3%)	n= 27 (52.9%)	*p*= 0.417[b]
A	*n= 61 (35%)*	*n= 13 (25%)*	
B	41 (23%)	n= 14 (27%)	
II (n = 94)	n= 70 (40.7%)	n= 24 (47.1%)	*p*= 0.586[b]
A	*n= 21 (12%)*	n= 10 (20%)	
B	n= 49 (28%)	n= 14 (27%)	
Survival			
2-yr survival	*n*= 148 (86%)	*n*= 43 (84%)	*p*= 0.722[b]

Data presented as n, % of respective data set (training or validation) [a] total number of cases with either adenocarcinoma or squamous cell histology, n [b] p represents the significance of the difference between the two sets

Image preprocessing

Image pre-processing included manual tumor identification, isotropic rescaling, and density normalization of input CT data. Segmentation of tumor regions was performed using a single-click seed-point placement technique. Here, a seed-point is placed in the center of the tumor region using the open-source 3D Slicer software (version 4.5.0-1, https://www.slicer.org/), after assessment of transverse sections slice by slice. We then extract 3D volumes around the seed-points and from this, 2D input tiles measuring 50 mm x 50 mm (**Figure S2** in Supplementary Material). Isotropic rescaling was performed on the image data with a linear interpolator to minimize distortion, applying scaling factors that allow for a uniform spatial representation of 1 mm × 1 mm for each isotropic pixel. Density normalization was also performed with mean subtraction and linear transformation.

Classification with deep convolutional neural-networks

In this exploratory analysis, CNNs were used for feature extraction and ultimate image classification. To address the challenge presented by the scarcity of curated medical data as well as the heterogeneous CT data normally encountered in routine clinical practice, we used a transfer learning approach, where robust models that are effective at performing other computer vision tasks are fine tuned to perform visual recognition on our imaging data. The VGG-16 (Visual Geometry Group) neural network architecture[17] pre-trained on a large natural image dataset (ImageNet) was assessed. We evaluated the network with fine-tuning of the last convolutional, pooling, and fully connected layers. Hyperparameter optimization was explored iteratively. Inputs of the VGG-16 model

were 50mm x 50mm image patches. The model had three input channels, all of which were fed grayscale images (that is, model inputs are identical stacked images). Fine-tuning was performed over 100 epochs with a subset of patients that had either ADC or SCC histology for our primary model, *model A*, and with a mix of all 3 histology types (ADC, SCC, and "Other") for the secondary model, *model B* **(Figure 1)**. Accordingly, the final prediction (softmax) layer was changed to 2 for *model A,* and 3 for *model B* **(Figure 2)**. The predictive performance of the models was evaluated with the area under the receiver operator curve (AUC), and other performance metrics outlined in the model assessment section.

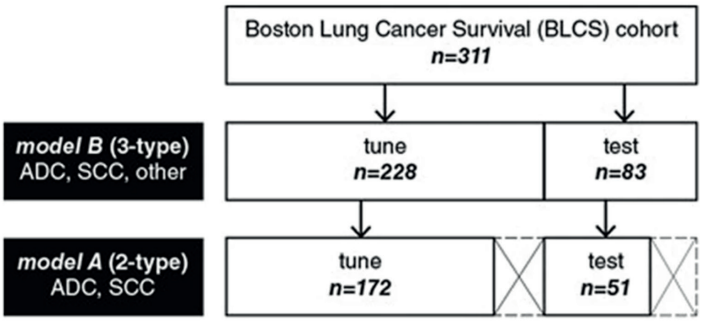

Figure 1. Dataset breakdown for model A and model B. Patients were categorized into three groups; ADC, SCC, and and "Other" category that comprised all other NSCLC histology subtypes. Similar to data presented in Table 2 for model A, model B was fine-tuned using the same BLCS dataset, but with the inclusion of all other histology types. This translated to a tuning-set with 120 ADC, 52 SCC, and 56 patients with "Other" histology types, and a test-set with 35 ADC, 16 SCC, and 32 patients with "Other" histology types.

Feature based analysis and classification

Many studies have shown that CNN-derived feature maps may outperform the original CNN in classification tasks when used with machine learning classifiers such as support vector machine (SVM) and random forest classifiers (RF)[18–20]. Unlike hand-crafted radiomics features, features from CNNs preserve global spatial information with the convolutional kernel operations on the input image (14). This gives them an advantage in fine-grained recognition, domain adaptation, contextual recognition as well as texture attribute recognition (14). CNNs are also less dependent on human curation which reduces bias. This provides rationale for an exploratory analysis using the "deep-radiomics" features from our models. For this, we generated features of the tumor regions as represented by the last pooling and the first fully connected layer of *model A*. The extracted descriptor feature vectors (512-D and 4096-D respectively) were normalized by subtracting the mean, and scaling to unit variance. This is essential to optimizing classification performance with discriminative machine learning classifiers, such as SVMs. Despite having flexible criteria, these methods may perform poorly if individual features deviate significantly from a normal distribution. In our data, individual features

appeared to follow Gaussian or Gaussian mixture distributions which validates this approach (**Figure S3** in Supplementary Material).

Compared to filtered feature reduction techniques which may eliminate important high order features, unsupervised feature reduction maintains the interaction among features, benefiting the model training process. Algorithms for unsupervised learning include principal component analysis (PCA) and auto-encoders, a generalized form of PCA. In our analysis, dimensionality reduction was performed using PCA to select independent features corresponding to a set threshold (>95%) of cumulative explained variance. The least absolute shrinkage and selection operator (LASSO) method was then used to select features that have the strongest association with the target types (shrinkage parameter, $\alpha = 0.01$). Four machine-learning classification models were independently evaluated on the extracted features: support vector machine (SVM) with both linear and non-linear kernels, k-nearest neighbors (kNN), as well as the random forest (RF) classifier[21,22].

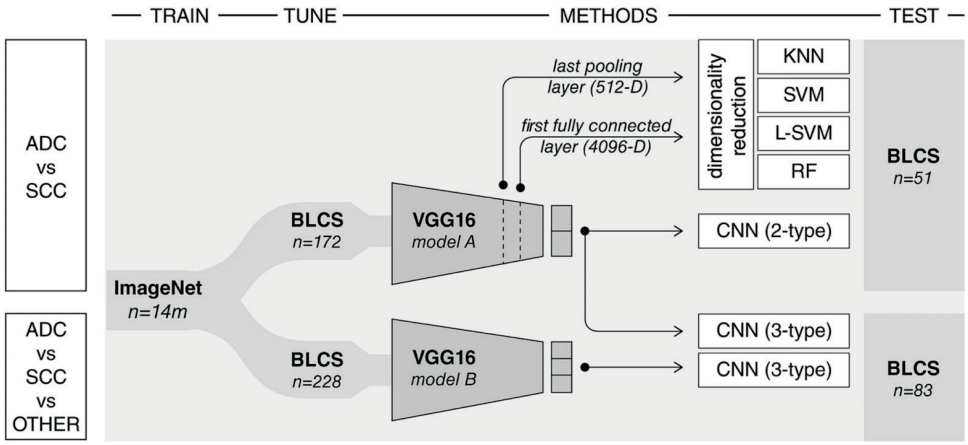

Figure 2. Experimental design. A convolutional neural network (VGG16) developed by the visual geometry group at Oxford and pre-trained on the large ImageNet dataset of more than 14 million hand-annotated natural images is employed in this analytical study. Model A is fine-tuned using a sample of 172 patients with either adenocarcinoma or squamous cell carcinoma and is used to predict future cases of these histology types using a held-out test set of 51 patients with adenocarcinoma or squamous cell carcinoma only. This model is also used as a fixed feature extractor for the assessment of machine learning classifiers (kNN, SVM, Linear-SVM, RF). These quantitative radiographic features are derived from the last pooling and first fully connected layers, corresponding to 512-D and 4096-D vectors, respectively. Model A is also used as a probabilistic classifier of histology and tested on a held-out test-set of 83 cases containing all histology types, grouped into adenocarcinoma, squamous cell carcinoma, or other. Model B is the fully connected VGG16 network tuned with a heterogenous sample of 228 cases with all histology types, and has as its output 3 different histology types, tested on the 83-patient sample as illustrated.

Model assessment

We assessed the discriminative power of *model A* in distinguishing the two most common histologies ADC vs SCC. Training for this and the feature based models was performed on the subset of patients with these histology types, translating to 172 for tuning and

51 for testing. Effects of hyper-parameter optimization e.g batch size were evaluated, as was the depth of fine tuning.

To assess the predictive performance of our models we used different descriptive indices including the area under the receiver operator curves (AUC), accuracy, sensitivity, and specificity. We also computed the Wilcoxon rank sum statistic for the binary predicted samples and a two-sided p-value of the test, with the assumption that these are samples from continuous distributions. Features or models with an AUC above 0.60 and a p-value below 0.05 are generally considered predictive in similar studies[23].

Neural network prediction probabilities and histological groups

In addition to noting *model A* performance in distinguishing ADC vs SCC, it may also be important to see how our CNN based biomarker performs on a dataset containing other histologies. For this we looked at a heterogeneous held-out test set of 83 patients containing ADC (n=48), SCC (n=18), and "Other" histologies (n =17). Using *model A* as a probabilistic classifier[24], the non-parametric Kruskal-Wallis H-test test was performed on the CNN-based prediction probabilities to assess the difference between the three independent samples of ADC, SCC, and "Other" on the test set. A p-value < 0.05 was considered as statistical significance. We also noted the model performance AUC and accuracy for the correct prediction of ADC in this heterogeneous data set (discriminative power).

For comparison, an identical network architecture, *model B* was fine-tuned using a non-overlapping composite dataset of 228 cases with all histologic types (ADC, SCC, Other). This seperate model was then tested on the same heterogeneous dataset of 83 patients. Given that three types exist for this model, micro-averaging of the predicted types was employed to binarize the ROC scores to either ADC vs all other histologies or SCC vs all other histologies.

Results

Clinical Characteristics

Our total patient cohort consisted of 311 patients diagnosed with early-stage NSCLC. A total of 186 (59.8%) patients had overall Stage I, and 125 (40.2%) had Stage II disease. Median follow-up from time of diagnosis was 3.5 years, with 86% 2-year survival. 155 (49.8%) patients had pathologist determined ADC, 68 (21.9%) of patients had SCC. The remaining 88 (28.3%) patients had all other histological subtypes, which included large cell and mixed histology, bronchoalveolar carcinoma, carcinoid, and cases with more than one primary tumor. Molecular testing for EGFR/KRAS mutation was done for 18 (5.8%) patients. Overall patient characteristics are summarized in **Table 1**. Model A fine-tuning and test cohort characteristics are summarized in **Table 2**.

Classification with CNNs

The VGG-16 based *model A* achieved significant predictive performance differentiating between ADC and SCC on a held-out test set of 51 patients with AUC of 0.71 (p =

0.018) **(Table 3)**. Similar fine-tuning and model evaluation was performed with another widely adopted ImageNet architecture, the ResNet50 network architecture [25]. There was no significant difference in its discriminative output and results from this analysis are included in the supplement.

Table3. Histology prediction probabilities for neural network vs PCA-derived feature-based classifiers

Method	AUC[a]	Accuracy	Specificity	Sensitivity	p
Fully Connected Neural Network Classifier					
VGG-16 (Model A)	0.709	68.6%	82.9%	37.5%	0.018
Machine learning classifiers on 512-D feature vectors					
kNN (k = 5) [b]	0.636	68.6%	77.1%	50%	0.123
Linear Support Vector Machine	0.616	70.6%	85.7	37.5%	0.187
Support Vector Machine	0.630	72.5%	88.6%	37.5%	0.138
Random Forest	0.613	72.5%	91.4%	31.3%	0.197
Machine learning classifiers on 4096-D feature vectors					
kNN (k = 5) [b]	0.71	76.5%	85.7%	56.3%	0.017
Linear Support Vector Machine	0.679	74.5%	85.7%	50%	0.042
Support Vector Machine	0.642	76.5%	97.1%	31.3%	0.107
Random Forest	0.571	66.7%	82.9%	31.3%	0.423

[a] Area under the ROC curve [b] k number of specified nearest neighbors, an even integer

Classification with CNN-derived features

With a threshold of 95% cumulative explained variance, PCA was able to perform dimensionality reduction of the 512-D and 4096-D feature space to 60 principal components. Feature selection with the LASSO (alpha = 0.01) yielded the 18 best performing features used in model building.

All models based on CNN-derived features were able to perform binary classification of tumor histology (ADC vs SCC). The 4096-D feature vector seemed to correlate with marginally better predictive performance with most machine learning classifiers, except with the RF classifier. The kNN model had the highest performance (AUC = 0.71, p = 0.017). This was on par with or better than the CNN (AUC = 0.71, p = 0.018). Other classifiers also showed significant predictive power, with an AUC of 0.68 (p = 0.042) for SVC with linear kernel (c = 0.1), AUC of 0.64 (p = 0.107) for non-linear SVC classifier. RF had the lowest predictive performance in all instances (AUC = 0.57, p =0.423), although this improved to an AUC of 0.61 (p = 0.197) with the 512-D feature vector. All models had higher specificity than sensitivity, while accuracy was again highest with the kNN model **(Table 3)**.

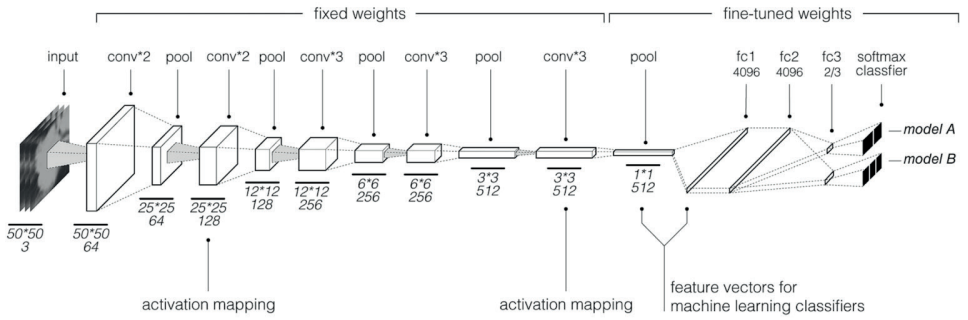

Figure 3. Model A and B schematic. This convolutional neural network architecture is based on the VGG architecture. With our transfer learning approach, weights of the last convolutional and pooling layers were fine-tuned using radiographic data. Model A, tuned on adenocarcinoma and squamous cell carcinoma tuning-set, had two classes as output in the softmax layer, while Model B which was tuned on a dataset containing all histology types had 3 type outputs.

Neural network prediction probabilities and histological groups

The 83 patient heterogeneous test set contained three histologic subgroups, ADC, SCC, and "Other". Looking at distributions of the prediction probabilities for each of these subgroups, based on our CNN biomarker, a statistically significant difference was noted for a comparison of all 3 groups (p=0.015). Post-hoc comparisons between groups showed that the difference was most pronounced between the ADC and SCC groups (p-value =0.003) **(Figure 4)**. There was a trend towards significance (p=0.235) between the predictions for the SCC and "Other" groups, however there was no statistically significant difference between the ADC and "Other" groups (p = 0.355). In keeping with the assumption that the test statistic H has a chi-square distribution, our sample sizes were all significantly greater than 5. Even in this heterogeneous test set, *model A* was still able to correctly predict ADC with an AUC of 0.66 (p = 0.013). The test specificity was 85% and sensitivity was 31% for ADC.

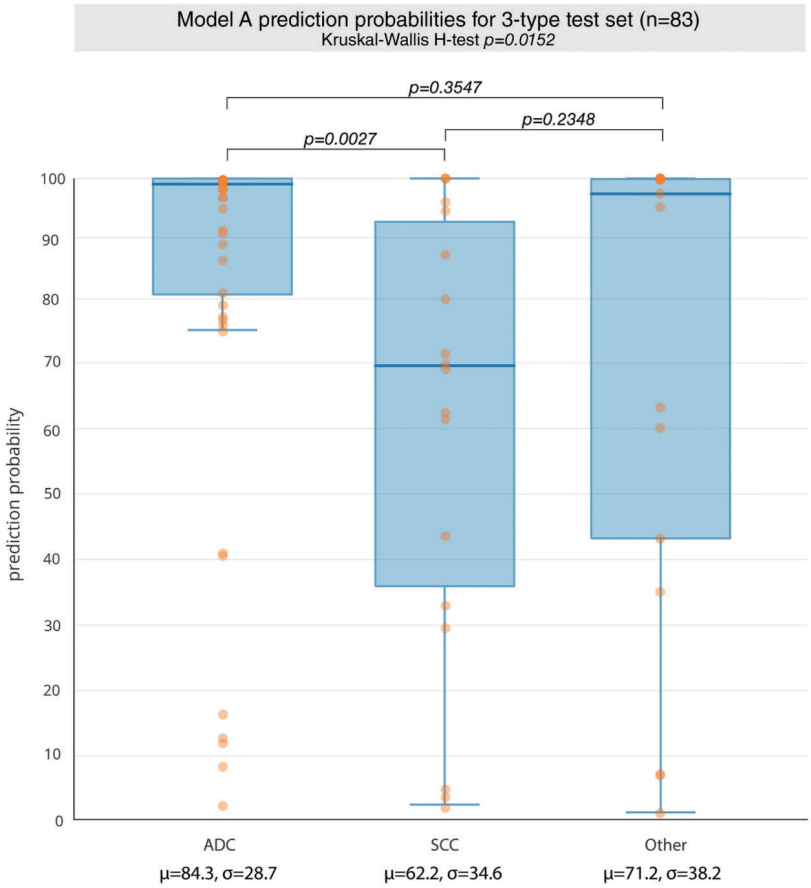

Figure 4. Model A as probabilistic classifier of non-small cell histology in 83 sample held-out test set containing all histology types. There is a statistically significant difference in predictions comparing all 3 histology groups: ADC, SCC, Other. Comparison of ADC vs SCC revealed a statistically significant difference with p-value of 0.003, while comparison of SCC vs Other had a p-value of p = 0.235, and ADC vs Other had a p-value of 0.355.

A separate analysis using an identical VGG network architecture, *model B* fine-tuned with a heterogeneous tuning set (n=228) containing all 3 histologic groups also had some predictive power when tested on the same 83 patient test set, albeit to a lesser extent. Using the ROC metric to evaluate classifier output quality for the 3-type model, ROC score when binarizing for SCC vs all other histologies was 0.62 (p = 0.127), and AUC = 0.58 (p=0.234) when binarizing ADC vs all other histologies. As such, the model trained on ADC and SCC alone outperformed one trained on all histologies in differentiating ADC histology from all other histology types (AUC = 0.66 compared to AUC =0.58).

Model Interpretability

Activation heat mapping was obtained using Gradient-weighted Class Activation Mapping[26]. Here we extracted heatmaps for all layers of our best performing model, *model A,* and selected representative examples **(Figure 5)**. This provided a spatial representation of areas within the input images that contribute the most to the model prediction. The first convolutional layers highlighted tumor edges. This is in line with what is observed when pre-trained models with similar architectures are applied to natural images, while deeper layers tend to pick up more abstract features, and in our experiment highlighted regions on or immediately around the tumor.

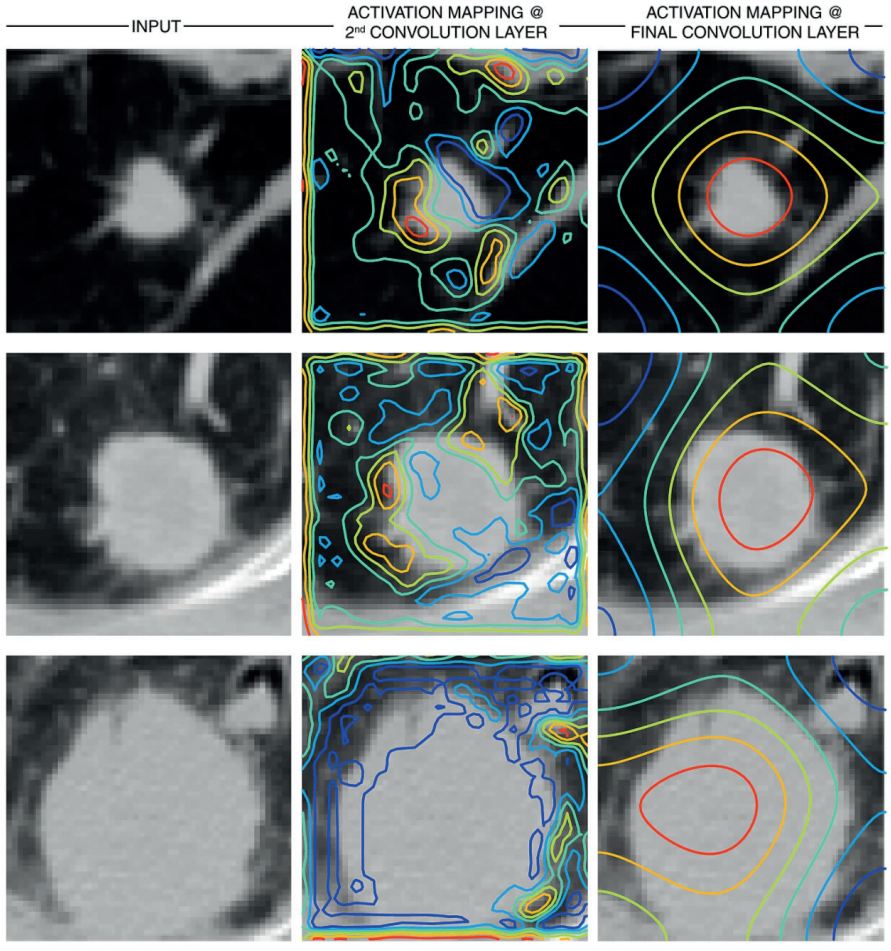

Figure 5. Gradient based class activation heat maps (Grad-CAM) for deep learning based model A. Visualization of image regions with the most discriminative value in type prediction as determined by the best performing convolutional neural network model. Here sample test input images are shown with overlaid activation contours, where red highlights regions with highest contribution and blue represents areas with the least value. The second and last convolutional layers in model A were used for generation of class activation maps as depicted by Figure 3.

Discussion

We investigated the utility of CNNs in predicting histology in early-stage NSCLC patients, using routinely acquired noninvasive radiologic images. We also assessed the association of CNN-derived quantitative radiographic image feature maps with histologic phenotype in this cohort. The goal of this work was to non-invasively predict cancer histology, and develop robust deep-learning based CT-radiomics models to aid pathologists in differentiating clinically important histologic subtypes in NSCLC.

We found that CNNs, which are effective at natural image recognition tasks, can be implemented to distinguish between the most common histopathologic subtypes in NSCLC. With enough labeled training data, they are able to detect subtle differences in the image features, which may not be apparent to the human observer, to aid in classification of clinically important groups, or predict the probability of certain phenotypes in future cases[10]. Using pre-trained models enabled us to build on previously learned low-/mid-level features in digital images (e.g., edges, shadows, texture etc). This reduces the likelihood of over-fitting, given the relatively large models , high dimensionality of features, and the limited size datasets. It also allows the model to more effectively decode heterogeneous image data as is commonly encountered in routine clinical practice, thereby aiding in the construction of models that are robust to these variations.

Our best performing model was able to detect adenocarcinoma with higher specificity than sensitivity, which could justify its potential role as a tool for computer assisted diagnosis. There is predictive and prognostic value in tumor histology, hence the ability to non-invasively predict this characteristic could provide significant cost and time saving benefits in addition to its ability to boost pathologist accuracy and productivity[10,12].

Prior studies have demonstrated the utility of CNNs as fixed feature extractors for image analysis and classification tasks, with many using the outputs from the last convolutional, pooling, or fully connected layers in VGG or related models[18–20,27]. We followed a similar approach in this work using the image feature representations from these layers in combination with various machine learning classifiers. These abstract high dimensional features are descriptive of the original image data with a great degree of redundancy. PCA was used for dimensionality reduction for its ability to preserve higher order features and their relationships, while eliminating a vast majority of the redundant features. Subsequently applying the LASSO led to the retention of only features with the strongest association with the predicted types[28,29]. Narrowing the dimensionality of the deep-radiomics feature space brings performance benefits and avoids overfitting. This was realized in this study with the kNN estimator which performed on par with the original neural network on the learned features, while other classifiers including SVM also showed significant predictive power with both feature sets. The findings suggest that dimensionality-reduction of CNN derived feature maps to summarize them with low-dimensional vectors, may serve as a robust multi-step alternative to fully-connected neural networks. This approach is in line with similar methods in the data science literature[18–20,30,31].

Both the 512-D and 4096-D feature space were successfully reduced to 18 best performing features. This suggests the same features were selected from both layers, which speaks to the reproducibility of the features. However, machine learning classifiers built around the 4096-D feature vector from the first fully connected layer seemed to correlate with marginally better predictive performance than from the 512-D feature vector. Neurons in a fully connected layer have full connections to all activations in the previous layer, whereas convolutional layers have connection to only the local features. This could help explain the marginally better performance with the fully connected layer **(FC1, Figure 3)**.

Looking at our CNN based biomarker as a probabilistic classifier of histology, we found that there is strong association between model prediction value and the likelihood of certain tumor phenotypes actually being present. That is, higher prediction certainty was associated with correct type prediction. For our analysis, because the histology group distribution was unbalanced, with more ADC than SCC and "Other", we favored using a group-based analysis of prediction probability distributions over directly assessing the association of certain types with percentiles of prediction probabilities. The ADC and SCC groups were found to have the most significant difference, which was expected, given our CNN biomarker was trained on distinguishing these two subtypes. No statistically significant difference existed between the ADC and "Other" groups, suggesting a significant overlap in radiographic phenotypes (or deep-radiomics features) in ADCs and the "Other" group. It is also in line with the widely reported misclassification of histologic subtypes in these broad umbrella groups, such as the notable misclassification of bronchoalveolar carcinoma (BAC) as adenocarcinoma or undifferentiated NSCLC or mixed phenotypes[32], with recent revised classification replacing the term BAC altogether[33]. As such, the "Other" group may contain a significant number of misclassified ADCs[2]. These findings not only demonstrate the validity of our CNN biomarker, but also open avenues for AI-enhanced methods to potentially drive paradigm shifts in histopathologic classification. In any case, adding these "Other" histologies to the test set introduced noise, reducing our model discriminative capacity. Furthermore, including "Other" histologies in the tuning cohort further reduces model performance, with the model trained on ADC and SCC alone outperforming one trained on all histologies in differentiating ADC histology from all others.

A well-recognized limitation of neural networks is their black-box nature. Looking at intermediate layers may help shed light into learned features, and further enhance the performance of our models. CNN interpretability is an area of increased investigation for the potential to not only help us understand how the models work, but also gain new insights into clinical data routinely encountered and be able to identify and predict failures. Here we found through gradient-based class activation heat mapping that our best performing model was activating on relevant image regions. In addition to the lesion of interest, our model also highlighted areas around the tumor, suggesting surrounding contextual information may also have predictive value. For lesions near the chest wall, the CNN appeared to still focus on the lesion and lung parenchyma, while placing less value on other structures including bone and soft tissue, which may

otherwise have similar CT density to tumor. This suggests an ability to learn complex and representative features. Overall these findings make intuitive sense, and importantly, provide reassurance that the model is detecting the right structures within our region of interest (ROI).

Access to the comprehensive BLCS cohort which has extensive clinical and biologic data was a unique strength of this study. Furthermore, our approach does not rely on accurate segmentation of tumor regions to work. This creates a less time intensive and more efficient work flow, whereas conventional approaches require precise tumor annotations, and are therefore more prone to human bias[34,35]. However, some limitations to the present study include small sample size and the lack of external validation. In addition, the interpretability exercise presented here is qualitative, and quantitative metrics may better validate future analyses.

The findings from this study provide a proof-of-concept that deep-learning based radiomics can identify histological phenotypes in lung cancer, and is a promising approach for non-invasive lung cancer histology classification. There is potential for clinical applicability of these models as decision support tools. While such methods are unlikely to replace the biopsy, they can help select those that may not require invasive diagnostic procedures. This can be achieved through radiomics' ability to stratify patients according to risk based on both abstract and well understood radiographic correlates. For example, while both SCC and ADC can cavitate, SCC cavitates more frequently. Small cell carcinoma, yet another important bronchogenic carcinoma, is never known to cavitate. Conversely, ADCs present with a characteristic ground glass appearance[36]. These well documented and easily identifiable distinctive imaging characteristics provide a firm theoretical basis for taking this approach a step further with radiomics. Similar studies have explored using CT texture analysis for histopathological grading in other disease sites including pancreatic ductal adenocarcinoma[37]. As such, deep-learning based radiomics has the potential to serve as both a decision-support tool and a corrective aid for the diagnostician.

Deep-learning based radiomics has the potential to create new paradigms in lung cancer risk assessment, by enabling us to transform the current rigid classification system into a more analytical and flexible model that includes radiological, biological, and other variables[11,13,15,37,38]. These methods can also potentially augment other emerging techniques, such as liquid biopsy; offering complementary information to guide clinical decision making. As molecular testing data becomes more widely available[39], future research may also help clarify the prognostic and predictive value of oncogenes such as KRAS in lung cancer. This additional information may also help establish stronger correlations between the deep learning based radiomics signatures and tumor biological data as it relates to histologically misclassified tumors.

Acknowledgements

The authors acknowledge support from the National Institutes of Health (NIH) with grant numbers (NIH-USA U24CA194354, NIH- USA U01CA190234, NIH-USA U01CA209414, and NIH-USA R35CA22052), the European Union—European Research Council (866504), and the Howard Hughes Medical Institute.

Supporting Information

https://www.nature.com/articles/s41598-021-84630-x#Sec17

References

1. Huang, T. *et al.* Distinguishing Lung Adenocarcinoma from Lung Squamous Cell Carcinoma by Two Hypomethylated and Three Hypermethylated Genes: A Meta-Analysis. *PLoS One* **11**, e0149088 (2016).
2. Davidson, M. R., Gazdar, A. F. & Clarke, B. E. The pivotal role of pathology in the management of lung cancer. *J. Thorac. Dis.* **5 Suppl 5**, S463–78 (2013).
3. Ilié, M. & Hofman, P. Pros: Can tissue biopsy be replaced by liquid biopsy? *Transl Lung Cancer Res* **5**, 420–423 (2016).
4. Zhao, B. *et al.* Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci. Rep.* **6**, 23428 (2016).
5. Kohl, S. K. *et al.* The College of American Pathologists and National Society for Histotechnology workload study. *Arch. Pathol. Lab. Med.* **135**, 728–736 (2011).
6. Sun, L., Wang, D., Zubovits, J. T., Yaffe, M. J. & Clarke, G. M. An improved processing method for breast whole-mount serial sections for three-dimensional histopathology imaging. *Am. J. Clin. Pathol.* **131**, 383–392 (2009).
7. Salto-Tellez, M., James, J. A. & Hamilton, P. W. Molecular pathology - the value of an integrative approach. *Mol. Oncol.* **8**, 1163–1168 (2014).
8. Fassan, M. Molecular Diagnostics in Pathology: Time for a Next-Generation Pathologist? *Arch. Pathol. Lab. Med.* **142**, 313–320 (2018).
9. Jansen, I. *et al.* Histopathology: ditch the slides, because digital and 3D are on show. *World J. Urol.* **36**, 549–555 (2018).
10. Djuric, U., Zadeh, G., Aldape, K. & Diamandis, P. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *NPJ Precis Oncol* **1**, 22 (2017).
11. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **278**, 563–577 (2016).
12. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).
13. Wu, W. *et al.* Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology. *Front. Oncol.* **6**, 71 (2016).
14. Ganeshan, B., Abaleke, S., Young, R. C. D., Chatwin, C. R. & Miles, K. A. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. *Cancer Imaging* **10**, 137–143 (2010).
15. Penzias, G. *et al.* Identifying the morphologic basis for radiomic features in distinguishing different Gleason grades of prostate cancer on MRI: Preliminary findings. *PLoS One* **13**, e0200730 (2018).
16. Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **15**, e1002711 (2018).
17. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv [cs.CV]* (2014).

18. Li, Z., Wang, Y., Yu, J., Guo, Y. & Cao, W. Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Sci. Rep.* **7**, 5467 (2017).

19. Notley, S. & Magdon-Ismail, M. Examining the Use of Neural Networks for Feature Extraction: A Comparative Analysis using Deep Learning, Support Vector Machines, and K-Nearest Neighbor Classifiers. *arXiv [cs.LG]* (2018).

20. Setiono, R. & Liu, H. Feature extraction via Neural networks. in *Feature Extraction, Construction and Selection: A Data Mining Perspective* (eds. Liu, H. & Motoda, H.) 191–204 (Springer US, 1998).

21. Hall, P., Park, B. U. & Samworth, R. J. Choice of neighbor order in nearest-neighbor classification. *Ann. Stat.* **36**, 2135–2152 (2008).

22. Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **46**, 175–185 (1992).

23. Coroller, T. P. *et al.* Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiother. Oncol.* **119**, 480–486 (2016).

24. Garg, A. & Roth, D. Understanding Probabilistic Classifiers. in *Machine Learning: ECML 2001* 179–191 (Springer Berlin Heidelberg, 2001).

25. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).

26. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. in *2017 IEEE International Conference on Computer Vision (ICCV)* 618–626 (2017).

27. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. in *Computer Vision – ECCV 2014* 818–833 (Springer International Publishing, 2014).

28. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer Science & Business Media, 2009).

29. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R*. (Springer Science & Business Media, 2013).

30. Mafarja, M. & Mirjalili, S. Whale optimization approaches for wrapper feature selection. *Appl. Soft Comput.* **62**, 441–453 (2018).

31. Islam, M. M. M., Islam, M. R. & Kim, J.-M. A Hybrid Feature Selection Scheme Based on Local Compactness and Global Separability for Improving Roller Bearing Diagnostic Performance. in *Artificial Life and Computational Intelligence* 180–192 (Springer International Publishing, 2017).

32. Raz, D. J. *et al.* Misclassification of bronchioloalveolar carcinoma with cytologic diagnosis of lung cancer. *J. Thorac. Oncol.* **1**, 943–948 (2006).

33. Gardiner, N., Jogai, S. & Wallis, A. The revised lung adenocarcinoma classification-an imaging guide. *J. Thorac. Dis.* **6**, S537–46 (2014).

34. Joskowicz, L., Cohen, D., Caplan, N. & Sosna, J. Automatic segmentation variability estimation with segmentation priors. *Med. Image Anal.* **50**, 54–64 (2018).

35. Zhao, B. *et al.* Exploring intra- and inter-reader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors on CT scans reconstructed at different slice intervals. *Eur. J. Radiol.* **82**, 959–968 (2013).

36. Austin, J. H. M. *et al.* Radiologic implications of the 2011 classification of adenocarcinoma of the lung. *Radiology* **266**, 62–71 (2013).

37. Qiu, W. *et al.* Pancreatic Ductal Adenocarcinoma: Machine Learning–Based Quantitative Computed Tomography Texture Analysis For Prediction Of Histopathological Grade. *CMAR* **11**, 9253–9264 (2019).

38. Coroller, T. P. *et al.* Radiographic prediction of meningioma grade by semantic and radiomic features. *PLoS One* **12**, e0187908 (2017).

39. Lindeman, N. I. *et al.* Updated Molecular Testing Guideline for the Selection of Lung Cancer Patients for Treatment With Targeted Tyrosine Kinase Inhibitors: Guideline From the College of American Pathologists, the International Association for the Study of Lung Cancer, and the Association for Molecular Pathology. *Archives of Pathology & Laboratory Medicine* vol. 142 321–346 (2018).

7

Chapter 7

Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging

Y Xu, *A Hosny*, R Zeleznik, C Parmar, T Coroller, I Franko, RH Mak & HJWL Aerts

Clinical Cancer Research 2019

Abstract

Purpose: Tumors are continuously evolving biological systems, and medical imaging is uniquely positioned to monitor changes throughout treatment. While qualitatively tracking lesions over space and time may be trivial, the development of clinically-relevant, automated radiomics methods that incorporate serial imaging data is far more challenging. In this study, we evaluated deep-learning networks for predicting clinical outcomes through analyzing time-series CT-images of locally advanced non-small cell lung cancer (NSCLC) patients.

Experimental Design: Dataset-A consists of 179 stage-III NSCLC patients treated with definitive chemoradiation, with pre- and post-treatment CT-images at 1, 3, and 6 months follow-up (581 scans). Models were developed using transfer-learning of convolutional neural-networks(CNNs) with recurrent-networks(RNN), using single seed-point tumor-localization. Pathologic-response validation was performed on Dataset-B, comprising 89 NSCLC patients treated with chemoradiation and surgery (178 scans).

Results: Deep-learning models using time-series scans were significantly predictive of survival and cancer-specific outcomes (progression, distant metastases and local-regional recurrence). Model performance was enhanced with each additional follow-up scan into the CNN model (e.g. 2-year overall-survival: AUC=0.74,p<0.05). The models stratified patients into low and high mortality risk-groups, which were significantly associated with overall-survival (HR=6.16, 95%CI[2.17,17.44],p<0.001). The model also significantly predicted pathological-response in Dataset B (p=0.016).

Conclusion: We demonstrate that deep-learning can integrate imaging-scans at multiple time-points to improve clinical outcome predictions. AI-based non-invasive radiomics biomarkers can have a significant impact in the clinic, given their low cost and minimal requirements for human input.

Translational Relevance

Medical imaging provides non-invasive means for tracking patients' tumor response and progression after treatment. However, quantitative assessment through manual measurements is tedious, time consuming, and prone to inter-operator variability as visual evaluation can be non-objective and biased.Artificial intelligence (AI) can perform automated quantification of radiographic characteristics of tumor phenotypes as well as monitor changes in tumors, before, during, and after treatment in a quantitative manner. In this study, we demonstrated the ability of deep learning networks to predict prognostic endpoints of patients treated with radiation therapy using serial CT imaging routinely obtained during follow-up. We also highlight their potential in accounting for and utilizing the available serial images to extract the relevant time-point and image features pertinent to the prediction of survival and response to treatment. This provides further insight into applications including the detection of gross residual disease without surgical intervention, as well as other personalized medicine practices.

Introduction

Lung cancer is one of the most common cancers worldwide and the highest contributor to cancer death in both the developed and developing worlds[1]. Among these patients, most are diagnosed with non-small-cell lung cancer (NSCLC) and have a five-year survival rate of only 18%[1,2]. Despite recent advancements in medicine spurring a large increase in overall cancer survival rates, this improvement is less consequential in lung cancer, as most symptomatic and diagnosed patients have late stage disease[3]. These late stage lesions are often treated with non-surgical approaches, including radiation, chemotherapy, targeted, or immunotherapies. This signals the dire need for monitoring therapy response using follow up imaging and tracking radiographic changes of tumors over time[4]. Clinical response assessment criterias, such as the response evaluation criteria in solid tumors (RECIST)[5], analyse time series data using simple size based measures such as axial diameter of lesions.

Artificial Intelligence (AI), allows for a quantitative, instead of a qualitative, assessment of radiographic tumor characteristics, a process also referred to as 'radiomics'[6]. Indeed, several studies have demonstrated the ability to non-invasively describe tumor phenotypes with more predictive power than routine clinical measures[7–10]. Traditional machine learning techniques involved the derivation of engineered-features for quantitative description of images with success in detecting biomarkers for response assessment and clinical outcome prediction[11–15]. Recent advancements in deep learning[6], have demonstrated successful applications in image analysis without human feature definition[16]. The use of convolutional neural networks (CNN) allows for the automated extraction of imaging features and identification of non-linear relationships in complex data. CNN networks which have been trained on millions of photographic images can be applied to medical images through transfer learning[17]. This has been demonstrated in cancer research with regards to tumor detection and staging[18]. AI developments can be clinically applicable to enhance patient care by providing accurate and efficient decision support[6,11].

The majority of quantitative imaging studies have focused on the development of imaging biomarkers for a single time-point[19,20]. However, the tumor is a dynamic biological system with vascular and stem cell contributions, which may respond, thus the phenotype may not be completely captured at a single time-point[21,22]. It may be beneficial to incorporate post-treatment CT scans from routine clinical follow-up as a means to tracking changes in phenotypic characteristics after radiation therapy. State of the art deep learning methods in video classification and natural language processing have utilized recurrent neural networks (RNN) to incorporate longitudinal data[23]. However, only a few studies have applied these advanced computational approaches in radiology[24].

In this study, we use AI in the form of deep learning, specifically CNNs and RNNs, to predict survival and other clinical endpoints of NSCLC patients by incorporating pre-treatment and follow-up CT images. Two datasets were analyzed containing patients with similar diagnosis of stage III lung cancer, but treated with different therapy regimens. In the first dataset, we developed and evaluated deep learning models in patients treated

with definitive chemoradiation therapy. The generalizability and further pathologic validation of the network was evaluated on a second dataset comprising patients treated with chemoradiation followed by surgery. For localization of the tumors, only single-click seed points were needed without volumetric segmentations, demonstrating the ease of incorporating a large number of scans at several time points into deep learning analyses. The CT imaging-based patient survival predictions can be applied to response assessment in clinical trials, precision medicine practices, and tailored clinical therapy. This work has implications for the use of AI-based imaging biomarkers in the clinic, as they can be applied noninvasively, repeatedly, at low cost, and requiring minimal human input.

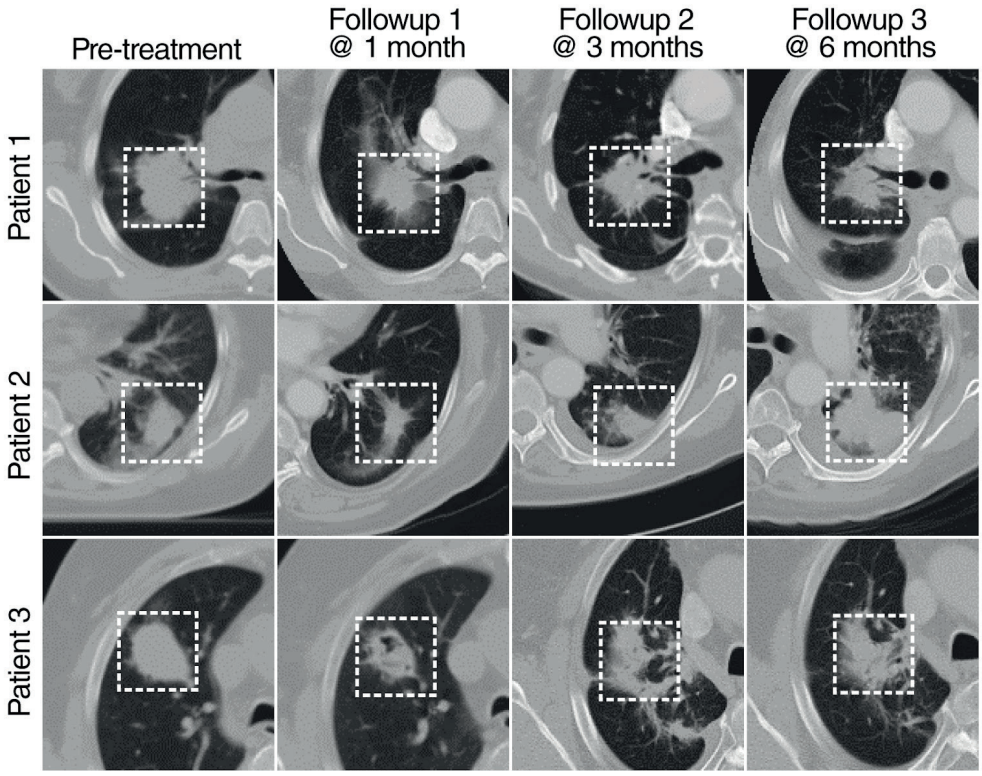

Figure 1. Serial Patient Scans. Representative computed tomography (CT) images of stage III non-surgical NSCLC patients before radiation therapy and one, three, and six months post radiation therapy. A single click seed point identifies the input image patch (defined by the dotted white line) that was inputted into the neural network.

Materials and Methods

Patient cohorts. We used two independent cohorts, Dataset A and Dataset B consisting of a total of 268 stage III NSCLC patients for this analysis. Dataset A contained 179 consecutive patients who were treated at Brigham and Women's/Dana-Farber Cancer Center between 2003 and 2014 with definitive radiation therapy and chemotherapy with Carboplatin/Paclitaxel (Taxol) or Cisplatin/Etoposide (chemoRT) and had at least one follow-up CT scan. We analyzed a total of 581 CT scans (average of 3.2; range 2-4 scans per patient, 125 attenuation CTs from PET and 456 diagnostic CTs) of pre-treatment and follow-up scans at 1, 3 and 6 months after radiation therapy for delta analysis of the serial scans **(Figure 1)**. The CT-PET scans were acquired without iodinated contrast, and the contrast administration of chest CT scans are patient specific and based on clinical guidelines. As a realistic representation of clinical settings, not all patients received imaging scans at all time points **(Figure S1)**. Patients with surgery prior to or after therapy were not included in this study. The main endpoint of this study was the prediction of survival and prognostic factors for stage III patients treated with definitive radiation **(Figure 2)**. Dataset A was randomly split 2:1 into training/tuning (n=107) and test (n=72). Overall, survival was assessed along with three other clinical endpoints for the definitive radiation therapy cohort: distant metastases, locoregional recurrence, and progression.

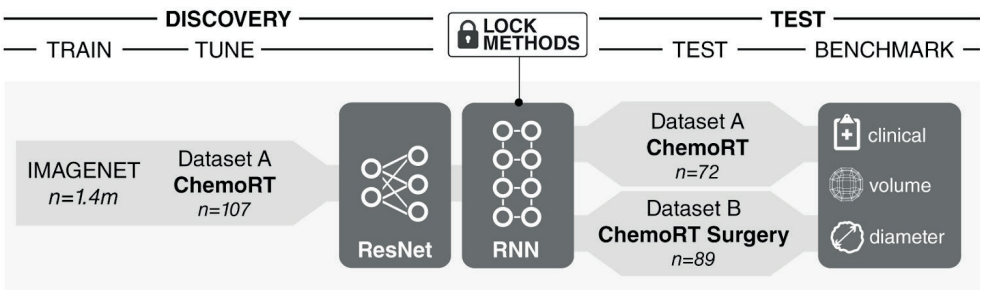

Figure 2. Analysis Design. Depiction of the deep learning based workflow with two datasets and additional comparative models. Dataset A included patients treated with chemotherapy and definitive radiation therapy, and was used to train and fine tune a ResNet convolutional neural network (CNN) combined with a recurrent neural network (RNN) for predictions of survival. A separate test set from this cohort was used to assess performance and compared with the performance of radiographic and clinical features. Dataset B included patients treated with chemotherapy and surgery. This cohort was used as an additional test set to predict pathological response, and the model predictions were compared to the change in volume.

An additional test was performed on Dataset B, a cohort of 89 consecutive, stage III NSCLC patients from our institution between 2001 and 2013, who were treated with neoadjuvant radiotherapy and chemotherapy prior to surgical resection (trimodality). The analysis of Dataset B was included for further validation with a range of standard of care treatment protocols. A total of 178 CT scans with two time points; scans taken

prior to radiation therapy and the scans after radiation were used, both taken prior to surgery. Patient exclusion included those who presented with distant metastasis or those with more than a 120 day delay between chemoradiation and surgery, as well as those without survival data. For both cohorts, no histological exclusions were applied. The endpoint of the additional test set of trimodality patients was the prediction of pathological response, validated at the time of surgery. The residual tumor was classified as responders (pathological complete response n = 14, and microscopic residual disease n = 28) or gross residual disease (n = 47) based on surgical pathological reports.

CT acquisition and image preprocessing. CTs were acquired according to standardized scanning protocols at our institution, using a GE "Lightspeed" CT scanner (GE Medical System, Milwaukee, WI, USA) for treatment, pre-treatment, and follow-up scans. The follow-up scans consisted of different axial spacing and a portion of the images are from PET-CT acquisitions. The input of the tumor image region is defined at the center of the identified seed point for the pre-treatment, and for the one, three, and six-month follow-up CT scans after definitive radiation therapy. The seed points were manually defined in 3D Slicer 4.8.1[25]. Due to the variability in slice thicknesses and in-plane resolution, the CT voxels were interpolated to $1 \times 1 \times 1$ mm^3 using linear and nearest neighbor interpolation. In order to have a stable input for the proposed architecture, it was necessary to interpolate the imaging data to homogeneous resolution. This was performed as the slice thicknesses were a maximum of 5mm and thus the 2D input images are taken at a slice not further than 2mm away from a non-interpolated slice. The linear interpolation was used to avoid potential perturbations from more complex interpolation methods which involves and may be dependent on several parameters and longer computation time. The fine scale was chosen to maintain the details of the tumor.

The three axial slices of 50×50 mm^2 centered on, 5 mm proximal to and 5 mm distal to the selected seed point were inputs to the model. 5 mm was the maximum slice thickness of the CT images. A transfer learning approach was applied using the pretrained ResNet CNN that was trained on natural RGB images. The three axial slices were used as input to the CNN network. Using three 2D slices gives the network information to learn from but keeps the number of features lower than a full 3D approach, as well as reduces GPU memory usage and training time as well as limits the overfitting. Image augmentation was performed on the training data, and involved image flipping, translation, rotation, and deformation, which is a conventional good practice and has shown to improve performance[26]. The same augmentation was performed on the pre-treatment and followup images, such that the network generates a mapping for the entire input series of images. The deformation was on the order of millimeters and did not noticeably change the morphology of the tumor or surrounding tissues. *Neural network structure.* The network structure was implemented in Python, using Keras with Tensorflow backend (Python 2.7, Keras 2.0.8, Tensorflow 1.3.0). The proposed network structure has a base ResNet convolutional neural network (CNN) trained on the ImageNet database containing over 14 million natural images **(Figure 3)**. One CNN was defined for each time point input, such that an input with scans at three time points would involve input into three CNNs. The output of the pretrained network model was then input into recurrent layers with gated recurrent units (GRU), which

takes the time domain into account. To ensure the network was able to handle missing scans[27,28], RNN algorithms were used which allowed for amalgamation of several time points and the ability to learn from samples with missed patient scans at a certain time points. The output of the pretrained network was masked to skip the time point when a scan was not available.Averaging and fully connected layers are then applied after the GRU with batch normalization[29] and dropout[30] after each fully connected layer to prevent overfitting. The final softmax layer allows for a binary classification output. To test a model without the input of follow-up scans the pre-treatment image alone was input into the proposed model, with the recurrent and average pooling layers replaced by a fully connected layer, as there was only one input time point.

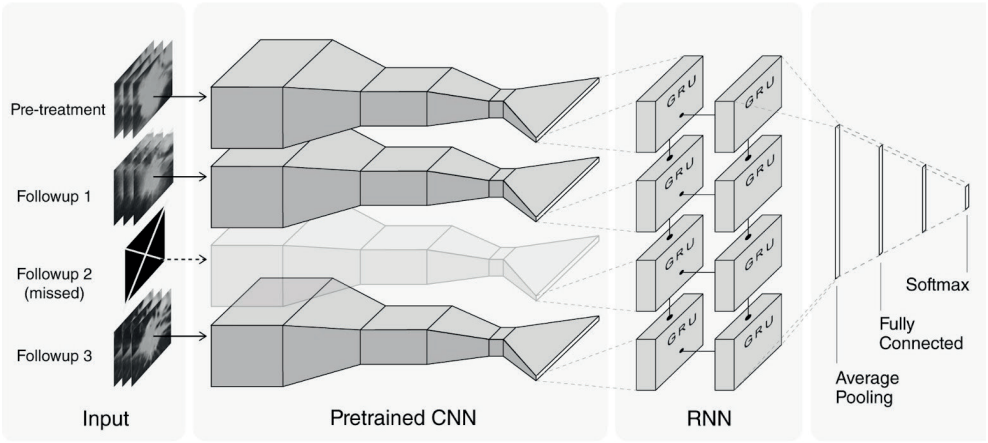

Figure 3. Deep Learning Architectures. The neural architecture includes ResNet convolutional neural networks (CNN) merged with a recurrent neural network (RNN), and was trained on baseline and follow-up scans. The input axial slices of 50 × 50 mm^2 centered on, 5 mm proximal to and 5 mm distal to the selected seed point. Deep learning networks are trained on natural RGB images, and thus need 3 image slices for input. The outputs of each CNN model are input into the RNN, with a gated recurrent unit (GRU) for time-varying inputs. Masking was performed on certain inputs of the CNN so that the recurrent network takes missed scans into account. The final softmax layer provides the prediction.

Transfer learning. Weights trained with ImageNet, a set of 14 million 2D color images, were used for the ResNet[31] CNN and the additional weights following the CNN were randomized at initialization for transfer learning. Dataset A was randomly split 2:1 into training/tuning and test.Training was performed with Monte Carlo cross validation, using 10 different splits (further 3:2 split of training: tuning) on 107 patients with class weight balancing for up to 300 epochs. The model was evaluated on an independent test set of 72 patients, who were not used in the training process. The surviving fractions for training/tuning (n=107) and test sets (n=72) were comparable **(Table S1)**. Only the pre-treatment image was input into the proposed model, and the recurrent and average pooling layers were replaced with a fully connected layer.

Statistical analysis. Statistical analyses were performed in Python version 2.7. All predictions were evaluated on the independent test set of Dataset A for survival and for prognostic factors after definitive radiation therapy. The clinical endpoints included distant metastasis, progression and locoregional recurrence as well as overall survival for one and two years following radiation therapy. The analyses were compared to a random forest clinical model with features of stage, gender, age, tumor grade, performance, smoking status and clinical tumor size (primary maximum axial diameter).

Statistical differences between positive and negative survival groups in Dataset A were assessed using the area under the receiver operator characteristic curve (AUC), and the Wilcoxon rank sums test (also known as the Mann–Whitney U test). Prognostic and survival estimates were calculated using the Kaplan-Meier method between low and high mortality risk groups, stratified at the median prediction probability of the training set and controlled using a Log-Rank test. Hazard ratios were calculated through the Cox Proportional-Hazards Model.

An additional test was performed on Dataset B, the trimodality cohort using the one-year survival model from the definitive radiation cohort with two time points. Survival predictions were made from the one-year survival model trained on Dataset A. The model predictions were used to stratify the trimodality patients based on survival and tumor response to radiation therapy prior to surgery. The groups were assessed using their respective AUC, and were tested with the Wilcoxon rank sums test. This was compared to the volume change after radiation therapy and a random forest clinical model with the same features used for Dataset A.

Results

Clinical characteristics. To evaluate the value of deep learning based biomarkers to predict overall survival using patient images prior and post radiation therapy **(Figure 1)**, a total of 268 stage III NSCLC patients with 739 CT scans were analyzed **(Figure 2)**. Dataset A consisted of 179 patients treated with definitive radiation therapy and was used as a cohort to train and test deep learning biomarkers **(Table S2)**. There was no significant difference between the patient parameters in the training and test sets of Dataset A (p>0.1, group summary values in Table S2). The patients were 52.8% females (median age of 63 years; age range 32 to 93 years) and were predominantly diagnosed as having stage IIIA (58.9%) NSCLC at the time of diagnosis, with 58.1% in the adenocarcinoma histology category. The median radiation dose was 66 Gy for the definitive radiation cohort (range 45 to 70 Gy, median follow-up of 31.4 months). Another cohort of 89 patients treated with trimodality served as an external test set (Dataset B). The median radiation dose for the trimodality patients was lower, at 54 Gy (range 50 to 70 Gy, median follow-up of 37.1 months).

Deep Learning based prognostic biomarker development and evaluation. To develop deep learning based biomarkers for overall survival, distant metastasis, disease progression, and locoregional recurrence, training was performed using the discovery part of Dataset

A **(Figure 2)**. To leverage the information from millions of photographic images, the ResNet CNN model was pre-trained on ImageNet and then applied to our dataset using transfer learning. The CNN extracted features of the CT images of each time point were fed into a recurrent network for longitudinal analysis. We observed that the baseline model with only pre-treatment scans demonstrated low performance for predicting two-year overall survival (AUC=0.58, p=0.3, Wilcoxon's test). Improved performance to predict two-year overall survival was observed with the addition of each follow-up scan; at 1 month (AUC=0.64, p=0.04), 3 months (AUC=0.69, p=0.007), and 6 months (AUC=0.74, p=0.001) **(Figure S2)**. We also observed the similar trend in performance for other clinical endpoints i.e. one-year, survival, metastasis, progression, and locoregional recurrence-free survival **(Figure S3)**. A clinical model, incorporating stage, gender, age, tumor grade, performance, smoking status and clinical tumor size, did not yield a statistically significant prediction of survival (two-year survival AUC = 0.51, p=0.93) or treatment response **(Table S3)**.

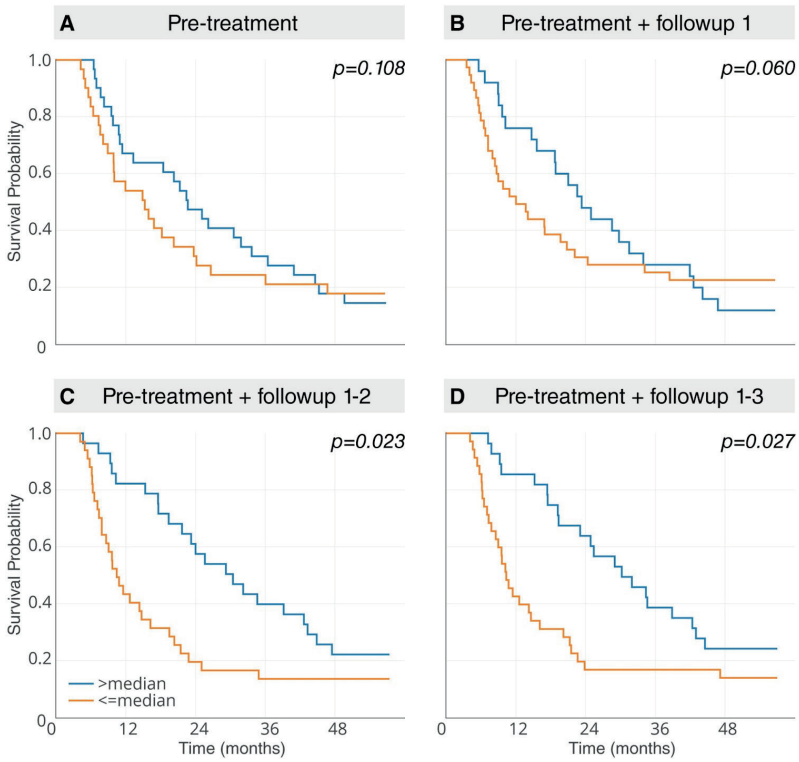

Figure 4. Performance Deep Learning Biomarkers on Validation Datasets. The deep learning models were evaluated on an independent test set for performance. The two-year overall survival Kaplan Meier curves were performed with median stratification (derived from the training set) of the low and high mortality risk groups with no follow-up, or up to three follow-ups at one, three and six months post treatment for Dataset A (72 definitive patients in the independent test set, log-rank test $p < 0.05$ for > one follow-up).

Further survival analyses were performed with Kaplan-Meier estimates for low and high mortality risk groups based on median stratification of patient prediction scores **(Figure 4)**. The models for two-year overall survival yielded significant differences between the groups with 2 (p=0.023, Log-Rank test) and 3 (p=0.027, Log-Rank test) follow-up scans. Comparable results were found for the following predictions with their respective Hazard ratios: one-year overall survival (6.16, 95% CI [2.17,17.44] p=0.0004), distant metastasis free (3.99, 95% CI [1.31,12.13] p=0.01), progression free (3.20, 95% CI [1.16,8.87] p=0.02) and no locoregional recurrence (2.74, 95% CI [1.18,6.34] p=0.02), each with significant differences at 3 follow-up time point scans.

Predicting Pathologic response. As an additional independent validation and to evaluate the relationship between delta imaging analysis and pathological response, the trimodality pre-radiation therapy and post-radiation therapy prior to surgery scans were input into the neural network model trained on dataset A. First for survival prediction evaluation, the model was tested on Dataset B. To match the number of input time points, the one-year survival model with the pre-treatment and first follow-up at one month was used. The model significantly predicted distant metastasis, progression, and local regional recurrence **(Table S4)**. Although, for overall survival there were a low number of events (30 of 89), the model was trending towards making a prediction for three-year overall survival in Dataset B.

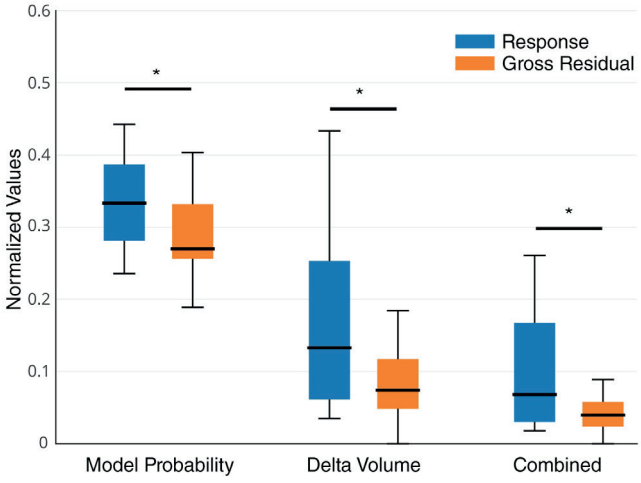

Figure 5. Pathological Response Prediction Validation. Model probability and the change in volume after radiation therapy was used for the prediction of pathologic response. The CNN survival model significantly stratified response and gross residual disease in the second test set Dataset B, comparable predictions were found with change in tumor volume, and the combination of the two parameters (n=89, Wilcoxon p < 0.05).

The predictions of the network were then used to categorize pathological response **(Figure 5)**, and were found to significantly distinguish between responders and gross residual disease, with an AUC of 0.65 (n=89, p=0.016, Wilcoxon's test), which was similar to the change in volume (AUC of 0.65; n=89; p=0.017, Wilcoxon's test). In order to investigate the additive performance, we built a combined model of the network probabilities and change in volume, which showed slightly higher performance (AUC of 0.67; n=89; p=0.006, Wilcoxon's test). The CNN probabilities and changes in the primary tumor volume were significantly correlated (p=0.0002), although with a Spearman's correlation value of 0.39. A clinical model, involving parameters of stage, gender, age, tumor grade, performance, smoking status and clinical tumor size, did not yield a statistically significant prediction for pathological response (p=0.42, Wilcoxon's test).

Discussion

Tracking tumor evolution for prediction of survival and response after chemotherapy and radiation therapy can be critical to treatment assessment and adaptive treatment planning for improving patient outcomes. Conventionally, clinical parameters are used to determine treatment type and to predict outcome[2], but this does not take into account phenotypic changes in the tumor. Medical imaging tracks this evolution of lesions non-invasively and provides a method for tracking the same region longitudinally through time, providing additional tumor characteristics beyond those obtained through static images at a single time point[5]. Follow-up CT scans are already a part of the clinical workflow, providing additional information regarding the patient. Using deep-learning approaches for tumor assessment allows for the extraction of phenotypic changes without manual and/or semi-automated contours or qualitative visual interpretations, which are prone to inter-observer variability. Additionally, prognostic predictions can potentially aid in the assessment of patient outcome in clinical trials to assess response and eventually dynamically adapting therapy.

Using a combined image-based CNN and a time encompassing RNN, the neural network was able to make survival and prognostic predictions at one year and two years for overall survival. As expected, with an increase in the number of time points and the amount of imaging data available to the network, there was an increase in performance. Although the performance varied between the predictions, there was a consistent increase in AUC, due to the increase in signal from each additional image of the primary tumor and the changes between the scans with time. In this cohort, using a single pre-treatment scan was not successful in making a prediction of survival. However, previous work in the field of radiomics using engineered[9,12,14,15] and deep learning[10] approaches using pretreatment imaging data only, were able to predict the endpoint of their interest with the use of anatomical CT or functional PET data. For the cohorts in this study, there is a trend towards significance of the deep learning model with the pre-treatment time point only. Using larger cohorts could improve the predictive power of the imaging markers. The clinical model, which included the clinical tumor size (longest axial diameter), was also not predictive of survival or the other prognostic factors.

The neural network was able to stratify patients into low and high mortality risk groups, with significant difference in overall survival **(Figure 4)**. This was also identified for the risk of locoregional recurrence with the input of two follow-up time points at around one and three months after the completion of definitive radiation therapy. The other outcomes, progression and distant metastasis needed the additional third follow-up at around 6 months for a significant stratification of the mortality risk groups. This may be due to a more defined set of early imaging phenotypes relating to survival and locoregional recurrence as compared to the other prognostic factors, or confounding phenotypes with regards to distant metastasis and progression, which the model cannot overcome unless the third follow-up is incorporated.

The two datasets within our study are inherently different as the cohorts are comprised of patients with different disease burdens and treatment modalities. The surgical patients are younger and healthier on average, with an earlier stage of disease, and well enough to tolerate surgery. It has been shown that the survival of surgical patients is dependent on the success of the surgical procedure and distant disease[32], where definitive RT survival is determined by local control[33]. There was also a higher proportion of stage IIIA in patients who also underwent surgical resection (Dataset B) compared to definitive RT patients (Dataset A).

Despite these differences, the survival CNN models trained on Dataset A predicted surrogates of survival in Dataset B including distant metastasis, progression, and locoregional recurrence. It was trending towards predicting survival and this may be due to the inherent differences between the cohorts, as well as the low number of events in the cohort and sample size. There was also only one follow-up scan available for Dataset B, thus less information was provided to the survival model. Although the model was designed to overcome the immortal time bias, there could still be an effect. With more time points, fewer patients are alive to have the scan performed and thus decrease the ability to predict survival.

Survival is associated with tumor pathological response[34,35]. Thus, we tested the relationship between the probabilities of the survival network model on similar stage III NSCLC patients who were in different treatment cohorts (definite radiation therapy and trimodality). Dataset B included the follow-up time point after radiation therapy and prior to surgery, for the prediction of response and for further validation of our model. This also serves as a test for generalizability in locally advanced NSCLC patients treated with different standard of care treatment protocols. To match the number of input time points, the one-year overall survival model with the pre-treatment and first follow-up at one month was used. The model was able to separate the pathologic responders from those with gross residual disease in the trimodality cohort. This was the case, even though the model development was completely blinded from this cohort.

This prediction was compared to a well-known prediction of response, the primary tumor size. The change in tumor volume also predicted the response in this cohort with a similar performance. However, the two measures, model probability and delta volume, were only weakly correlated and the combined model showed a slight improvement in

performance. The proposed model was able to predict pathologic response in a different cohort, with only the image and a seed point for input. There is also a weak correlation between the values, which suggests that the image based neural network model is detecting radiographic characteristics other than tumor size.

The use of a CNN based network captures the tumor region and the immediate tumor environment. Previous techniques focused on providing the machine learning algorithm with accurate manual delineations or semi-automated methods which may not incorporate surrounding tissue[36,37].CNN image input includes the boundary between the tumor and the normal tissue environment. This may provide additional indications for tumor response and infiltration to the surrounding tissue. Image augmentation was performed on the training tumor region, as conventional practice in the field of deep learning and biomedical image processing[38], to improve performance and the small-scale deformations were applied to prevent overfitting[39] on our relatively small training set. The use of conventional ResNet CNN for image characterization allows for the incorporation of pre-treatment weights on natural images[26]. This mediated the application of deep neural networks on medical images, with cohorts much smaller than the millions of samples used in other artificial intelligence solutions.

The number of samples available for most radiological studies are not on the same order of magnitude as those used for deep learning applications. For instance, a facial recognition deep learning application was developed by training on 87 thousand images and testing on 5 thousand images[40]. However, transfer learning can be used to leverage common low-level CNN parameters from databases such as ImageNet, which contains over 14 million natural images[26]. It would be ideal to incorporate the whole tumor volume by using a network pre-trained on 3D radiographic images or 3D images in general, however the number of images available are not near the order of magnitude of which are in photographic images. If available, a model pre-trained in 3D CT images with samples on the order of thousands of images will likely be overfitted to the patient cohort, the institution, and the outcome the network was trained to predict. The use of transfer learning has demonstrated its effectiveness on improving the performance of lung nodule detection in CT images[18]. Our study contained a sample size not on the order of studies based on photographic images, but the current performance was made possible with the incorporation of pre-trained networks on ImageNet. Transfer learning may also be used to test the feasibility of clinically applicable utilities prior to the collection of a full cohort for analysis.

The incorporation of follow-up time points to capture dynamic tumor changes was key to the prediction of survival and tumor prognosis. This was feasible with the use of RNNs, which allowed for amalgamation of several time points and the ability to learn from samples with missed patient scans at a certain time point, which is inevitable in retrospective studies such as this one. Although this type of network has not been applied to medical images, similar network architectures have demonstrated success in image and time dependent analyses, as in video classification and description applications[41]. The model was structured to overcome the immortal time bias[42]. The pooling of CNN without the RNN has been previously applied[43] , but in this case would result in bias

classifications for an event when the last patient scan is missed. The RNN was set to not learn from inputs where there is a missing scan[44]. GRU RNNs were used as they contain an update gate and a reset gate, which decides the weighting of the information passed on to the network output[45]. This captures the pertinent information from each time point for the survival and prognostic predictions.

Previous work has demonstrated the feasibility of using CT imaging features to make associations and predictions in lung cancer[7]. Several studies used radiomics approaches involving manual delineation of the tumor volume and user-defined calculated features to make predictions of survival and pathological response[12–15]. Recent applications of deep learning on lung cancer have focused on lung nodule classification as benign or metastatic and they focus on a single scan for the model input. The study by Kumar et. al. depended on manual delineation of lung nodules with feature extraction using an autoencoder and classification with decision trees[46]. Hua et. al. used 2D region of the tumor lesion on the axial slice for classification, also performed at one time point[47]. Our study differs mainly in the incorporation of multiple time points in the prediction of survival and prognostic factors. For further validation, we also applied our developed model on a different cohort for the prediction of pathologic response, an important clinical factor. In comparison to previous studies, our model only takes a seed point and creates a 50 × 50 mm^2 region around the seed point, which is used as input. In order to compute handcrafted radiomic features, an accurate tumor delineation is required[9], which is susceptible to inter-reader segmentation variability and also is time-consuming. Recently, deep learning has been shown to have higher performance than conventional radiomics[39]. Our approach only required a seed point within a tumor and hence is more efficient and robust to manual inference. Additional clinical and pathological evaluations are not always available. Morphological parameters dependent on manual and semi-automated contours of the whole tumor volume or RECIST[5] measurements are prone to inter-operator variability and can be costly to acquire.

Ideally, after training on a larger diverse population and after extensive external validation and benchmarking with current clinical standards, quantitative prognostic prediction models can be implemented in the clinic[48]. There are several lung nodule detection algorithms available in the literature and with the aid of the pretreatment tumor contours routinely delineated by the radiation oncologist, the location of the tumor on the follow up images can be detected automatically[49]. The input of our model would simply be the bounding box surrounding the detected tumor and can be cropped automatically as well. The trained network can generate probabilities of prognosis within a few seconds, and thus would not hinder current clinical efficiency. The probabilities can then be presented to the physician along with other clinical images and measures, such as the RECIST criteria[5], to aid in the process of patient assessment.

This proof of principle study has its limitations, one of which is the sample size of the study cohorts. Thus, a pre-trained CNN was used to improve predictive power. Using a deep-learning technique has its limitations. Previous associations were found for risk of distant metastases with the pre-treatment scan only, with machine learning techniques[15]. It has been demonstrated that machine learning based on engineered features out

performs deep learning with small sample sizes. Perhaps with a larger cohort, we could potentially achieve better performance deep learning. The probabilities are essentially calculated with a black box for a specific task, thus are less practical than engineered features, which could potentially be reused for other applications. Neural networks can be prone to overfitting, even with the techniques we have used to mitigate this[29,30], thus images were resampled to a common pixel spacing. Our model used three 2D slices due to the predefined parameters necessary for transfer learning. However, a 3D image volume may better represent tumor biology and thus increase performance. Our survival models are based purely on the CT image and could potentially benefit from the incorporation of patient specific parameters, such as age, sex, histology, smoking cessation and radiation therapy parameters, with a larger cohort of patients. With these limitations, our deep learning model was able to make predictions of survival and perhaps with a larger dataset and finer more consistent axial spacing, higher and more clinically relevant performance may be feasible.

Deep learning is a flexible technique which has been successfully implemented in several fields[16]. However, the theory behind how the network functions has yet to be established[50]. The input and output of the model can be quite intuitive, but as suggested by the term, the hidden middle layers are not. It is therefore very challenging to determine the reasoning behind a network's performance and whether certain parameters have a positive or negative impact. Unlike engineered features built to capture certain characteristics of the image, the interpretation of deep learning features can be ambiguous. To circumvent this in the field of image-based CNN, activation maps have been generated to capture highly weighted portions of the image with respect to the network's predictions. This can be visualized in the form of heat maps, generated over the final convolutional layer. Also, how to incorporate the domain knowledge into these abstract features is a very important question that needs to be addressed. Further research in this direction could make these automatically learned feature representations more interpretable.

Conclusion

This study demonstrated the impact of deep learning on tumor phenotype tracking before and after definitive radiation therapy through pretreatment and CT follow-up scans. There were increases in performance of survival and prognosis prediction with incorporation of additional time points using CNN and RNN networks. This was compared to the performance of clinical factors, which were not significant. The survival neural network model could predict pathological response in a separate cohort with trimodality treatment after radiation therapy. Although the input of this model consisted of a single seed point in put at the center of the lesion, without the need for volumetric segmentation our model had comparable predictive power compared to tumor volume, acquired through time consuming manual contours. Non-invasive tracking of the tumor phenotype predicted survival, prognosis and pathological response, which can have potential clinical implications on adaptive and personalized therapy.

Acknowledgements

Authors acknowledge financial support from the National Institute of Health (NIH-USA U24CA194354, and NIH-USA U01CA190234); https://grants.nih.gov/funding/index.htm. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supporting Information

https://clincancerres.aacrjournals.org/content/suppl/2019/09/21/1078-0432.CCR-18-2495.DC1

References

1. Torre, L. A. *et al.* Global cancer statistics, 2012. *CA Cancer J. Clin.* **65**, 87–108 (2015).
2. Ettinger, D. S. *et al.* Non-small cell lung cancer. *J. Natl. Compr. Canc. Netw.* **10**, 1236–1271 (2012).
3. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA Cancer J. Clin.* **66**, 7–30 (2016).
4. Goldstraw, P. *et al.* The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *J. Thorac. Oncol.* **11**, 39–51 (2016).
5. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
6. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Hugo J W. Artificial intelligence in radiology. *Nat. Rev. Cancer* (2018) doi:10.1038/s41568-018-0016-5.
7. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Hugo J W. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci. Rep.* **5**, (2015).
8. Aerts, H. J. W. L. Data Science in Radiology: A Path Forward. *Clin. Cancer Res.* **24**, 532–534 (2018).
9. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).
10. Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **15**, e1002711 (2018).
11. Parmar, C., Barry, J. D., Hosny, A., Quackenbush, J. & Aerts, H. J. Data Analysis Strategies in Medical Imaging. *Clin. Cancer Res.* (2018) doi:10.1158/1078-0432.CCR-18-0385.
12. Coroller, T. P. *et al.* Radiomic-Based Pathological Response Prediction from Primary Tumors and Lymph Nodes in NSCLC. *J. Thorac. Oncol.* **12**, 467–476 (2017).
13. Huynh, E. *et al.* CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. *Radiother. Oncol.* **120**, 258–266 (2016).
14. Coroller, T. P. *et al.* Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiother. Oncol.* **119**, 480–486 (2016).
15. Coroller, T. P. *et al.* CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother. Oncol.* **114**, 345–350 (2015).
16. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
17. Shin, H.-C. *et al.* Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).
18. Shin, H.-C. *et al.* Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).

19. Dandil, E. *et al.* Artificial neural network-based classification system for lung nodules on computed tomography scans. in *2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)* (2014). doi:10.1109/socpar.2014.7008037.

20. Shin, H.-C. *et al.* Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).

21. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).

22. Hermann, P. C. *et al.* Distinct populations of cancer stem cells determine tumor growth and metastatic activity in human pancreatic cancer. *Cell Stem Cell* **1**, 313–323 (2007).

23. Donahue, J. *et al.* Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 677–691 (2017).

24. Cierniak, R. A New Approach to Image Reconstruction from Projections Using a Recurrent Neural Network. *Int. J. Appl. Math. Comput. Sci.* **18**, 147–157 (2008).

25. Fedorov, A. *et al.* 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **30**, 1323–1341 (2012).

26. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).

27. Rubins, J., Unger, M. & Colice, G. L. Follow-up and Surveillance of the Lung Cancer Patient Following Curative Intent Therapy. *Chest* **132**, 355S–367S (2007).

28. Calman, L. *et al.* Survival benefits from follow-up of patients with lung cancer: a systematic review and meta-analysis. *J. Thorac. Oncol.* **6**, 1993–2004 (2011).

29. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv [cs.LG]* (2015).

30. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

31. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). doi:10.1109/cvpr.2016.90.

32. Albain, K. S. *et al.* Phase III study of concurrent chemotherapy and radiotherapy (CT/RT) vs CT/RT followed by surgical resection for stage IIIA(pN2) non-small cell lung cancer (NSCLC): Outcomes update of North American Intergroup 0139 (RTOG 9309). *J. Clin. Orthod.* **23**, 7014–7014 (2005).

33. Tsujino, K. *et al.* Predictive value of dose-volume histogram parameters for predicting radiation pneumonitis after concurrent chemoradiation for lung cancer. *International Journal of Radiation Oncology*Biology*Physics* **55**, 110–115 (2003).

34. Hellmann, M. D. *et al.* Pathological response after neoadjuvant chemotherapy in resectable non-small-cell lung cancers: proposal for the use of major pathological response as a surrogate endpoint. *Lancet Oncol.* **15**, e42–50 (2014).

35. Pataer, A. *et al.* Histopathologic response criteria predict survival of patients with resected lung cancer after neoadjuvant chemotherapy. *J. Thorac. Oncol.* **7**, 825–832 (2012).

36. Parmar, C. *et al.* Robust Radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One* **9**, e102107 (2014).

37. Mackin, D. *et al.* Measuring Computed Tomography Scanner Variability of Radiomics Features. *Invest. Radiol.* **50**, 757–765 (2015).

38. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Lecture Notes in Computer Science* 234–241 (2015).

39. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).

40. Sun, Y., Wang, X. & Tang, X. Deep Learning Face Representation from Predicting 10,000 Classes. in *2014 IEEE Conference on Computer Vision and Pattern Recognition* 1891–1898 (IEEE, 2014).

41. Joe Yue-Hei Ng *et al.* Beyond short snippets: Deep networks for video classification. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015). doi:10.1109/cvpr.2015.7299101.

42. Suissa, S. Immortal time bias in pharmaco-epidemiology. *Am. J. Epidemiol.* **167**, 492–499 (2008).

43. Karpathy, A. *et al.* Large-Scale Video Classification with Convolutional Neural Networks. in *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014). doi:10.1109/cvpr.2014.223.

44. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci. Rep.* **8**, 6085 (2018).

45. Cho, K. *et al.* Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014). doi:10.3115/v1/d14-1179.

46. Kumar, D., Wong, A. & Clausi, D. A. Lung Nodule Classification Using Deep Features in CT Images. in *2015 12th Conference on Computer and Robot Vision* (2015). doi:10.1109/crv.2015.25.

47. Hua, K.-L., Hsu, C.-H., Hidayati, S. C., Cheng, W.-H. & Chen, Y.-J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *Onco. Targets. Ther.* **8**, 2015–2022 (2015).

48. Lehman, C. D. *et al.* Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. *Radiology* 180694 (2018).

49. Valente, I. R. S. *et al.* Automatic 3D pulmonary nodule detection in CT images: A survey. *Comput. Methods Programs Biomed.* **124**, 91–107 (2016).

50. Wang, G. A Perspective on Deep Imaging. *IEEE Access* **4**, 8914–8924 (2016).

8

Chapter 8

Clinical Validation of Deep Learning Algorithms for Lung Cancer Radiotherapy Targeting

A Hosny, DS Bitterman, CV Guthier, H Roberts, S Perni, A Saraf, JM Qian, LC Peng, IM Pashtan, Z Ye, BH Kann, D Kozono, P Catalano, HJWL Aerts & RH Mak

Submitted 2021

Abstract

Background: Artificial intelligence (AI) and deep learning (DL) methods have demonstrated great potential in streamlining clinical tasks, including radiation treatment planning. However, most studies are confined to *in silico* validation in small internal cohorts, lacking data on real-world clinical utility.

Purpose: In this study, we developed a multifaceted strategy for the clinical validation of DL models for assisted and fully-automated segmentation of primary non-small cell lung cancer (NSCLC) tumors and involved lymph nodes in computed tomography (CT) images - a time intensive radiation treatment planning step with large variability among experts.

Materials and Methods: CT images and expert segmentations were collected from eight independent internal and external sources totalling 2208 NSCLC patients: 787 for model discovery and 1421 for validation. Our clinical validation strategy consisted of benchmarking, primary validation, functional validation, and end-user testing. Primary validation consisted of stepwise testing on increasingly external and heterogeneous datasets using volumetric (VD) and surface (SD) dice metrics among others. Functional validation explored model stability and accuracy in test-retest and phantom images, dosimetric evaluation, and failure mode analysis. End-user testing with eight radiation oncologists was carried out to test automated segmentations in a simulated clinical setting.

Results: *Benchmarking:* Models showed an improvement over the interobserver benchmark ($P < .01$), and were within the intraobserver benchmark. *Primary Validation*: Performance on internal data segmented by the same expert was VD 0.83 [0.82,0.85], within the interobserver benchmark. Performance on internal data segmented by other experts was VD 0.70 [0.67,0.73], worse than the interobserver benchmark ($P < .0001$). Performance on the RTOG-0617 clinical trial data was VD 0.71 [0.69,0.73], with similar results on diagnostic radiology datasets. *Functional*

Validation: Models were found to be stable across separate images of the same subject, and yielded planning target volumes with equivalent radiation dose coverage compared to human experts. *End-user testing:* We found non-significant differences between *de novo* expert and AI-assisted segmentations for both VD and SD. AI-assistance led to a 65% reduction in segmentation time ($P < .0001$) and 32% reduction in interobserver variability ($P < .05$).

Conclusion: We developed and validated a high performing automated segmentation model for a difficult clinical task with high interobserver variability. Our validation strategy may help assess clinical utility beyond the proof-of-concept stage, provide sufficient confidence in pursuing prospective clinical trials, and guide future research in this domain.

Introduction

Lung cancer is the leading cause of cancer-related mortalities worldwide[1], while being the second most commonly diagnosed cancer in both men and women[2]. Non-small-cell lung cancer (NSCLC) is the most common type of lung cancer, accounting for 85% of all diagnoses[3]. Radiation therapy (RT) plays a key role in treating NSCLC, with one fifth and one half of early and late stage patients, respectively, receiving this treatment modality[4]. RT is also highly versatile as it may be administered as a sole treatment, with systemic agents, precede or follow surgery, and play a role in palliation[5].

RT's time- and cost-effectiveness is impacted by an expensive upfront investment: RT planning. RT planning is crucial in maximizing and minimizing radiation to cancer and normal cells, respectively. After a clinical decision for RT treatment and image acquisition, planning steps include image registration, target and adjacent organ segmentation, and dose distribution design among others. The manual segmentation of the target i.e. primary tumor and involved lymph nodes, is one of the most time consuming planning tasks performed by radiation oncologists[6,7]. This meticulous task requires interpreting images on a voxel-by-voxel basis to delineate the target volume, unlike diagnostic interpretation where reporting image-level findings is often sufficient[8]. The advent of advanced RT planning and delivery techniques such as intensity modulated RT (IMRT) and image guidance have enabled smaller margins and less dose to surrounding organs, but require higher segmentation accuracy[9]. Additionally, physicians' personal style and preferences contribute to a large and well documented interobserver variability in target segmentation[10–12], even in RT clinical trials with pre-specified parameters[13]. Finally, the accuracy of target segmentation can directly affect patient outcomes where under-segmentation can result in underdosing and decreased tumor control, while over-segmentation can result in overdosing and increased toxicity risks[14,15].

Multiple computer-aided tools have been proposed to help streamline RT planning[6]. For segmentation tasks, semi-automated approaches that incorporate knowledge from a collection of reference images, known as segmentation atlases, have had varying degrees of clinical utility[16]. Curating atlases requires substantial time and effort on the physician's end, and the heterogeneity of its contents may diminish performance[17]. More recently, artificial intelligence (AI) methods - deep learning (DL) specifically - have been proposed as promising alternatives[18]. Unlike prior methods, DL algorithms are able to automatically learn feature representations from data, ultimately contributing to superior performance across multiple tasks[19]. The versatility of these algorithms has also led to widespread applications across imaging modalities, tissue types, and disease sites[20]. While many studies have explored the use of DL to automate RT target segmentation and improve its accuracy and consistency[21], most remain at the proof-of-concept stage. As such, they are often confined to *in silico* validation in small internal cohorts. Within a sea of promising results, only a few efforts demonstrate the clinical impact of these automated systems[22,23].

In this study, we present a generalizable clinical validation strategy for therapeutic AI algorithms with the aim of bridging early proof-of-concept studies and prospective

clinical trials, while providing sufficient confidence in pursuing the latter. The strategy comprises four main components including developing benchmarks, performing primary and functional validation, as well as conducting end-user testing **(Fig. 1)**.

To demonstrate the application of this strategy, we present a study in clinically validating DL models for RT targeting. We performed an integrative analysis on eight independent datasets (2208 patients). Utilizing a discovery cohort of 787 patients, we developed multiple DL models for localizing and segmenting primary NSCLC tumors and involved lymph nodes in CT images. We then established an interobserver benchmark across six radiation oncologists, followed by an intraobserver benchmark across images segmented by the same radiation oncologist. Primary validation was carried out across 1421 patients including both internal and external cohorts, RT clinical trial data, as well as diagnostic radiology images. Functional validation was conducted across multiple datasets including test-retest and thorax phantom images. Therein, we assessed the dosimetric impact of AI segmentations, measured their stability and accuracy, as well as identified failure modes. Finally, in order to gauge the clinical utility of AI segmentations, we carried out end-user testing. In a simulated clinical setting, eight radiation oncologists from our institution were asked to perform the segmentation task *de novo* as well as rate and edit a provided AI segmentation. Taken together, these studies comprehensively assess the performance of the DL models both standalone and in their intended use setting for robust validation prior to prospective trials and ultimate clinical integration.

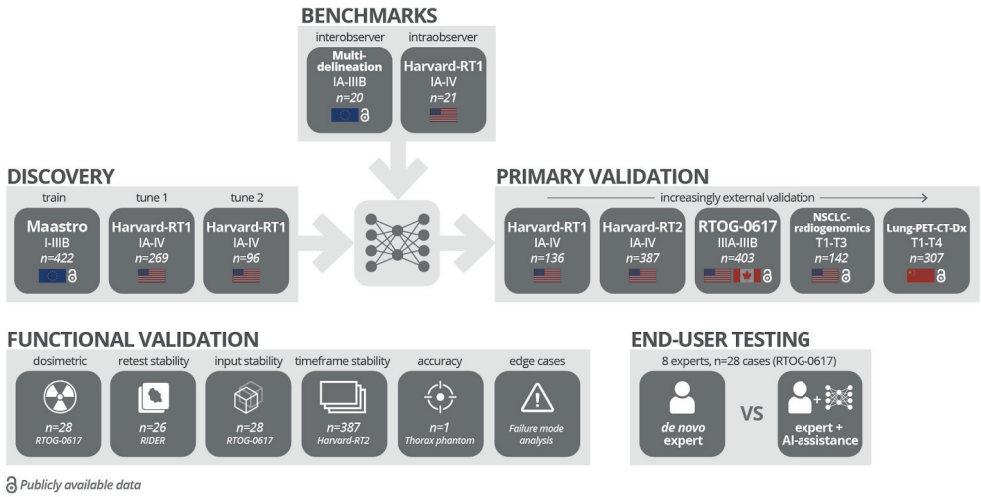

Figure 1. Clinical validation framework and experimental setup.
We performed an integrative analysis on 8 independent datasets totaling 2208 patients to assess the performance of deep learning models in localizing and segmenting primary NSCLC tumors and involved lymph nodes in CT images. Two datasets, Maastro and Harvard-RT1, were used to train fully convolutional neural network models to perform the tasks. Both interobserver and intraobserver benchmarks were established. Primary model validation was carried out on five datasets, two of which are internal and three are external. Functional validation was also conducted to assess the dosimetric impact of AI segmentations and measure their stability and accuracy among others. Finally, end-user testing was carried out in a simulated clinical setting in order to gauge the clinical utility of AI segmentations when provided to physicians.

Materials & Methods

Discovery

The following datasets were used for model development:

- **Maastro:** 422 patients (stages I-IIIB; 290 male, 132 female; mean age 68) with histologically proven NSCLC and treated with radiotherapy alone (n=196) or with chemo-radiation (n=226). Patients were treated at MAASTRO Clinic, Maastricht, The Netherlands between 2004 and 2010, and imaged with CT - with or without intravenous contrast. See Supplementary Table 8. This dataset is publicly available at https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics

- **Harvard-RT1:** 501 patients (stages IA-IV; 263 male, 236 female, 2 unspecified; median age 73) with histologically proven NSCLC referred for radiotherapy between 2001 and 2015 at the Dana-Farber Cancer Institute and Brigham and Women's Hospital, Boston, Massachusetts, US. Patients were imaged with CTwith or without intravenous contrast. Target volumes were delineated by a single radiation oncologist (R.H.M., R1). 269, 96, and 136 patients from this dataset were used for training, tuning, and testing the segmentation models respectively. The test set is identical to that used in a previously published tumor segmentation study[8]. See Supplementary Table 9.

Due to varying CT slice thickness across datasets **(Supplementary Fig. 39)**, preprocessing involved resampling all data to a common voxel spacing of 1*1*3 mm^3. This was achieved using linear and nearest neighbor interpolations for CT images and segmentations, respectively. Images were normalized to a 0 to 1 range (–1,024 to 3,071 Hounsfield units). Use of intravenous contrast in images was detected using a published algorithm[24]. During model training, data augmentation included transformations (scaling, rotating, mirroring), addition of gaussian noise and blur, as well as brightness and contrast adjustments. No testing-time augmentation was applied. Our assisted and automated pipelines consist of four 3D U-Net models - closely following the original implementation[25,26] - for the localization and segmentation of lungs, primary tumor, as well as involved thoracic lymph nodes. The assisted pipeline requires a user-placed seed point within the tumor volume, while the automated pipeline is fully autonomous. For pipeline schematics, see supplementary Fig. 17. Each model comprised 2 blocks per level along both the encoder and decoder. Each block contained a convolutional layer with instance normalization[27] and leaky ReLU activation[28]. Strided and transposed convolutions were used to downsample and upsample the images respectively. Number of feature maps started at 32 and was doubled and halved at every level along the encoder and decoder respectively. For model specifications, see Supplementary Table 6. For training, we used the stochastic gradient descent (SGD) optimizer with Nesterov momentum (μ = 0.99) and an initial learning rate of .01 (decay using polynomial policy[29]) for a maximum of 1000 epochs. The loss function used was dice coefficient combined with cross entropy[30]. Pytorch[31] was used for model development, and nnU-

Net[20] for hyper-parameter tuning. Multiple segmentation metrics were used for model validation **(Supplementary Table 7)**.

Benchmarks

The interobserver benchmark was developed using the Multi-delineation dataset. 20 patients (stages IA–IIIB; 12 male, 8 female; median age 67) with histologically proven NSCLC referred for radiotherapy at Maastro Clinic, Maastricht, The Netherlands[32]. Manual tumor delineations were performed on pre-treatment CT images by five radiation oncologists: three specializing in thoracic oncology and two residents. Tumor volumes were also delineated by R1 to create the interobserver benchmark for a total of six experts. See Supplementary Fig. 43 and Supplementary Table 10. This dataset is publicly available at https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics-Interobserver1

The intraobserver benchmark was developed using 21 randomly sampled patients from the Harvard-RT1 test set. R.H.M. performed the segmentation task twice with a three months washout period in between.

Primary Validation

In addition to testing on the Harvard-RT1 test set, further validation was performed on the following increasingly external datasets:

- **Harvard-RT2:** 387 patients (stages IA-IV; 165 male, 222 female; median age 69) with histologically proven NSCLC referred for radiotherapy between 2011 and 2017 at the Dana-Farber Cancer Institute and Brigham and Women's Hospital, Boston, Massachusetts, US. Patients were imaged with CT with or without intravenous contrast. Tumor volumes were delineated by multiple physicians. See Supplementary Fig. 11 and Supplementary Table 11.

- **RTOG-0617:** 403 NSCLC patients (stages IIIA-IIIB; 223 male, 155 female, 25 unspecified; median age 64) from the clinical trial RTOG-1617 (NCT00533949)[33,34], "High-Dose or Standard-Dose Radiation Therapy and Chemotherapy With or Without Cetuximab in Treating Patients With Newly Diagnosed Stage III Non-Small Cell Lung Cancer That Cannot Be Removed by Surgery". Patients were treated between 2007 and 2011 at 185 institutions across the USA and Canada. Thoracic CT data were obtained within 6 weeks of trial registration. See Supplementary Fig. 44 and Supplementary Table 12. This dataset is publicly available at https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=33948334

- **NSCLC-radiogenomics:** 142 (total n=162) early stage NSCLC patients (pathological stages T1-T3, N0-N2, M0-M1; 124 male, 38 female; mean age 68) referred for surgical treatment at Stanford University School of Medicine (n=69) and Palo Alto Veterans Affairs Healthcare System (n=93)[35] in California, US. Patients were treated between 2008 and 2012. Data consisted of preoperative CT

images with automated tumor segmentations that were edited and reviewed by two thoracic radiologists. See Supplementary Fig. 45 and Supplementary Table 13. This dataset is publicly available at https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics

- **Lung-PET-CT-Dx:** 307 NSCLC patients (clinical stages T1-T4, N0-N3, M0-M3; 163 male, 144 female; mean age 61) imaged at the Second Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang Province, China[36]. Data consisted of CT images, together with PET/CT images for a subset of patients. Tumor location was annotated using per-slice bounding rectangles by five academic thoracic radiologists with expertise in lung cancer. See Supplementary Fig. 46 and Supplementary Table 14. This dataset is publicly available at https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70224216

Functional Validation

Data utilized in the dosimetric analysis is a random quartile-based 28 patient subset of the RTOG-0617 clinical trial dataset **(Supplementary Fig. 18)**. While the models described herein predict the gross tumor volume (GTV), it is the planning target volume (PTV) that accounts for uncertainty and is ultimately used for dose calculation in RT planning. AI PTVs were generated from AI GTV as follows. First, a uniform expansion of 5mm was applied to the GTV to generate the clinical target volume (CTV). This represents the lower bound of the 5mm to 10mm range specified in RTOG-0617 clinical trial protocol[37]. The CTV was further uniformly expanded by the mean margin between CTV and PTV segmentation used for each patient in the trial independently. Statistics of this margin were min=4.6mm, mean=8.1mm, max=12.1mm. The same treatment plans and radiation dose distributions used in the trial were utilized. Dose volume histograms and other dose calculations were performed in 3Dslicer[38] using the SlicerRT extension[39].

Test-retest stability was assessed using RIDER. 26 (total n=32) NSCLC patients (primary tumor >=1cm; 16 men, 16 women; mean age 62) each of whom underwent two CT scans of the chest within 15 minutes[40]. Images were acquired between January 2007 and September 2007 at the Memorial Sloan-Kettering Cancer Center, New York, USA. Tumor segmentations were initially performed by an automated segmentation algorithm and inspected by two thoracic radiologists. See Supplementary Fig. 47. This dataset is publicly available at https://wiki.cancerimagingarchive.net/display/Public/RIDER+Lung+CT. Additionally, we tested the assisted models' stability as a function of variation in input data by simulating multiple readers' placement of seed points (n=50 simulations). Stability across 3D and 4D CT data was tested in Harvard-RT2 comprising n=186 single timeframe 3D CT with GTV annotations, against n=201 multi timeframe 4DCT with internal gross target volume (iGTV) annotations. A GTV was predicted at each of the 10 timeframes and combined to produce an iGTV that compensates for the tumor's physiological movements.

Model accuracy was assessed using a CT study of a thorax phantom containing 12 synthetic lesions (10 and 20 mm in effective diameter) inserted into the lungs - 6

lesions per lung[41,42]. The phantom was scanned at Columbia University Medical Center, NewYork, USA. See imaging information Supplementary Table 15. This dataset is publicly available at https://wiki.cancerimagingarchive.net/display/Public/Lung+Phantom

End-user testing

We recruited eight radiation oncologists from our institution with varying degrees of experience (three attendings and five residents, Supplementary Table 2 and 3). All readers consented to the testing. Data used in this analysis was a random quartile-based 28 patient subset of the RTOG-0617 clinical trial dataset, which was not used in model development. This subset was further divided into two random quartile-based groups of 14 patients each (Supplementary Fig. 18). For group A patients, readers were asked to perform the primary tumor and lymph node segmentation task *de novo*. For group B patients, readers were asked to rate and edit a provided segmentation blinded to its source. For 10 patients, automated segmentations from the assisted pipeline were provided (AI-assisted). For 4 patients, clinical segmentations from the RTOG-0617 clinical trial were provided (expert-assisted, Supplementary Fig. 48, Supplementary Table 4). The testing was conducted in a simulated clinical setting. A workflow was set up in MIM®, the software typically used for this task at our institution **(Supplementary Fig. 49)**. Readers were provided the following information for each patient: age, gender, Zubrod score, histology, stage, and primary tumor lung lobe. All answers to survey questions were collected before, during, and after each case within the same software environment. Time for task completion was also recorded automatically in the background.

Statistics

All statistical tests conducted were non-parametric, with a two-tailed $P < .05$ indicating significance. For two dependent groups, the Wilcoxon matched-pairs signed rank test was used. For two independent groups, the Mann-Whitney U rank test was used. For three or more independent groups, the Kruskal-Wallis H-test was used. For measuring correlation between two groups, the Spearman rank-order correlation coefficient was used.

Results

Discovery

DL models were developed to localize and segment primary NSCLC tumors and involved thoracic lymph nodes in pretreatment CT images, either assisted by a user-placed seed point or fully automated (Methods, Supplementary Fig. 17). Models were first trained using the Maastro dataset, then fine tuned using 365 patients from Harvard-RT1 with tumor segmentations performed by the main expert radiation oncologist from this study (R.H.M., R1). Multiple segmentation metrics were used to evaluate model performance (Methods). The most common of these, namely volumetric dice (VD) and surface dice (SD), are reported hereafter (median [95%CI]).

Benchmarks

Interobserver benchmark was VD 0.83 [0.82,0.84]; SD 0.72 [0.7,0.75] **(Supplementary Fig. 1)**. AI vs R1 yielded VD 0.91 [0.84,0.92]; SD 0.86 [0.74,0.91], a significant improvement over the benchmark with VD, $P < .01$; SD, $P < .001$ **(Supplementary Fig. 2 and 3)**. Additionally, AI vs R1 was found to be inversely correlated with the interquartile range of variability among all 6 readers, Spearman R -0.74, $P < .001$ **(Supplementary Fig. 4)**. With AI segmentations as reference, non-significant differences were detected between residents and attendings **(Supplementary Fig. 5)**.

Intraobserver benchmark was VD 0.88 [0.84,0.9]; SD 0.85 [0.81,0.93] (Supplementary Fig. 6). AI vs R1's first read yielded VD 0.86 [0.83,0.87]; SD 0.79 [0.74,0.88], with similar results for the second read. Non-significant differences were observed when both results were compared to the benchmark **(Supplementary Fig. 7 and 8)**.

Primary validation

First, we tested on 136 patients held out from the internal Harvard-RT1 dataset, which were also segmented by R1. Assisted primary tumor segmentation results were VD 0.86 [0.85,0.87]; SD 0.83 [0.80,0.85], a significant improvement over previously published contest results on the same data[8] ($P < .0001$, Supplementary Fig. 10). Results for primary tumor and lymph node segmentation were VD 0.83 [0.82,0.85]; SD 0.79 [0.75,0.81] for the assisted model and VD 0.82 [0.80,0.83]; SD 0.74 [0.71,0.76] for the automated model (2% localization failure rate, Supplementary Table 1, Fig. 2).

Second, we tested on the internal Harvard-RT2 dataset comprising segmentations by other radiation oncologists in our institution. Results for primary tumor and lymph node segmentation were VD 0.70 [0.67,0.73]; SD 0.50 [0.47,0.54] for the assisted model and VD 0.63 [0.61,0.67]; SD 0.44 [0.40,0.48] for the automated model (10% localization failure rate, Supplementary Table 1, Fig. 2).

Third, we tested on data collected as part of the RTOG-0617 clinical trial[34,37,43]. Results for primary tumor and lymph node segmentation were VD 0.71 [0.69,0.73]; SD 0.47 [0.45,0.49] for the assisted model and VD 0.69 [0.67,0.72]; SD 0.44 [0.42,0.47] for the automated model (0.5% localization failure rate, Supplementary Table 1, Fig. 2). Non-significant differences in performance were observed between trial arms: high vs low radiation dose **(Supplementary Fig. 13)**, as well as between RT treatment techniques: 3D conformal (3D-CRT) vs IMRT **(Supplementary Fig. 14)**.

Finally, we tested on two diagnostic datasets. For NSCLC-radiogenomics[35], tumor segmentation results were VD 0.68 [0.63,0.73]; SD 0.61 [0.54,0.74] for the assisted model and VD 0.64 [0.59,0.69]; SD 0.55 [0.47,0.66] for the automated model (6% localization failure rate, Supplementary Table 1, Fig. 2). Non-significant differences were observed between lung lobes (Supplementary Fig. 15). For Lung-PET-CT-Dx dataset[36], results were VD 0.66 [0.64,0.68]; SD 0.31 [0.29,0.34] for the assisted model and VD 0.61 [0.59,0.64]; SD 0.27 [0.24,0.29] for the automated model (9% localization failure rate, Supplementary Table 1, Fig. 2).

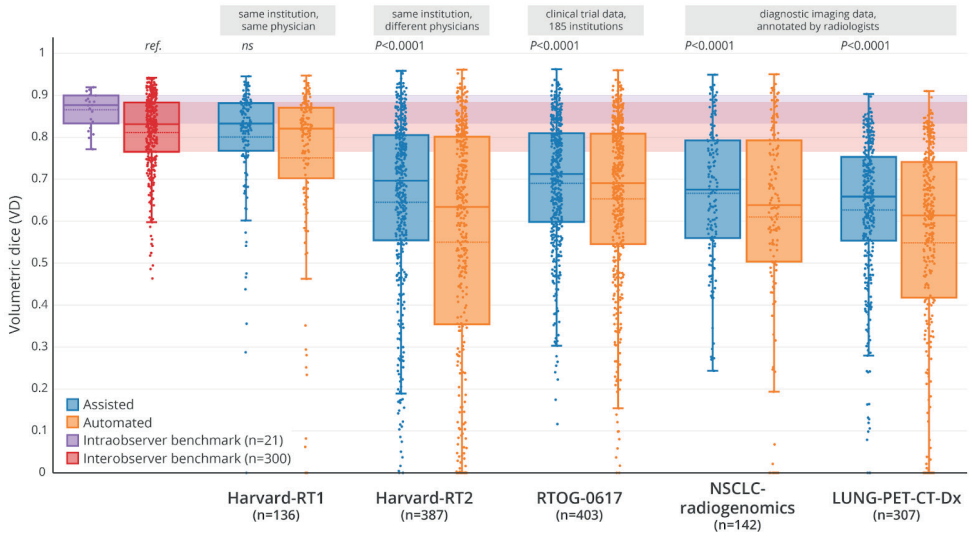

Figure 2. Primary validation results and comparison with benchmarks.
Deep learning model performance in localizing and segmenting primary NSCLC tumors and involved lymph nodes, as validated on five increasingly external datasets using the volumetric dice (VD) metric. First, the intraobserver and interobserver benchmarks were established, depicted here in purple and red respectively. Validation first started on Harvard-RT1, the dataset that most resembles that training data i.e. from the same institution and annotated by the same physician. Next is Harvard-RT2 also from the same institution but annotated by other physicians. This was followed by validation on RTOG-0617, a clinical trial dataset collected from 185 institutions. Final validation was conducted on diagnostic data annotated by radiologists. Blue box plots depict the seed point assisted models, while orange box plots depict the fully automated models. The Mann-Whitney U rank test was used, with a two-tailed $P < .05$ indicating significance. See Supplementary Fig. 9 for other segmentation metrics including surface dice (SD) and precision.

Functional validation

To assess changes in radiation dose delivered as a result of using AI-generated segmentations in RT treatment planning, we performed a dosimetric analysis **(Supplementary Fig. 19)**. Non-significant differences were observed between the clinical and AI PTV segmentations across two common dose coverage metrics: V95, or percent target volume that received at least 95% of the prescription dose, and D95, or dose covering 95% of the target volume **(Supplementary Fig. 20)**.

Model stability across two separate CT scans of the same subject were assessed using the RIDER dataset[40] **(Supplementary Fig. 23)**. AI vs radiologist on the first scan was non-significantly different from the same comparison on the second scan (Supplementary Fig. 24). Radiologists' variation in tumor volume across the two scans was non-significantly different from that of the AI models **(Supplementary Fig. 25)**. Additionally, we tested the assisted models' stability as a function of variation in seed point placement. Median model predictions showed high stability with an interquartile range of 0.02 for both VD and SD **(Supplementary Fig. 26)**. With regards to stability across CT timeframes,

non-significant differences in performance were observed between 3D and 4D input CT data for both the assisted and automated models **(Supplementary Fig. 12)**.

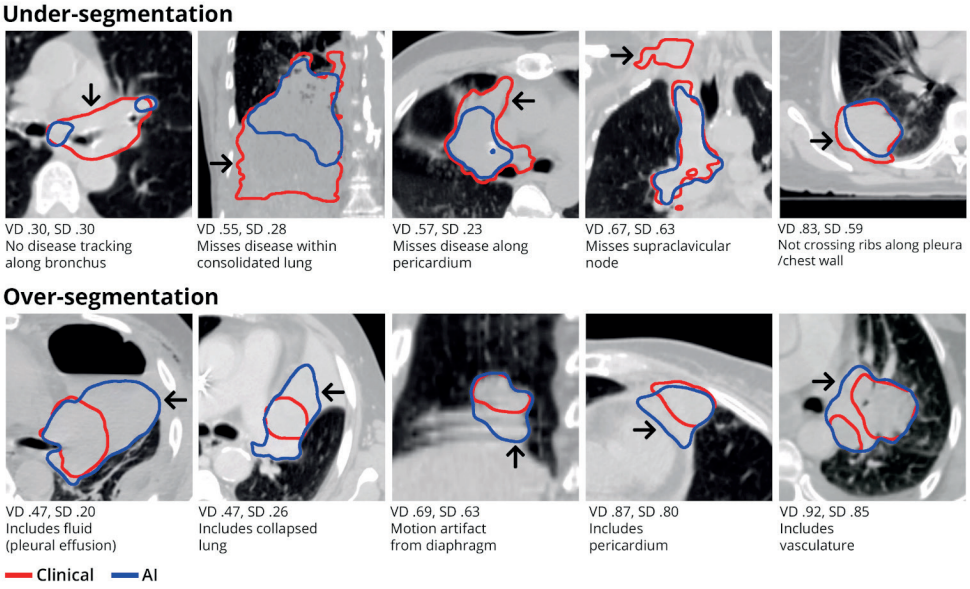

Figure 3. Examples of model failure modes
10 representative examples of model failures for both under- and over-segmentation scenarios (5 cases each). Cases are ordered left to right in increasing model performance metrics.

To gauge model accuracy, we tested on a CT scan of a thorax phantom containing 12 nodules of known volume[42] **(Supplementary Fig. 27)**. On average, models were found to underestimate nodule volume by 0.4cc, or 12% of known volume. Three previously published models also showed a similar trend when tested on the same data[41] **(Supplementary Fig. 28)**.

Finally, model failure modes were examined **(Fig. 3)** through review by clinical experts. Examples of these included missing thoracic nodal stations originally undersampled in the discovery data e.g. supraclavicular nodes **(Supplementary Fig. 42)**, over-segmentation into pericardium and collapsed lungs, as well as susceptibility to motion artifacts around the diaphragm.

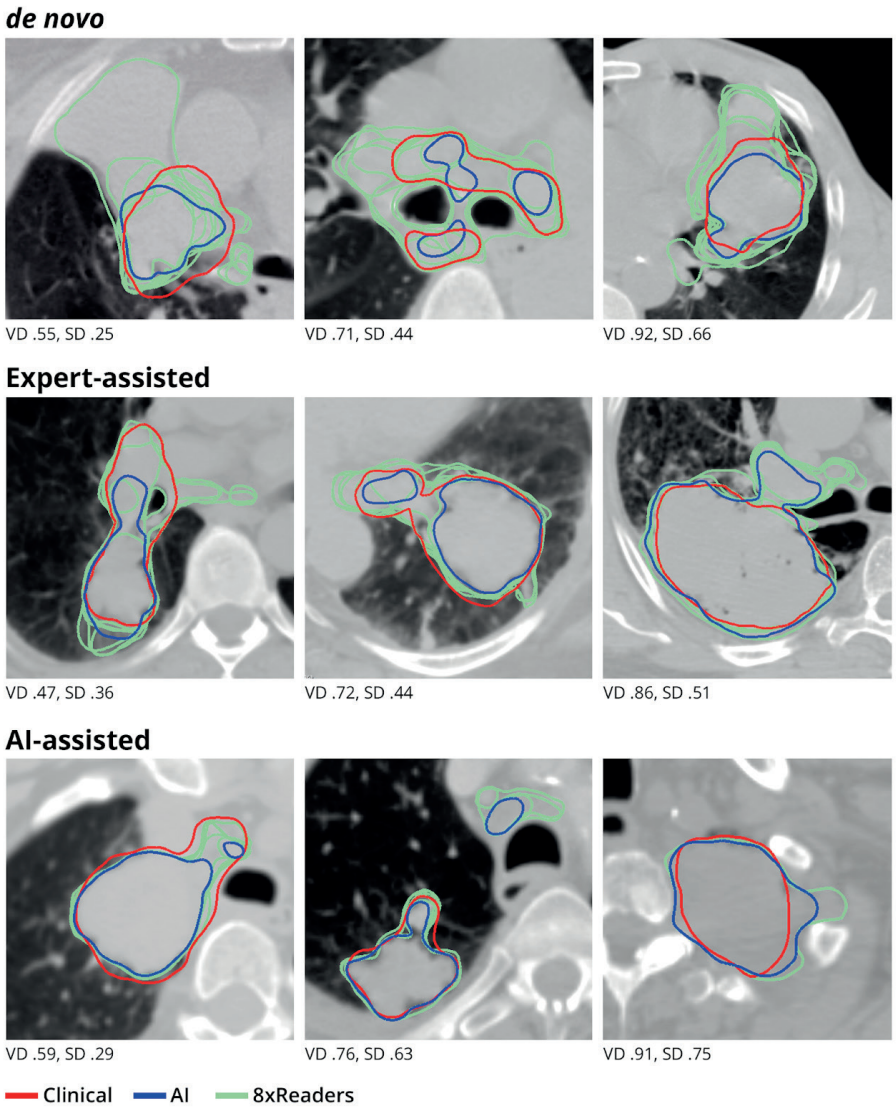

Figure 4. Nine representative examples from the end-user testing.
Nine examples from the de novo, expert-assisted (clinical segmentation provided), and AI-assisted (AI-generated segmentation provided) segmentations (three examples each).

End-user testing

We conducted end-user testing where eight readers were asked to perform the segmentation task *de novo*, or rate and edit a provided segmentation blinded to its source (Fig. 4, Methods). Provided segmentations were either clinical (expert-assisted) or AI-generated (AI-assisted). Clinical (RTOG-0617 clinical trial), AI, and our readers' segmentation

are shown in red, blue, and green respectively. Depicted scores are calculated between clinical and AI segmentations. VD=volumetric dice, SD=surface dice.

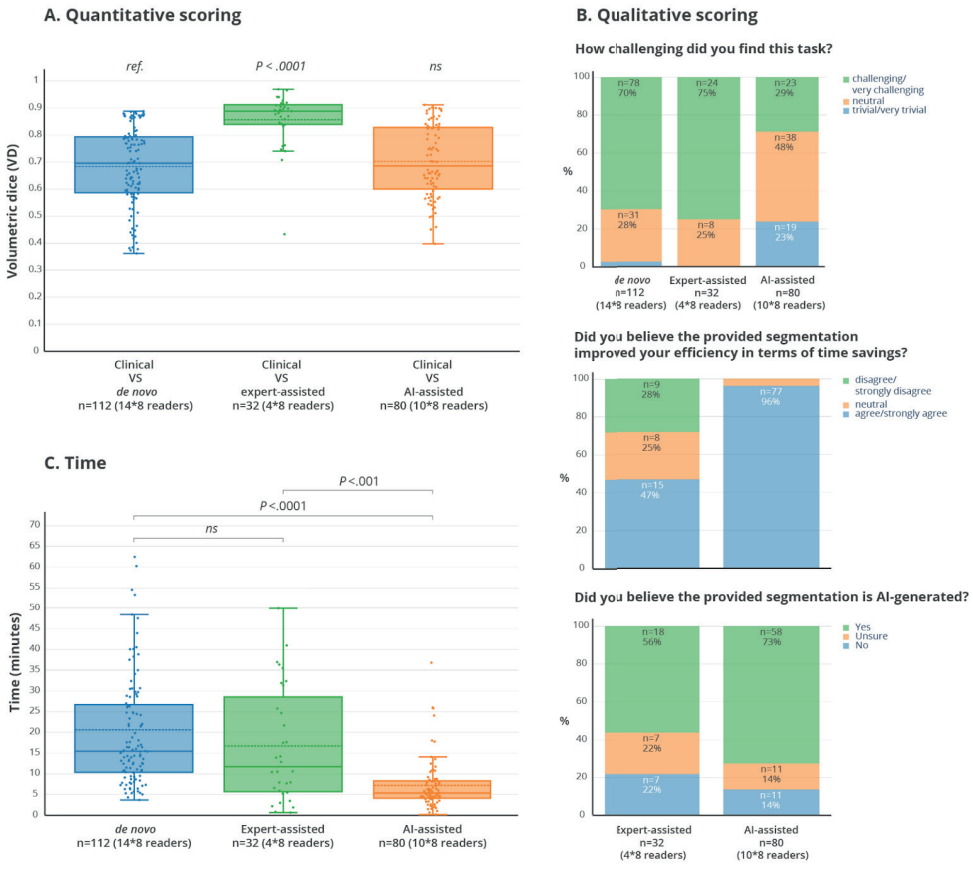

Figure 5. Results from the end-user testing.
Panel A reports the volumetric dice score between clinical trial segmentations and each of *de novo*, expert-assisted (clinical segmentation provided), and AI-assisted (AI-generated segmentation provided) segmentations. See Supplementary Fig. 29 for surface dice. Panel B reports answers to qualitative questions asked to readers during the end-user testing. Panel C reports the time needed to complete the segmentation task. The Mann-Whitney U rank test was used, with a two-tailed $P < .05$ indicating significance.

Using clinical segmentations as reference, we found non-significant differences between *de novo* VD 0.7 [0.64,0.75]; SD 0.43 [0.4,0.48] and AI-assisted VD 0.69 [0.65,0.75]; SD 0.38 [0.36,0.53] segmentations **(Fig. 5A, Supplementary Fig. 29)**. Similar results were obtained across individual readers **(Supplementary Fig. 30 and 31)**. When compared to the *de novo* median segmentation time of 15.5 minutes, expert-assistance led to an non-significant 24% reduction (11.7 minutes) while AI-assistance led to a significant 65% reduction (5.4 minutes, $P < .0001$, Fig. 5C). Non-significant differences were detected between *de novo* segmentations by residents and attendings **(Supplementary Fig. 32)**. When compared to the *de novo* interquartile range of interobserver variability, AI-assistance led to a non-significant 53% reduction for VD and a significant 32% reduction for SD ($P < .05$, Supplementary Fig. 33 and 34).

Qualitative data were also collected during the testing. For 96% of AI segmentations, readers agreed that the provided segmentations improved their efficiency. 74% of AI segmentations failed a Turing test-like setup as they were identified as being AI-generated **(Fig. 5B)**. 79% of AI segmentations were rated as "acceptable with minor modifications" by readers. Finally, we found that VD and SD metrics did not correlate with the time required to edit AI segmentations, nor did they significantly stratify subgroups based on qualitative measures including segmentation rating and perceived task difficulty **(Fig. 6A)**.

Discussion

In this study, we developed a multifaceted strategy for the clinical validation of DL models for RT targeting. Beyond establishing inter- and intraobserver benchmarks, we performed multi-tiered validation on internal and external datasets including clinical trial and diagnostic radiology data. We also carried out additional dosimetric validation and measured the models' stability and accuracy. Finally, we conducted end-user testing to measure clinical utility and physician acceptance.

Clinical validation strategy

Our validation strategy is aimed at closing the translational gap that falls in between early *in silico* validation and larger scale prospective clinical trials[44,45]. This strategy may provide the high levels of confidence needed to pursue AI clinical trials in medicine[46], uncover model weaknesses that would have been otherwise overlooked, generate preliminary data on human factors given our incomplete understanding of this area[47], as well as help quantify the time and effort needed to bring AI outputs to clinically acceptable levels.

Herein, we demonstrated the application of this strategy across four components (Fig. 1). *A. Benchmarks: Developing clinical benchmarks to understand the current standard of care.* This is reflected in our work on quantifying inter and intraobserver variability. *B. Primary Validation: Validation in large external cohorts to understand models' generalizability profile.* We conducted our study using large heterogeneous multi-institutional data (n=2208) from across multiple geographies. 60% (n=1320) of our data is publicly

available[48], allowing for future improvements on the same data. In addition to testing on pretreatment CT used in RT planning, we also opted to test on diagnostic CT to better understand the models' utility in different clinical contexts. *C. Functional Validation: Studying the models' impact on related clinical tasks and downstream clinical endpoints.* This is reflected in our experiments to better understand the dosimetric impact of AI-generated segmentations, and measure model stability and accuracy. Finally, *D. End-user testing in simulated clinical settings beyond in silico testing.* Our end-user tests aimed at understanding real-world clinical performance of the model, physicians' interactions with AI outputs, as well as overall satisfaction. This evaluation of human-machine interaction highlights the importance of studying clinical AI models under their intended use in the clinic, which is most commonly decision support and not full automation[49].

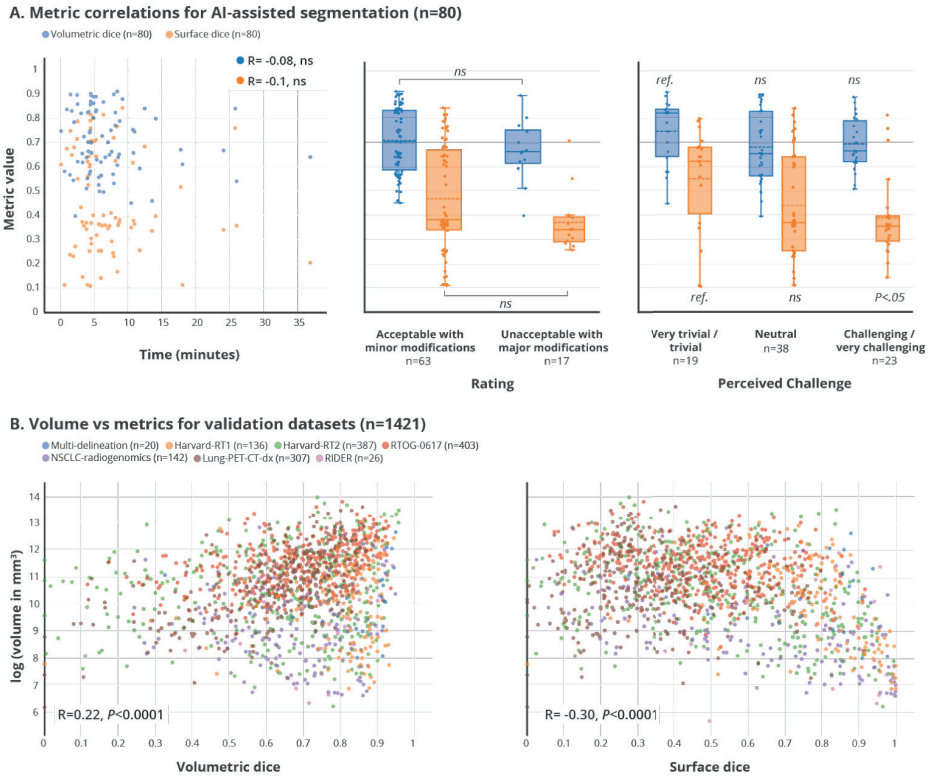

Figure 6. Analysis of segmentation metrics.
Panel A reports the correlation of segmentation metrics (volumetric dice in blue and surface dice in orange) with time needed to edit AI segmentations, qualitative rating provided by readers, as well as the perceived challenge. Data used for this analysis is from the end-user testing when an AI segmentation is provided to readers (n=80, 10 cases * 8 readers). The Mann-Whitney U rank test was used, with a two-tailed $P < .05$ indicating significance. Panel B reports the correlation of segmentation metrics (left, volumetric dice and right, surface dice) with tumor volume (displayed on log scale). We used all validation datasets for this analysis (n=1421). For comparisons between tumor volume and other metrics (precision, recall, jaccard, and segmentation score), see Supplementary Fig. 38. For correlation results on a per dataset basis, see **Supplementary Table 5.** The Spearman rank-order correlation coefficient was used for measuring correlation between the two groups.

This effort builds upon a large body of DL applications in medical imaging generally[19], and RT specifically[6]. Image segmentation is a major component of the RT treatment planning workflow, and deep learning has shown superior performance over prior state of the art methods[50,51]. Recent studies have described DL systems for organ segmentation in head & neck[52,53], renal[54], thoracic[54], prostate[55], and liver[56] cancer RT patients. Other efforts have explored tumor segmentation for nasopharyngeal[57], lung[8], and oropharyngeal[58] cancers. Most studies remain at the proof of concept stage, often validated in <100 samples, and lack external validation. Additionally, most are confined to *in silico* testing with only a few evaluating model performance in clinical settings[22,23,55].

Benchmarks, validation, and end-user testing

Our benchmarking results underscore the model's ability in identifying challenging cases with large interobserver variability. While our study showed no difference in performance between residents and attending physicians, further work is needed to understand the impact of training level on human-AI interaction and the potential of such models to augment physician training and standardize RT planning.

Testing on multiple external datasets of varying characteristics is crucial in understanding a given model's generalizability profile. Our tiered validation process starts with a single-reader internal test data (Harvard-RT1) that most resembles the training data, and expands to multi-reader clinical trial data as well as diagnostic radiology data. The performance drop at the multi-reader internal test data (Harvard-RT2) in the context of its relative stability on subsequent increasingly external datasets suggest that segmentation variability may be a function of treating physician style or preference. Results on 4DCT data within the Harvard-RT2 dataset imply the models' relevance toward modern imaging practices.

Studies have shown significant differences in protocol compliances of target and organ segmentation in RT clinical trials[59]. Results on the RTOG-0617 clinical trial dataset suggest that automated target segmentation models may be used as quality assurance tools for these trials. AI Models can detect and flag subpar segmentations during the trial, thereby acting as a triage mechanism for the time-consuming and expensive human peer review[60].

Results on the diagnostic datasets highlight known differences between radiologists (anatomical knowledge) and radiation oncologists (therapeutic goals) in defining tumor boundaries **(Supplementary Fig. 16 and 35)**. These differences are likely synergistic, and emphasize the importance of radiologist input in RT planning[61,62], especially when professional overlap is inevitable e.g. post-operative RT[63]. This also stimulates further discussions around the "off-label" use of AI where applications developed within one speciality are deployed in another.

In terms of metrics that best assess AI-generated segmentations, there is no consensus thus far[64]. Our functional validation studies underscore the importance of evaluating AI-generated RT segmentations beyond the common scope of geometric measures[65]. Similar to prior studies[66,67], our dosimetric analysis showed no correlation between

geometric and dosimetric measures **(Supplementary Fig. 36)**. We also found that geometric measures may fail to accurately mirror time savings and other qualitative measures **(Fig. 6A)**. Our results also highlight undesired correlations between metrics and tumor volume **(Fig. 6B, Supplementary Fig. 37)**, and echo VD's bias towards larger tumors[52,68]. As VD is the most common metric used in medical image segmentation[69], there is an unmet need for new metrics that combine qualitative physician evaluation with geometric, dosimetric, and time-related measures to accurately reflect meaningful clinical outcomes[64,70,71].

A closer look into model failure modes **(Fig. 3)** may help guide implementation. Such modes may be automatically detected and flagged to the physician together with a warning that model outputs may be compromised, thereby bringing much needed trust into automated systems[72]. The co-development of both assisted and automated models provides the flexibility to address a variety of clinical scenarios. In terms of tumor localization failures, our models failed in 87 (6%) of the 1421 validation cases **(Supplementary Table 1)**, thereby requiring fall back onto the seed point assisted models. Alternatively, future models may be augmented through the automated extraction of rough anatomic tumor location from other existing RT data sources such as clinical notes[73].

The exact effects of imaging contrast on model performance remain unclear. This is especially true as our models significantly over-performed on contrast enhanced images (VD $P < .05$; SD $P < .01$, Supplementary Fig. 41), despite being trained primarily on non-contrast data (Supplementary Fig. 40). Finally, further work on understanding models' accuracy is needed, especially given known sensitivities to imaging parameters including CT slice thickness and reconstruction algorithms[42,74].

Limitations

Several limitations should be noted. Both our *in silico* and end-user testing are limited by their retrospective nature. Much of our discovery data relied on a single human expert. While this enabled us to highlight the model's ability to encapsulate the skills of a given expert, it also implies our models may have acquired a natural bias. Our dosimetric analysis may not always reflect clinical reality as it does not allow for manually editing the intermediate volume between GTV and PTV, namely the clinical target volume (CTV). The design of our end-user tests did not allow for testing AI effects on intraobserver variability, nor did it did not incorporate PET imaging - a commonly used modality in guiding RT planning for NSCLC patients[75]. Additionally, while blinding readers to the source of the provided segmentations allowed for a more fair evaluation of the AI model, this design did not allow for testing human bias towards a clinical AI algorithm[76]. Such a bias may have ramifications for real-world adoption and use.

Implications for Cancer Care

In addition to assisting day-to-day RT planning, automated segmentation may also help propel more modern RT practices. Specifically, image-guided adaptive RT may benefit from automated continuous target segmentation to account for anatomical variations during irradiation[77]. Such tools may also enable RT applications in the global health

context. As expert human knowledge is embedded into AI tools they have the potential to transcend borders and bring much needed expertise to medically underserved communities suffering from health professional shortages[78]. Beyond RT, cancer imaging in clinical trials may also benefit from automated tumor segmentation for consistent volumetric response assessment beyond current 2D methods (e.g. RECIST)[79], as well as for developing volumetric imaging biomarkers[80,81].

Conclusion

Early testing of AI tools in clinical environments is crucial for translation to clinic. Our four-component validation strategy allows for uncovering downstream consequences of clinical AI implementation that may otherwise go unnoticed in typical *in silico* validation studies. We encourage the broader adoption of similar validation strategies that help close the translational gap for clinical AI applications.

Acknowledgements

Authors acknowledge financial support from the U.S. National Institutes of Health (U24CA194354, U01CA190234, and U01CA209414); https://grants.nih.gov/funding/index.htm. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank the patients whose imaging data were used in this study. We thank the team behind The Cancer Imaging Archive (TCIA) for hosting all the publicly available data used in this study. We thank Dan Darkow and Joe Meyers of MIM Software for their help in setting up workflows to conduct the end-user testing, together with Alex Cruz of our institution's information systems. This manuscript was prepared using data from Datasets NCT00533949-D1/D2/D3 from the NCTN Data Archive of the National Cancer Institute's (NCI's) National Clinical Trials Network (NCTN). Data were originally collected from clinical trial NCT number NCT00533949 "High-Dose or Standard-Dose Radiation Therapy and Chemotherapy With or Without Cetuximab in Treating Patients With Newly Diagnosed Stage III Non-Small Cell Lung Cancer That Cannot Be Removed by Surgery". 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, or the NCI.

Supporting Information

https://docs.google.com/document/d/11JBbS972zqVnEKHXH2bjFJv6LKUnUXeXJ6GEgCaMiF8/edit?usp=sharing

https://docs.google.com/document/d/1p4f4k_KhzsJ9CGjn8csoM47axzTISt jSjAHZPiUEeI/edit?usp=sharing

https://docs.google.com/document/d/1rrxuNAEQEHNMA3pt4LvhH3nTdxPHx2FtUallrhAo04Q/edit?usp=sharing

References

1. Torre, L. A. *et al.* Global cancer statistics, 2012. *CA Cancer J. Clin.* **65**, 87–108 (2015).
2. Society, A. C. Cancer facts & figures. *American Cancer Society* (2016).
3. Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E. & Adjei, A. A. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin. Proc.* **83**, 584–594 (2008).
4. Miller, K. D. *et al.* Cancer treatment and survivorship statistics, 2016. *CA Cancer J. Clin.* **66**, 271–289 (2016).
5. Jumeau, R., Vilotte, F., Durham, A.-D. & Ozsahin, E.-M. Current landscape of palliative radiotherapy for non-small-cell lung cancer. *Transl Lung Cancer Res* **8**, S192–S201 (2019).
6. Huynh, E. *et al.* Artificial intelligence in radiation oncology. *Nat. Rev. Clin. Oncol.* **17**, 771–781 (2020).
7. Vorwerk, H. *et al.* Protection of quality and innovation in radiation oncology: The prospective multicenter trial the German Society of Radiation Oncology (DEGRO-QUIRO study). *Strahlenther. Onkol.* **190**, 433–443 (2014).
8. Mak, R. H. *et al.* Use of Crowd Innovation to Develop an Artificial Intelligence-Based Solution for Radiation Therapy Targeting. *JAMA Oncol* **5**, 654–661 (2019).
9. Chan, C. *et al.* Intensity-modulated radiotherapy for lung cancer: current status and future developments. *J. Thorac. Oncol.* **9**, 1598–1608 (2014).
10. Van de Steene, J. *et al.* Definition of gross tumor volume in lung cancer: inter-observer variability. *Radiother. Oncol.* **62**, 37–49 (2002).
11. Cui, Y. *et al.* Contouring variations and the role of atlas in non-small cell lung cancer radiation therapy: Analysis of a multi-institutional preclinical trial planning study. *Pract. Radiat. Oncol.* **5**, e67–e75 (2015).
12. Fotina, I., Lütgendorf-Caucig, C., Stock, M., Pötter, R. & Georg, D. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. *Strahlenther. Onkol.* **188**, 160–167 (2012).
13. Ohri, N. *et al.* Radiotherapy protocol deviations and clinical outcomes: A meta-analysis of cooperative group clinical trials. *Journal of Clinical Oncology* vol. 30 181–181 (2012).
14. Peters, L. J. *et al.* Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. *J. Clin. Oncol.* **28**, 2996–3001 (2010).
15. Eaton, B. R. *et al.* Institutional Enrollment and Survival Among NSCLC Patients Receiving Chemoradiation: NRG Oncology Radiation Therapy Oncology Group (RTOG) 0617. *J. Natl. Cancer Inst.* **108**, (2016).
16. Delpon, G. *et al.* Comparison of Automated Atlas-Based Segmentation Software for Postoperative Prostate Cancer Radiotherapy. *Front. Oncol.* **6**, 178 (2016).
17. Ciardo, D. *et al.* Atlas-based segmentation in breast cancer radiotherapy: Evaluation of specific and generic-purpose atlases. *Breast* **32**, 44–52 (2017).
18. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).

19. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).

20. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).

21. Sheng, K. Artificial intelligence in radiotherapy: a technological review. *Front. Med.* **14**, 431–449 (2020).

22. Lin, L. *et al.* Deep Learning for Automated Contouring of Primary Tumor Volumes by MRI for Nasopharyngeal Carcinoma. *Radiology* **291**, 677–686 (2019).

23. Bi, N. *et al.* Deep Learning Improved Clinical Target Volume Contouring Quality and Efficiency for Postoperative Radiation Therapy in Non-small Cell Lung Cancer. *Front. Oncol.* **9**, 1192 (2019).

24. Ye, Z. *et al.* Deep learning-based detection of intravenous contrast in computed tomography scans. *arXiv [eess.IV]* (2021).

25. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241 (Springer, Cham, 2015).

26. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* 424–432 (Springer International Publishing, 2016).

27. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv [cs.CV]* (2016).

28. Maas, A. L., Hannun, A. Y., Ng, A. Y. & Others. Rectifier nonlinearities improve neural network acoustic models. in *Proc. icml* vol. 30 3 (Citeseer, 2013).

29. Mishra, P. & Sarawadekar, K. Polynomial Learning Rate Policy with Warm Restart for Deep Neural Network. in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)* 2087–2092 (2019).

30. Jadon, S. A survey of loss functions for semantic segmentation. in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* 1–7 (2020).

31. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) vol. 32 (Curran Associates, Inc., 2019).

32. van Baardwijk, A. *et al.* PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. *Int. J. Radiat. Oncol. Biol. Phys.* **68**, 771–778 (2007).

33. High-Dose or Standard-Dose Radiation Therapy and Chemotherapy With or Without Cetuximab in Treating Patients With Newly Diagnosed Stage III Non-Small Cell Lung Cancer That Cannot Be Removed by Surgery. https://clinicaltrials.gov/ct2/show/NCT00533949.

34. NSCLC-Cetuximab (RTOG-0617). https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=33948334.

35. Bakr, S. *et al.* A radiogenomic dataset of non-small cell lung cancer. *Sci Data* **5**, 180202 (2018).

36. A large-scale CT and PET/CT dataset for lung cancer diagnosis (lung-PET-CT-dx) - the cancer imaging archive (TCIA) public access - cancer imaging archive wiki. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70224216.

37. Bradley, J. D. *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.* **16**, 187–199 (2015).

38. Kikinis, R., Pieper, S. D. & Vosburgh, K. G. 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. in *Intraoperative Imaging and Image-Guided Therapy* (ed. Jolesz, F. A.) 277–289 (Springer New York, 2014).

39. Pinter, C. *et al.* Performing radiation therapy research using the open-source SlicerRT toolkit. in *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada* 622–625 (Springer International Publishing, 2015).

40. Zhao, B. *et al.* Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non–Small Cell Lung Cancer. *Radiology* **252**, 263–272 (2009).

41. Kalpathy-Cramer, J. *et al.* A Comparison of Lung Nodule Segmentation Algorithms: Methods and Results from a Multi-institutional Study. *J. Digit. Imaging* **29**, 476–487 (2016).

42. Zhao, B., Tan, Y., Tsai, W. Y., Schwartz, L. H. & Lu, L. Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study. *Transl. Oncol.* **7**, 88–93 (2014).

43. High-Dose or Standard-Dose Radiation Therapy and Chemotherapy With or Without Cetuximab in Treating Patients With Newly Diagnosed Stage III Non-Small Cell Lung Cancer That Cannot Be Removed by Surgery. https://clinicaltrials.gov/ct2/show/NCT00533949.

44. Group, T. D.-A. S. & The DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nature Medicine* vol. 27 186–187 (2021).

45. Kann, B. H., Hosny, A. & Aerts, H. J. W. L. Artificial intelligence for clinical oncology. *Cancer Cell* **39**, 916–927 (2021).

46. Topol, E. J. Welcoming new guidelines for AI clinical research. *Nat. Med.* **26**, 1318–1320 (2020).

47. Sujan, M. *et al.* Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inform* **26**, (2019).

48. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).

49. Bitterman, D. S., Aerts, H. J. W. L. & Mak, R. H. Approaching autonomy in medical artificial intelligence. *Lancet Digit Health* **2**, e447–e449 (2020).

50. Chen, W. *et al.* Deep learning vs. atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images. *Radiat. Oncol.* **15**, 176 (2020).

51. Ahn, S. H. *et al.* Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer. *Radiat. Oncol.* **14**, 213 (2019).

52. Nikolov, S. *et al.* Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. *J. Med. Internet Res.* **23**, e26151 (2021).

53. Ibragimov, B. & Xing, L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med. Phys.* **44**, 547–557 (2017).

54. Lustberg, T. *et al.* Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother. Oncol.* **126**, 312–317 (2018).

55. Oktay, O. *et al.* Evaluation of Deep Learning to Augment Image-Guided Radiotherapy for Head and Neck and Prostate Cancers. *JAMA Netw Open* **3**, e2027426 (2020).

56. Ibragimov, B., Toesca, D., Chang, D., Koong, A. & Xing, L. Combining deep learning with anatomical analysis for segmentation of the portal vein for liver SBRT planning. *Phys. Med. Biol.* **62**, 8943–8958 (2017).

57. Men, K. *et al.* Deep Deconvolutional Neural Network for Target Segmentation of Nasopharyngeal Cancer in Planning Computed Tomography Images. *Front. Oncol.* **7**, 315 (2017).

58. Cardenas, C. E. *et al.* Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function. *Int. J. Radiat. Oncol. Biol. Phys.* **101**, 468–478 (2018).

59. Tsang, Y. *et al.* Assessment of contour variability in target volumes and organs at risk in lung cancer radiotherapy. *Tech Innov Patient Support Radiat Oncol* **10**, 8–12 (2019).

60. Men, K., Geng, H., Biswas, T., Liao, Z. & Xiao, Y. Quality Assurance of Contouring for NRG Oncology/RTOG 1308 Clinical Trial Based on Automated Segmentation with Deep Active Learning. *Int. J. Radiat. Oncol. Biol. Phys.* **105**, S22–S23 (2019).

61. Braunstein, S., Glastonbury, C. M., Chen, J., Quivey, J. M. & Yom, S. S. Impact of Neuroradiology-Based Peer Review on Head and Neck Radiotherapy Target Delineation. *AJNR Am. J. Neuroradiol.* **38**, 146–153 (2017).

62. Burnet, N. G., Thomas, S. J., Burton, K. E. & Jefferies, S. J. Defining the tumour and target volumes for radiotherapy. *Cancer Imaging* **4**, 153–161 (2004).

63. Ng, S. P. *et al.* A prospective in silico analysis of interdisciplinary and interobserver spatial variability in post-operative target delineation of high-risk oral cavity cancers: Does physician specialty matter? *Clin Transl Radiat Oncol* **12**, 40–46 (2018).

64. Sherer, M. V. *et al.* Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother. Oncol.* **160**, 185–191 (2021).

65. Vinod, S. K., Jameson, M. G., Min, M. & Holloway, L. C. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiother. Oncol.* **121**, 169–179 (2016).

66. Tsuji, S. Y. *et al.* Dosimetric evaluation of automatic segmentation for adaptive IMRT for head-and-neck cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **77**, 707–714 (2010).

67. Cao, M. *et al.* Analysis of Geometric Performance and Dosimetric Impact of Using Automatic Contour Segmentation for Radiotherapy Planning. *Front. Oncol.* **10**, 1762 (2020).

68. van der Veen, J., Gulyban, A., Willems, S., Maes, F. & Nuyts, S. Interobserver variability in organ at risk delineation in head and neck cancer. *Radiat. Oncol.* **16**, 120 (2021).

69. Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* **15**, 29 (2015).

70. Zhu, J. *et al.* Evaluation of Automatic Segmentation Model With Dosimetric Metrics for Radiotherapy of Esophageal Cancer. *Front. Oncol.* **10**, 564737 (2020).

71. Gooding, M. J. *et al.* Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test. *Med. Phys.* **45**, 5105–5115 (2018).

72. Asan, O., Bayrak, A. E. & Choudhury, A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J. Med. Internet Res.* **22**, e15154 (2020).

73. Si, Y. & Roberts, K. A Frame-Based NLP System for Cancer-Related Information Extraction. *AMIA Annu. Symp. Proc.* **2018**, 1524–1533 (2018).

74. Brendle, C. *et al.* Is the standard uptake value (SUV) appropriate for quantification in clinical PET imaging? - Variability induced by different SUV measurements and varying reconstruction methods. *Eur. J. Radiol.* **84**, 158–162 (2015).

75. Grootjans, W. *et al.* PET in the management of locally advanced and metastatic NSCLC. *Nat. Rev. Clin. Oncol.* **12**, 395–407 (2015).

76. Gaube, S. *et al.* Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med* **4**, 31 (2021).

77. Kong, V., Wenz, J., Craig, T. & Milosevic, M. Image-Guided Adaptive Radiotherapy – Delivering Personalized Radiation Medicine to Improve Treatment Quality and Patients' Outcome. *Journal of Medical Imaging and Radiation Sciences* **44**, 55–56 (2013).

78. Hosny, A. & Hugo J W. Artificial intelligence for global health. *Science* **366**, 955–956 (2019).

79. Welsh, J. L. *et al.* Comparison of response evaluation criteria in solid tumors with volumetric measurements for estimation of tumor burden in pancreatic adenocarcinoma and hepatocellular carcinoma. *Am. J. Surg.* **204**, 580–585 (2012).

80. Buckler, A. J. *et al.* The use of volumetric CT as an imaging biomarker in lung cancer. *Acad. Radiol.* **17**, 100–106 (2010).

81. Dercle, L. *et al.* Vol-PACT: A Foundation for the NIH Public-Private Partnership That Supports Sharing of Clinical Trial Data for the Development of Improved Imaging Biomarkers in Oncology. *JCO Clin Cancer Inform* **2**, 1–12 (2018).

III

PART III

AI Methods and Best Practices

9

Chapter 9

Handcrafted Versus Deep Learning Radiomics for Prediction of Cancer Therapy Response

A Hosny, HJWL Aerts & RH Mak

The Lancet Digital Health 2019

Commentary

In The Lancet Digital Health, Bin Lou and colleagues[1] apply deep learning methods to analyse pre-treatment CT scans in a retrospective cohort study of 944 patients (849 in the internal study cohort and 95 in the independent validation cohort) treated with stereotactic body radiation therapy, a form of high-dose, pinpoint radiation therapy for lung tumours. The study presents a novel analysis by integrating traditional radiomics features through multi-task learning, applying a time-based survival analysis, and incorporating new deep learning methods including a three-dimensional (3D) convolutional neural network to analyse lung tumours before treatment. The authors input pre-therapy lung CT images into Deep Profiler, a multi-task deep neural network that has radiomics incorporated into the training signal. They combined these data with clinical variables to derive iGray, an individualised radiation dose that estimates the probability of treatment failure to be below 5%. Models that included Deep Profiler and clinical variables predicted treatment failures with a concordance index of 0·72 (95% CI 0·67–0·77), a significant improvement compared with traditional radiomics ($p<0·0001$) or clinical variables ($p<0·0001$) alone. The potential clinical applications of such models include identifying tumours at the highest risk of resistance to radiation therapy, and personalised dosing of radiation therapy to maximise likelihood of tumour control.

This study is representative of a major turning point in the underlying radiomics methodologies used in treatment response prediction and prognosis, specifically in radiation therapy with broader implications across other cancer therapies. Traditional radiomics makes use of handcrafted features and has been studied extensively as an imaging biomarker to predict cancer outcomes and responses to therapy[2,3]. The handcrafted radiomics approach involves manual segmentation of the region of interest (eg, the tumour) on medical imaging, and extraction of thousands of human-defined and curated quantitative features from the region of interest, which describe tumour shape and texture among other characteristics. In the final step, the approach involves application of machine learning methods to identify the imaging features that are associated with a given clinical endpoint. However, the human-derived nature of traditional radiomics methods has been criticised for introducing a source of human bias into the process[4]; there have been concerns of reproducibility[5] due to the intra-reader and inter-reader variability that results from the reliance on manual segmentation of the tumour, and due to variation in imaging and pre-processing techniques for feature extraction. Moreover, the value of traditional radiomics has recently come under question with the advent of deep learning methods and consequent proof-of-principle applications in predicting cancer outcomes[6,7]. For many of the deep learning radiomics applications, region of interest definition is based on a single point placement within the tumour volume, essentially replacing full tumour segmentations with approximate localisation and minimising the need for human input. Additionally, deep learning methods allow for automated learning of relevant radiographic features without the need for previous definition by researchers. In turn, these abstract representations have enabled a larger learning capacity, boosting generalisability and accuracy while reducing potential bias[8].

Some key caveats remain for clinical use of the deep learning model proposed by Lou and colleagues[1]. Firstly, the radiation dose delivered via stereotactic body radiation therapy for lung cancers represents the upper limit of what can be safely delivered to treat cancer in the human body with current technological capabilities. Of note, other tumours are often treated at substantially lower biological doses, and this study does not capture that range of radiation dose and tumour response curves. Secondly, radiation regimens used for stereotactic body radiation therapy are typically achievable only for localised (eg, stage I lung cancers) and small tumours (e.g. <5 cm diameter), and thus these dose predictions are not easily generalisable to more advanced tumours. Lastly, the model is built on a relatively rare event (3-year cumulative incidence of local failure was 13·5% in the overall population) which is an advantage to patients because it means stereotactic body radiation therapy works well, but a disadvantage for predictive model building because of the increased risk of over-fitting.

In this study, the authors chose to identify handcrafted radiomics features as ground truth while comparing them to features identified by deep learning methods. The level of agreement between these two sets of features was then used as a cost function to train and optimise the predictive model. This method was understandably chosen as a means to provide a connection to the previous traditional radiomics landscape and greater interpretability. However, we believe that deep learning can emerge as an independent methodology that does not need to rely on handcrafted radiomics to move forward. Combining traditional radiomic features into deep learning models risks incorporating the aforementioned known human biases into the model. Additionally, a combined approach does not address the interpretability problem since even most mathematically-derived handcrafted features capture uninterpretable imaging characteristics that cannot be discerned by the human eye. Nevertheless, the challenges of traditional radiomics approaches such as lack of reproducibility and interpretability as well as over-fitting on small datasets will only be amplified in deep learning-driven prediction models of cancer outcome. Fortunately, interpretability of features learned through neural networks is an active area of research[9], while sharing and transparency initiatives are paving the way for larger curated cancer imaging repositories[10].

Deep learning may also allow the decoding of new insights from cancer images and non-intuitive information that is uncharted thus far. We look with great interest at the saliency mapping in figure 5 of the Article, which identifies the regions of the CT scan in and around the tumour that are most associated with the predicted outcome of local tumour recurrence. Our group identified similar peri-tumoural localisation when performing activation mapping for a 3D convolutional neural network trained for a prognostication task in non-small lung cancer patients, which suggests potentially important imaging characteristics at the cancer-normal tissue interface[7]. Although these findings are preliminary and qualitative in nature, future work to understand the biology of this interface in relation to cancer therapy response prediction, and perhaps more importantly applying deep learning radiomics to target localised cancer therapies such as radiation therapy and surgery, represent a truly exciting new frontier of cancer care.

References

1. Lou, B. *et al.* An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction. *The Lancet Digital Health* **1**, e136–e147 (2019).
2. Coroller, T. P. *et al.* CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother. Oncol.* **114**, 345–350 (2015).
3. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).
4. Chalkidou, A., O'Doherty, M. J. & Marsden, P. K. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. *PLoS One* **10**, e0124165 (2015).
5. Traverso, A., Wee, L., Dekker, A. & Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int. J. Radiat. Oncol. Biol. Phys.* **102**, 1143–1158 (2018).
6. Xu, Y. *et al.* Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clin. Cancer Res.* **25**, 3266–3275 (2019).
7. Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **15**, e1002711 (2018).
8. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
9. Chakraborty, S. *et al.* Interpretability of deep learning models: A survey of results. in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* 1–6 (IEEE, 2017).
10. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).

10

Chapter 10

The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards

*A Hosny**, S Holland*, S Newman, J Joseph & K Chmielinski

Data Protection and Privacy: Data Protection and Democracy 2020

Abstract

Data is a fundamental ingredient in building Artificial Intelligence (AI) models, and there are direct correlations between data quality and model robustness, fairness, and utility. A growing body of research points to AI systems deployed in a wide range of use cases, where algorithms trained on biased, incomplete, or ill-fitting data produce problematic results. Despite the increased critical attention, data interrogation continues to be a challenging task with many issues being difficult to identify and rectify. Algorithms often come under scrutiny only after they are developed and deployed, which exacerbates this problem and underscores the need for better data vetting practices earlier in the development pipeline. We introduce the Dataset Nutrition Label (the Label), a diagnostic framework providing a distilled yet comprehensive overview of dataset "ingredients". The label is designed to be flexible and adaptable; it comprises a diverse set of qualitative and quantitative modules generated through multiple statistical and probabilistic modelling backends. Working with the ProPublica dataset "Dollars for Docs", we developed an open source tool consisting of seven sample modules. Consulting such a label prior to AI model development promotes vigorous data interrogation practices, aids in recognizing inconsistencies and imbalances, provides an improved means to selecting more appropriate datasets for specific tasks, and subsequently increases the overall quality of AI models. We also explore some challenges of the Label, including generalizing across diverse datasets, as well as discuss research and public policy agendas to further advocate its adoption and ultimately improve the AI development ecosystem.

Introduction

Data driven decision making systems play an increasingly important role in our lives. These frameworks are built on increasingly sophisticated artificial intelligence (AI) systems and are created and tuned by a growing population of data specialists[1] to arrive at a diversity of decisions: from movie and music recommendations to digital advertisements and mortgage applications[1]. These systems deliver untold societal and economic benefits, but they can also be harmful to individuals and society at large.

Data is a fundamental ingredient of AI, and the quality of a dataset used to build a model will directly influence the outcomes it produces. An AI model trained on problematic data will likely produce problematic outcomes. Examples of these include gender bias in language translations surfaced through natural language processing[2], and skin shade bias in facial recognition systems due to non-representative data[3]. Typically, the model development pipeline **(Figure 1)** begins with a question or goal. Within the realm of supervised learning, for instance, a data specialist will curate a labeled dataset of previous answers in response to the guiding question. Such data is then used to train a model to respond in a way that accurately correlates with past occurrences. In this way, past answers are used to forecast the future. This is particularly problematic when outcomes of past events are contaminated with (often unintentional) bias.

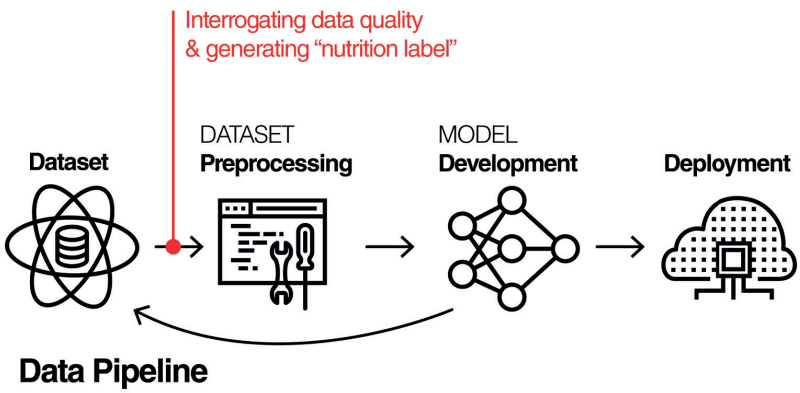

Figure 1. Model Development Pipeline

Models often come under scrutiny only after they are built, trained, and deployed. If a model is found to perpetuate a bias - for example, over-indexing for a particular race or gender - the data specialist returns to the development stage in order to identify and address the issue. This feedback loop is inefficient, costly, and does not always mitigate harm; the time and energy of the data specialist is a sunk cost, and if in use, the

1 The term "data specialist" is used instead of "data scientist" in the interest of using a general term that is broadly scoped to include all professionals utilizing data in automated decision making systems such as data scientists, data analysts, and artificial intelligence engineers and researchers. It also includes those who create, label, and receive datasets.

model deployment may have already produced problematic outcomes. Some of these issues could be avoided by undertaking thorough interrogation of data at the outset of model The term "data specialist" is used instead of "data scientist" in the interest of using a general term that is broadly scoped to include all professionals utilizing data in automated decision making systems such as data scientists, data analysts, and artificial intelligence engineers and researchers. It also includes those who create, label, and receive datasets.development. However, this is still not a widespread practice within AI model development efforts.

We conducted an anonymous online survey **(Figure 2)**, the results of which further lend credence to this problem. Although many (47%) respondents report conducting some form of data analysis prior to model development, most (74%) indicate that their organizations do not have explicit best practices for such analysis. 59% of respondents reported relying primarily on experience and self-directed learning (through online tutorials, blogs, academic papers, Stack Overflow, and online data competitions) to inform their data analysis methods and practices. This survey indicates that despite limited current standards, there is widespread interest to improve data analysis practices and make them accessible.

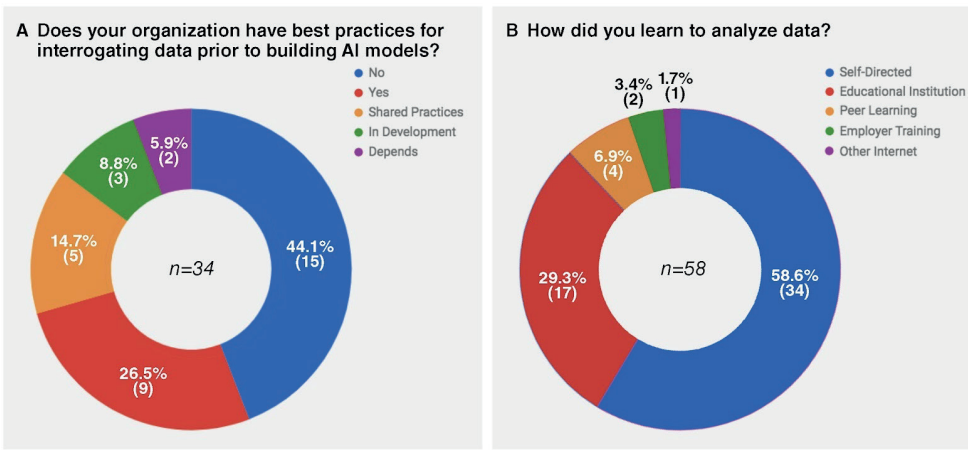

Figure 2. (A) Survey results about data analysis best practices in respondents' organizations and (B) Survey results about how respondents learned to analyze data

To improve the accuracy and fairness of AI systems, it is imperative that data specialists are able to more quickly assess the viability and fitness of datasets, and more easily find and use better quality data to train their models. As a proposed solution, we introduce the Dataset Nutrition Label, a diagnostic framework to address and mitigate some of these challenges by providing critical information to data specialists at the point of data analysis. The Label thus acts as a first point of contact where decisions regarding the utility and fitness of specific datasets can be made. This is achieved through allowing

the recognition of dataset inconsistencies and exclusions as well as promoting dataset interrogation as a crucial and inevitable procedure in the AI model development pipeline - with the ultimate goal of improving the overall quality of AI systems.

We begin with a review of related work, largely drawing from the fields of nutrition and privacy where labels are a useful mechanism to distill essential information, enable better decision-making, and influence best practices. We then discuss the Dataset Nutrition Label prototype, our methodology, demonstration dataset, and key results. This is followed by an overview of the benefits of the tool, its potential limitations, and ways to mitigate those limitations. We then briefly summarize some future directions, including research and public policy agendas that would further advance the goals of the Label. Lastly, we discuss implementation of the prototype and key takeaways.

Labels in Context and Related Work

To inform the development of our prototype and concept, we surveyed the literature for labeling efforts. Labels and warnings are utilized effectively in product safety[4], pharmaceuticals[5], energy[6], and material safety[7]. We largely draw from the fields of nutrition, online privacy, and algorithmic accountability as they are particularly salient for our purposes. The former is the canonical example and a long standing practice subject to significant study while the latter provides valuable insights in the application of a "nutrition label" in other domains, particularly in subjective contexts and where there is an absence of legal mandates and use is voluntary. Collectively, they elucidate the impacts of labels on audience engagement, education, and user decision making.

In 1990, Congress passed the Nutrition Labeling and Education Act (P.L. 101 - 535), which includes a requirement that certain foodstuffs display a standardized "Nutrition Facts" label[8]. By mandating the label, vital nutritional facts were communicated in the context of the "Daily Value" benchmark, and consumers could quickly assess nutrition information and more effectively abide by dietary recommendations at the moment of decision[8–10]. In the nearly three decades since its implementation, several studies have examined the efficacy of the now ubiquitous "Nutrition Facts" label; these studies include analyses of how consumers use the label[9,11], and the effect it has had on the market[12]. Though some cast doubt on the benefits of the mandate in light of its costs[13], most research concludes that the "Nutrition Facts" label has positive impact[14,15]. Surveys demonstrate widespread consumer awareness of the label, and its influence in decision making around food, despite a relatively short time since the passage of the Nutrition Labeling and Education Act[16]. According to the International Food Information Council, more than 80% of consumers reported they looked at the "Nutrition Facts" label when deciding what foods to purchase or consume, and only four percent reported never using the label[17]. Five years after the mandate, the Food Marketing Institute found that about one-third of consumers stopped buying food because of what they read on the label[16]. With regard to the information contained on the label and consumer understanding, researchers found that "label format and inclusion of (external) reference value information appear to have (positive) effects on consumer

perceptions and evaluations"[18] but consumers indicated confusion about the "Daily Value" comparison, suggesting that more information about the source and reliability of ground truth information would be useful[17]. The literature focuses primarily on the impact to consumers rather than on industry operations such as production and advertising. However, the significant impact of reported sales and marketing materials on consumers[12] provides a foundation for further inquiry into how this has affected the greater food industry.

In the field of privacy and privacy disclosures, the nutrition label serves as a useful point of reference and inspiration[19]. Researchers at Carnegie Mellon and Microsoft created the "Privacy Nutrition Label" to better surface essential privacy information to assist consumer decision making with regard to the collection, use, and sharing of personal information[20]. The "Privacy Nutrition Label" operates much like "Nutrition Facts" and sits atop existing disclosures. It improves the functionality of the Platform for Privacy Notices, a machine readable format developed by the World Wide Web Consortium, itself an effort to standardize and improve legibility of privacy policies[21]. User surveys that tested the "Privacy Nutrition Label" against alternative formats found that the label outperformed alternatives with "significant positive effects on the accuracy and speed of information finding and reader enjoyment with privacy policies," as well as improved consumer understanding[20,21].

Ranking and scoring algorithms also pose challenges in terms of their complexity, opacity, and sensitivity to the influence of data. End users and even model developers face difficulty in interpreting an algorithm and its ranking outputs, and this difficulty is further compounded when the model and the data on which it is trained is proprietary or otherwise confidential, as is often the case. "Ranking Facts" is a web-based system that generates a "nutrition label" for scoring and ranking algorithms based on factors or "widgets" to communicate an algorithm's methodology or output[22]. Here, the label serves more as an interpretability tool than as a summary of information as the "Nutrition Facts" and "Privacy Nutrition Label" operate. The widgets work together, not modularly, to assess the algorithm on author-created categories of transparency, fairness, stability, and diversity. The demonstration scenarios for using real datasets from college rankings, criminal risk assessment, and financial services establish that the label is potentially applicable to a diverse range of domains. This lends credence to the potential utility in other fields as well, including the rapidly evolving field of AI.

More recently, in an effort to improve transparency, accountability, and outcomes of AI systems, AI researchers have proposed methods for standardizing practices and communicating information about the data itself.

The first draws from computer hardware and industry safety standards where datasheets are an industry-wide standard. In datasets, however, they are a novel concept. Datasheets are functionally comparable to the label concept and, like labels that by and large objectively surface empirical information, can often include other information such as recommended uses which are more subjective. "Datasheets for Datasets" a proposal from researchers at Microsoft Research, Georgia Tech, University of Maryland, and the

AI Now Institute seeks to standardize information about public datasets, commercial APIs, and pretrained models. The proposed datasheet includes dataset provenance, key characteristics, relevant regulations and test results, but also significant yet more subjective information such as potential bias, strengths and weaknesses of the dataset, API, or model, and suggested uses[23]. As domain experts, dataset, API, and model creators would be responsible for creating the datasheets, not end users or other parties.

We are also aware of a forthcoming study from the field of natural language processing (NLP), "Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science"[24]. The researchers seek to address ethics, exclusion, and bias issues in NLP systems. Borrowing from similar practices in other fields of practice, the position paper puts forward the concept and practice of "data statements" which are qualitative summaries that provide detailed information and important context about the populations the datasets represent. The information contained in data statements can be used to surface potential mismatches between the populations used to train a system and the populations in planned use prior to deployment, to help diagnose sources of bias that are discovered in deployed systems, and to help understand how experimental results might generalize. The paper's authors suggest that data statements should eventually become required practice for system documentation and academic publications for NLP systems and should be extended to other data types (e.g. image data) albeit with tailored schema.

We take a different, yet complementary, approach. We hypothesize that the concept of a "nutrition label" for datasets is an effective means to provide a scalable and efficient tool to improve the process of dataset interrogation and analysis prior to and during model development. In supporting our hypothesis, we created a prototype, the Dataset Nutrition Label (the Label). Three goals drive this work. First, to inform and improve data specialists' selection and interrogation of datasets and to prompt critical analysis. Consequently, data specialists are the primary intended audience. Second, to gain traction as a practical, readily deployable tool, we prioritize efficiency and flexibility. To that end, we do not suggest one specific approach to the Label, or charge one specific community with creating the Label. Rather, our prototype is modular, and the underlying framework is one that anyone can utilize. Lastly, we leverage probabilistic computing tools to surface potential corollaries, anomalies, and proxies. This is particularly beneficial because resolving these issues requires excess development time, and can lead to undesired correlations in trained models.

Methods

Some assumptions are made to focus our prototyping efforts. Only tabular data is considered. Additionally, we limit our explorations to datasets <10k rows. This allows for a narrower scope and deeper analysis. The Label's first contribution lies in the standard format it provides for metadata communication. This works to address weaknesses in the most common format for tabular data exchange: comma separated values, or the ".csv" format. Despite its simple plain-text nature, portability, and interoperability[25], the

lack of additional .csv metadata describing how data should be interpreted, validated, and displayed, is perhaps its biggest drawback. As early as 2015, the World Wide Web Consortium published recommendations on "Metadata Vocabulary for Tabular Data" and "Access methods for CSV Metadata"[26,27]. However, the adoption of these recommendations within the data science community is not widespread. The Label also builds on existing data science practices: directly following the acquisition of a dataset, most data specialists often enter an "exploratory phase". This can be seen, for instance, on web-hosted machine learning competition platforms such as Kaggle, and involves understanding dataset distributions through histograms and other basic statistics. The Label attempts to provide these statistics "out of the box," with the hopes of shortening model development lead times. The Label also aims to provide insights from advanced probabilistic modelling backends for more advanced users. While targeted mainly at a professional audience, many of the modules do not require expert training for interpretation and can thus be utilized in a public-facing version of the Label.

Modular Architecture

The Label is designed in an extensible fashion with multiple distinct components that we refer to as "modules" **(Table 1)**. The modules are stand-alone, allowing for greater flexibility as arrangements of different modules can be used for different types of datasets. This format also caters to a wide range of requirements and information available for a specific dataset. During label generation and subsequent updates, it also accommodates data specialists of different backgrounds and technical skill levels.

Modules **(Table 1 & 2)** range from the purely non-technical, such as the Metadata module, to the highly technical, such as the Probabilistic Computing module. Some modules require manual effort to generate, such as those that provide qualitative descriptions of the data (Metadata, Provenance, Variables), while others can ideally be the result of an automated process (Statistics, Pair Plots). Modules also vary in their subjectivity, especially where there exists a reliance on the Label author to identify which questions should be asked of the data and in what way (e.g. Probabilistic Computing). Many of the example modules are also interactive, highlighting a crucial benefit of a label living on a platform (such as a web page) that supports user interaction. This allows Label users to interrogate various dataset aspects with great flexibility and free of preconceived notions developed during Label generation. Lastly, some modules could be designed to act as proxies for their corresponding dataset as they do not expose the underlying data. This could be key when dealing with proprietary datasets, as much of this data will not or cannot be released to the public based on intellectual property or other constraints. Other modules expose information such as distribution metrics which, in theory, would allow adversaries to approximate the dataset contents. The choice of module(s) is thus based on the availability of information, level of willingness and effort volunteered to document the dataset, and privacy concerns.

Table 1. Table illustrating 7 modules of the Dataset Nutrition Label, together with their description, role, and contents.

Module Name	Description	Contents
Metadata	Meta information. This module is the only required module. It represents the absolute minimum information to be presented	Filename, file format, URL, domain, keywords, type, dataset size, % of missing cells, license, release date, collection range, description
Provenance	Information regarding the origin and lineage of the dataset	Source and author contact information with version history
Variables	Descriptions of each variable (column) in the dataset	Textual descriptions
Statistics	Simple statistics for all variables, in addition to stratifications into ordinal, nominal, continuous, and discrete	Least/most frequent entries, min/max, median, mean,.etc
Pair Plots	Distributions and linear correlations between 2 chosen variables	Histograms and heatmaps
Probabilistic Model	Synthetic data generated using distribution hypotheses from which the data was drawn - leverages a probabilistic programming backend	Histograms and other statistical plots
Ground Truth Correlations	Linear correlations between a chosen variable in the dataset and variables from other datasets considered to be "ground truth", such as Census Data	Heatmaps

The list of modules currently examined in this study, while not exhaustive, provides a solid representation of the kinds of flexibility supported by the Label framework. Other modules considered for future iterations or additional datasets include but are not limited to: a comments section for users to interact with authors of the Label for feedback or other purposes; an extension of the Provenance section that includes the versioning history and change logs of the dataset and associated Labels over time, similar to Git; a privacy-focused module that indicates any sensitive information and whether the data was collected with consent; and finally, a usage tracking module that documents data utilization and references using some form of identifier, similar to the Digital Object Identifier[28] and associated citation systems in scientific publishing.

Table 2. Variability of attributes across prototype modules highlights the potential diversity of information included in a Label

	Module Characteristic - Level Required				
Module Name	**Technical Expertise**	**Manual Effort**	**Subjectivity**	**Interactivity**	**Data Exposure**
Metadata	Low	High	Low	Low	Low
Provenance	Low	High	Low	Low	Low
Variables	Low	High	Medium	Low	Medium
Statistics	Medium	Low	Low	Low	Medium
Pair Plots	Medium	Low	Low	High	High
Probabilistic Modeling	High	Medium	High	Low	High
Ground Truth Correlations	Medium	Medium	Low	Low	High

Web-Based Application

The label is envisioned as a digital object that can be both generated and viewed by web-based applications. The label ecosystem comprises two main components: a label maker and a label viewer **(Figure 3)**. Given a specific dataset, the label maker application allows users to select the desired modules and generate them. While the generation of some modules is fully automated, some require human input **(Table 2)**. For instance, the Metadata module mainly requires explicit input, while the Pair Plots module can be generated automatically from the dataset. The Label generator pre-populates as many fields as possible and alerts users to those requiring action. The Label itself lives in a .json format, as one that is human readable and well supported. The Label can then be viewed within the label viewer application where formating is carried out to achieve the desired user interface and user interaction effects. In terms of visual appearance and design, format and typeface requirements of the "Nutrition Facts" label [29] is used. These guidelines, such as the all black font color on white contrasting background, are optimized for clarity and conciseness. Design changes are anticipated in further iterations, and should be informed by user testing.

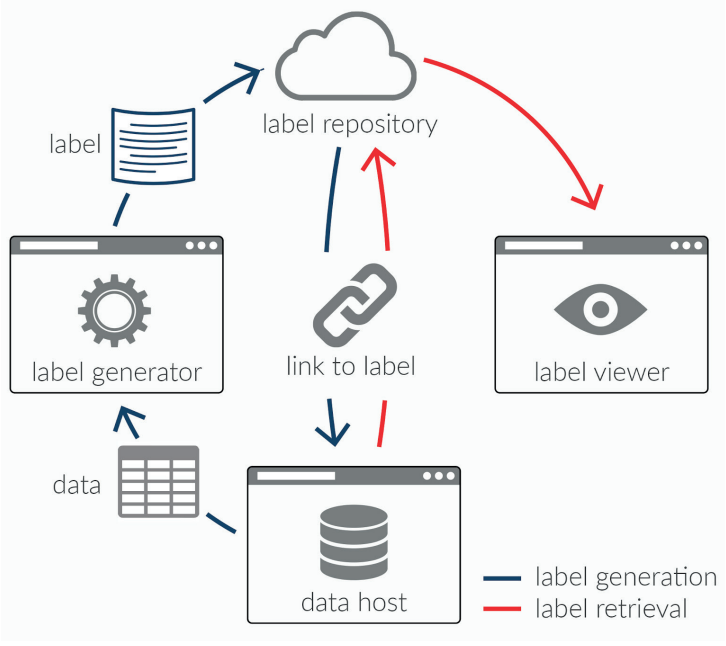

Figure 3. Architecture of the proposed Data Nutrition Label ecosystem.

Backends

Simple statistical analyses involving the generation of histograms, distribution information, and linear correlations are carried out directly in the browser, given tabular datasets of <100K rows. Server-side processing is thus reserved for more specialized and sophisticated analyses requiring additional computational power. Such processing could run multiple backends with the ultimate aim of providing the Label authors with a diverse set of options, fueled by the plethora of tools developed by research groups for automating the generation of summaries, insights, and understandings of datasets. The Label thus becomes a medium for the continuous deployment and testing of these tools. A somewhat recent and particularly powerful example of this is probabilistic computing, and specifically, BayesDB[30], an open source platform developed by researchers at MIT. With minimal modeling and programming effort, BayesDB enables inference of a model that captures the structure underlying the data and generates statistical summaries based on such structure.

Results

To test the concept generally and the modular framework specifically, we built a prototype with a dataset that included information about people and was maintained by an organization invested in better understanding the data. This combination of factors provides necessary information and access to build a wide variety of modules, including those that require full knowledge of the data and the ability to contact the organization that maintains the dataset. We were granted access to the "Dollars for Docs" database from ProPublica, an independent, nonprofit newsroom that produces investigative journalism in the public interest[2]. The dataset, which contains payments to doctors and teaching hospitals from pharmaceutical and medical device companies over a two-year time period (August 2013 - December 2015), was originally released by the U.S. Centers for Medicare and Medicaid Services (CMS) and compiled by ProPublica into a single, comprehensive database.

The resulting prototype successfully demonstrates how disparate modules can be built on a specific dataset in order to highlight multiple, complementary facets of the data, ideally to be leveraged for further investigation by data specialists through the use of additional tools and strategies. The prototype Label includes seven modules **(Table 1, 2)**. The Metadata, Provenance, and Variables modules **(Supp. Figure 1)** provide as-is dataset information. They mirror information submitted by the Label authors as well as provide a standard format for both the generation and consumption of such data. The Statistics module **(Supp. Figure 2)** starts to offer a glimpse into the dataset distributions. For instance, the skewness of a 500 row dataset subset towards a particular drug "Xarelto" can be quickly identified as the most frequent entry under the variable "product_name", and "Aciphex" as the least frequent entry. The Pair Plot module **(Figure 4)** starts to introduce interactivity into the label where the viewer is able to choose the variable pair being compared to one another. A specialist building a model predicting marketing

2 https://projects.propublica.org/docdollars/

spend in each state, for example, may choose to compare "recipient_state" and "total_amount_of_payment_usdollars," and will observe that some states (CA, NY) are more highly correlated with spend. In this case, the specialist would probably normalize the population as the next step beyond consulting the Label in order to identify anomalous spending trends.

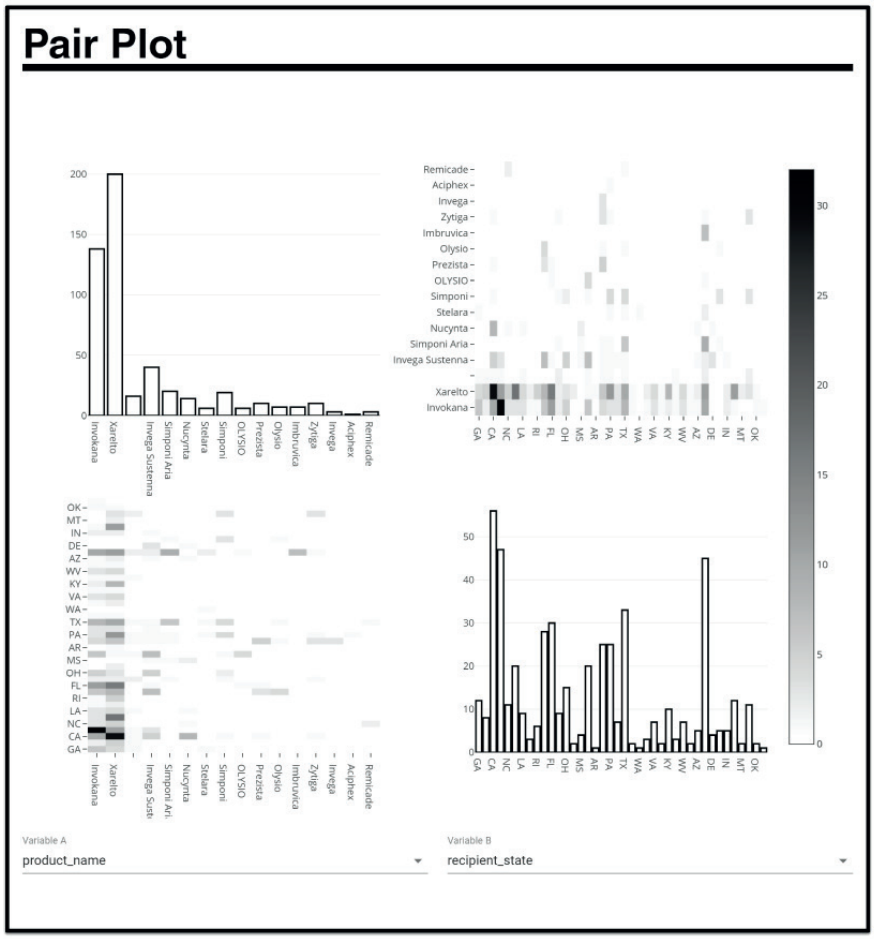

Figure 4. Prototype Label demonstrating the Pair Plot module and highlighting the interactive dropdown menus for selecting variables.

While all modules thus far investigate the dataset itself, the Probabilistic Model module **(Figure 5)** attempts to generate synthetic data by utilizing the aforementioned BayesDB backend. Computed from an inferred generative model, this module allows for the full benefits of Bayesian analysis[31], such as interpretability of inferences, coping with missing data, and robustness to outliers and regions of sparse data. In this specific use

case, an underrepresented drug is chosen from the dataset and the probability of this drug receiving a payment in different states is inferred. With the inevitable variation in the representation of different groups in datasets, such analyses are of great utility in extracting insights - even from relatively small sample sizes. A quick toggle indicates that the top few states for marketing spend are likely the same few states - with a few exceptions, including that NJ is likely to receive much more money for marketing activities relating to the drug Xarelto. Again, this information only acts as a flag for the "what"; specialists will ideally continue to investigate the data in order to identify the "why".

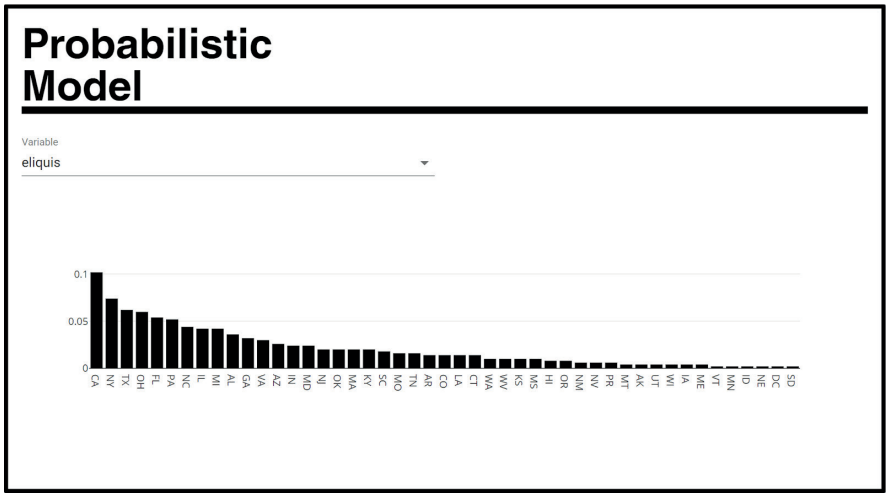

Figure 5. Prototype Label demonstrating the Probabilistic Model module and showcasing a hypothetical distribution for payments made towards the drug "Eliquis" across different states.

It is unavoidable that datasets collected from the real-world have relationships to demographics that the data specialist or other entities do not wish to propagate into the learned model and the inferences produced from it. For example, is a variable or an aggregate of a variable strongly correlated with the Hispanic population in a given region? To surface relationships like this, it is often necessary to explicitly compute a comparison between the dataset and demographic "ground truth" data, which is a task that can be both time consuming and challenging. The Ground Truth Correlation module **(Figure 6)** provides the data specialist initial evidence as to whether such relationships are likely, thus warranting further analysis. In order to surface any anomalies in the demographic distribution of these variables, we selected the 2010 US Census data as "ground truth" for zip code and race. The module then correlates zip code Census data with the dataset and calculates the Pearson correlation between demographics and field aggregates. To demonstrate its utility, the Label **(Figure 6, top)** highlights the negative correlations between the (sum of the) amount of payment field and demographics. A second example **(Figure 6, bottom)**, highlights the positive correlation between a "spend_per_person" aggregate and demographics. This module demonstrates, in a straightforward way,

specific anomalous relationships in the data that the data specialist should pay attention to during model training. In the prototype, we observe a slight positive correlation between white zip codes and payments, and a slight negative correlation between rural zip codes and payments. Toggling to per person spend underscores similar overall trends.

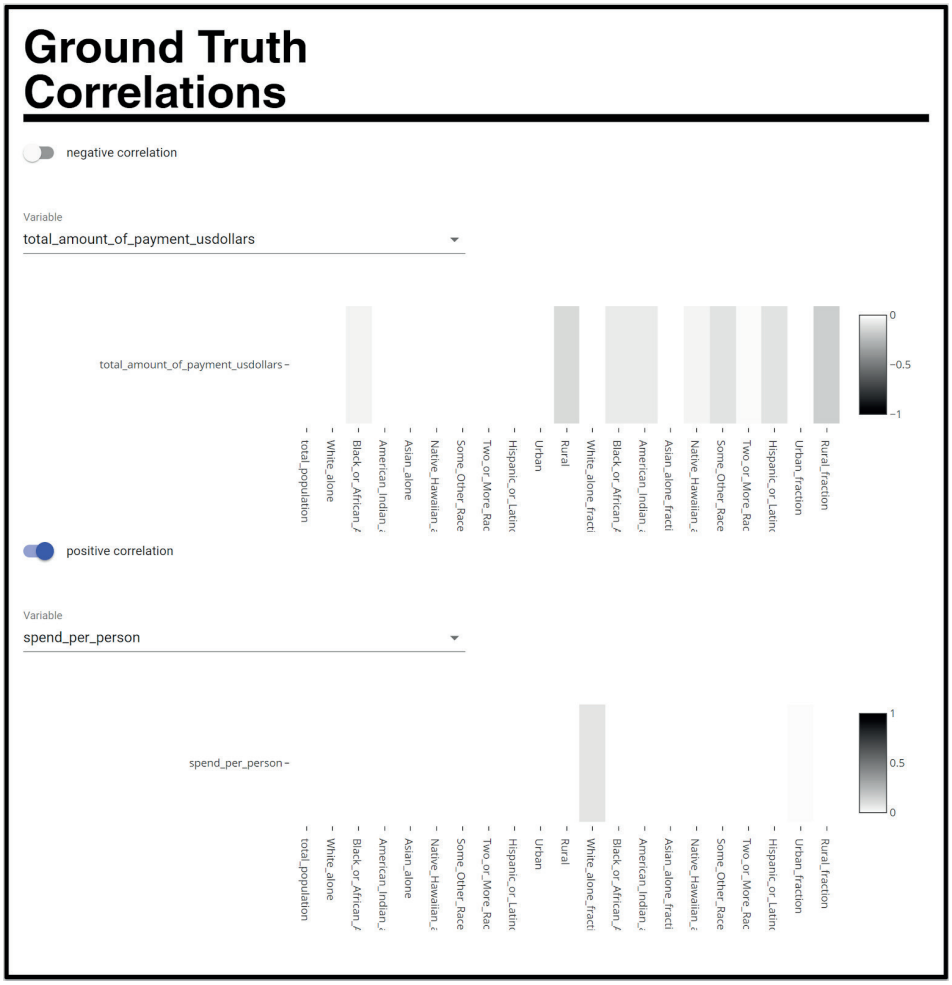

Figure 6. The negative (top) and positive (bottom) correlations to demographics produced by the Ground Truth Correlations module.

Discussion

The Label offers many benefits. Overall, it prompts critical questions and interrogation in the preprocessing phase of model development. It also expedites decision making, which saves time in the overall model development phase without sacrificing the quality or thoroughness of the data interrogation itself, perhaps encouraging better practices at scale. These benefits apply across the spectrum of data specialists' skill and experience, but are particularly useful for those new to the field or less attuned to concerns around bias and algorithmic accountability. First, the Label creates a pre-generated "floor" for basic data interrogation in the data selection phase. It also indicates key dataset attributes in a standardized format. This gives data specialists a distilled yet comprehensive overview of the "ingredients" of the dataset, which allows for a quick and effective comparison of multiple datasets before committing to one for further investigation. It also enables the data specialist to better understand and ascertain the fitness of a dataset by scanning missing values, summary statistics of the data, correlations or proxies, and other important factors. As a result, the data specialist may discard a problematic dataset or work to improve its viability prior to utilizing it.

Improved dataset selection affords a secondary benefit: higher quality models. The Label provides data specialists improved means by which to interrogate the selected dataset during model development, previously a costly and onerous enterprise. The Ground Truth Correlation module, in particular, provides a helpful point of reference for the data specialist before model completion, and surfaces issues such as surprising variable correlations, missing data, anomalous data distributions, or other factors that could reinforce or perpetuate bias in the dataset. Addressing these factors in the model creation and training phase saves costs, time, and effort, and also could prevent bad outcomes early on, rather than addressing them after the fact.

The Label is built with scalability in mind, and with an eye towards standardization. The modular framework provides flexibility for dataset authors and publishers to identify the "right" kind and amount of information to include in a Label; over time, this could become a set of domain-specific best practices. The interactivity of the Label also permits flexibility, as insights about the dataset may arise over time. For example, the ground truth data used for comparison could evolve, rendering a previously unsuitable dataset suitable. Interactive Labels also give data specialists the ability to dive further into anomalous data, rather than simply accepting static information provided by the Label author. With some modules more subjective in nature, and with a range of domain expertise across data specialists, this is particularly important. For advanced data specialists, the flexible Label backend makes it easy to "plug-in" more complex engines. Such complex backends can provide different statistical tools; for example, the Probabilistic Computing module makes it possible to investigate low frequency variables by generating synthetic data. Synthetic data gives data specialists the ability to address incomplete data, and opens a potential path to privacy preserving data usage[32].

Lastly, the Label functioning as a proxy for the dataset itself is an intriguing, even if distant, possibility. Increased calls for accountability of AI systems demand investigation

of datasets used in training models, but disclosing those datasets, even to a limited audience, may pose risks to privacy, security, and intellectual property that calls this approach into question. If the Label is able to convey the information essential for accountability in the dataset without disclosing the data itself, it would provide a valuable and much needed auditing tool[33] for AI systems while still preserving privacy, security, and proprietary information.

Limitations and Mitigations

There are challenges to our approach. The extensive variety of datasets used to build models raises important questions around whether the Label can generalize across data and dataset type, size, composition, and in different domains, and furthermore, whether a data specialist or domain expert will need to be involved in the creation of a Label across these different datasets. This could arise in an instance where important semantic information is atypically labeled and would be challenging to interpret automatically, such as if the field for zip code in a dataset had a custom field for "geographic area." A data specialist or domain expert may also be required when building a Label for sensitive or proprietary data, which may be accessible only to those who built the dataset and not accessible to the public. Building the Label as a modular system somewhat mitigates the complication of requiring input from a domain expert, as the framework can adapt to domain-specific best practices, and can easily support the generation of different types of Labels based on access. Within the Provenance module, it may be necessary or helpful to surface who made the Label, and what relationship they have to the dataset.

The veracity and usefulness of the Ground Truth Comparison module depends on the accuracy of the "Ground Truth" dataset, which serves as a benchmark standard and is considered objective, accurate, and provable, and with clear provenance. However, problematic ground truth data may lead to futile or even harmful comparisons. Without a realistic way to eliminate bias in all datasets, a mitigating step is to build Labels for ground truth datasets themselves. If these Labels include community feedback and comment modules, dataset authors can address the issues directly.

Further investigation is necessary to understand the feasibility and desirability of using the Label as a proxy for proprietary datasets. This would likely require that the dataset creator or controller create the Label. Another challenge is that the Label might not prompt the right questions or reactions for the data specialist, leaving certain biased data undetected. Analyses of machine bias indicate that zip codes often proxy for race, but many other proxies still exist, especially as the models themselves approach levels of complexity that are difficult or impossible for humans to comprehend and new or unexpected proxies emerge. Integrating new methods or tools to help identify proxies will be important to the industry, and our hope is that the Label will be flexible in such a way that these tools can be leveraged to create additional modules as they become available.

Finally, design of the label itself will require additional attention to determine the appropriate amount of information for presentation, comprehension, and adoption. As Kelley et al. made clear in their work on Privacy Nutrition Label[20], design is a key element in the efficacy of the label. It is worth investigating and testing the most effective presentation to drive adoption.

Future Directions

This paper and prototype are the first step toward bringing together a wide range of data specialists, from those who are creating and publishing datasets to those utilizing datasets to build models, in order to improve the quality of datasets used in AI-generated models.

Deeper research and iteration will be necessary as we continue to build additional prototypes of the Label. Creating a "nutrition label" for datasets is nascent and requires additional investigations about what information (in the form of modules or otherwise) is useful and practical to include. Based on the relatively small reach of our survey, we also recommend that a more rigorous survey be conducted to more accurately identify needs, as the survey we administered was limited in its reach, and disproportionately indexed to American and European respondents working in the private sector. The information pertinent to a data specialist will also shift based on the domain of the data, necessitating the building of additional prototypes for different kinds of datasets. The opportunities afforded by complex machine learning tools such as BayesDB in the creation of additional modules deserve more fulsome exploration to maximize the usability and usefulness of the Label.

Through building relationships with dataset publishers and circulating the Label, we hope to identify not only additional datasets for prototypes, but also to launch our Label on open datasets so that we can study the impact of the Label on the use of and conversation around the data. We will consider collaborations with colleagues from industry and academia to further drive this work, building knowledge around the impediments to adoption and considering ways that regulatory frameworks could further support the creation of a best practice or standard.

In terms of the Label ecosystem, the existence of a label for any given dataset could be notated using a mark or symbol, such as the "Conformité Européene" (CE) mark used by the European Union[34], on the author's or dataset host's webpage. Clicking on the mark would then navigate to the label viewer application and fetch the corresponding Label from a central repository where all Labels are hosted. Such a centralized archive of Labels would allow for generating usage statistics, least and most used modules, and eventually help inform future Label iterations. More importantly, a repository of this sort could act as an index of datasets without hosting the datasets themselves. For instance, API calls to such a repository could help locate datasets with queries like "MIT license dataset for facial recognition with >100k samples."

Beyond its utility as a tool, the Label could also drive a change in norms. Through using the Label, data specialists will build a habit around questioning datasets through analysis and interrogation techniques, even if a particular dataset does not include a Label. In time, the Label will facilitate an environment that encourages a broad spectrum of dataset creators, cleaners, publishers, and users to create Labels to publish alongside their datasets. This would lead to better identification of issues with data and bias, or inappropriate data collection practices, which in turn would increase data and dataset quality overall.

Looking beyond the Label itself, there are longer term opportunities for this framework and the data science community. Decisions made around the authorship and ownership model for the Label will be critical to the overall direction of the project; who will create these Labels going forward, and who will maintain them? Will there be a single place where all labels live or from where they are all linked? Additional future directions could include: building a public consortium or governing body to consider standards across the industry; creating curriculum for those collecting and working with datasets; and further exploration of appropriate ground truth data.

Conclusions

In an effort to improve the current state of practice of data analysis, we created the Dataset Nutrition Label, a diagnostic framework that provides a concise yet robust and standardized view of the core components of a dataset. We use the ProPublica Dollars for Docs dataset to create the Label prototype.

The Label serves as a proof of concept for several conceptual questions, beginning with the general feasibility of an extensible and diverse modular framework. It also confirms the possibility of mixing qualitative and quantitative modules that leverage different statistical and probabilistic modelling backend technologies in the same overall user experience. The Label integrates both static and interactive modules, underscoring the importance of using an interactive platform (such as a website) for the distribution of the Label itself. Together, this promises flexibility, scalability, and adaptability.

With the Label, data specialists can efficiently compare, select, and interrogate datasets. Additionally, certain modules afford the ability to check for issues with the dataset before and during model development, surface anomalies and potentially dangerous proxies, and find new insights into the data at hand. As a result, data specialists have a better, more efficient process of data interrogation, which will produce higher quality AI models. The Label is a useful, practical, timely, and necessary intervention in the development of AI models, and a first step in a broader effort toward improving the outcomes of AI systems that play an increasingly central role in our lives.

Acknowledgments

We are grateful to the ProPublica team, including Celeste LeCompte, Ryann Jones, Scott Klein, and Hannah Fresques, for their generosity in providing the Dollars for Docs dataset and for their assistance throughout prototype development. We are also grateful to the BayesDB team in the Probabilistic Computing Group at MIT, including Vikash Mansinghka, Sara Rendtorff-Smith, and Ulrich Schaechtle for their valuable time, assistance, and ongoing advice in regards to integrating the BayesDB backend with the Label. Thanks also go to Patrick Gage Kelley for bringing key work to our attention and for his constructive feedback, as well as the 2018 Assembly Cohort and Advisory Board, in particular Matt Taylor, Jack Clark, Rachel Kalmar, Kathy Pham, James Mickens, Andy Ellis, and Nathan Freitas; the City of Boston Office of New Urban Mechanics; and Eric Breck and Mahima Pushkarna of Google Brain for productive and insightful discussions. This work was made possible by the Assembly program led by Jonathan Zittrain of the Berkman Klein Center for Internet & Society and Joi Ito of the MIT Media Lab.

Supporting Information

https://arxiv.org/pdf/1805.03677.pdf

References

1. Davenport, T. H. & Harris, J. G. Automated decision making comes of age. *MIT Sloan Management Review* **46**, 83 (2005).
2. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. in *Advances in Neural Information Processing Systems* 4349–4357 (2016).
3. Buolamwini, J. & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (eds. Friedler, S. A. & Wilson, C.) vol. 81 77–91 (PMLR, 2018).
4. U.S. Consumer Product Safety Commission. Office of the General Counsel. *Compilation of statutes administered by CPSC*. (U.S. Consumer Product Safety Commission, 1998).
5. McClure, F. D. & United States. Food and Drug Administration. *FDA Nutrition Labeling Manual: A Guide for Developing and Using Databases*. (U.S. Food and Drug Administration, 1993).
6. Union, E. Directive 2009/28/EC of the European Parliament and of the Council of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC. *Official Journal of the European Union* **5**, 2009 (2009).
7. Safety, O., Administration, H. & Others. Hazard communication standard: safety data sheets. *OSHA Brief* (2012).
8. United States. Congress. House. Committee on Energy and Commerce. *Nutrition Labeling and Education Act of 1990: Report (to Accompany H.R. 3562) (including Cost Estimate of the Congressional Budget Office)*. (U.S. Government Printing Office, 1990).
9. Balasubramanian, S. K. & Cole, C. Consumers' search and use of nutrition information: The challenge and promise of the nutrition labeling and education act. *J. Mark.* **66**, 112–127 (2002).
10. Moorman, C. A Quasi Experiment to Assess the Consumer and Informational Determinants of Nutrition Information Processing Activities: The Case of the Nutrition Labeling and Education Act. *Journal of Public Policy & Marketing* **15**, 28–44 (1996).
11. Guthrie, J. F., Fox, J. J., Cleveland, L. E. & Welsh, S. Who uses nutrition labeling, and what effects does label use have on diet quality? *J. Nutr. Educ.* **27**, 163–172 (1995).
12. Silverglade, B. A. The Nutrition Labeling and Education Act: Progress to Date and Challenges for the Future. *Journal of Public Policy & Marketing* **15**, 148–150 (1996).
13. Petruccelli, P. J. Consumer and Marketing Implications of Information Provision: The Case of the Nutrition Labeling and Education Act of 1990. *Journal of Public Policy & Marketing* **15**, 150–153 (1996).
14. Teisl, M. F., Levy, A. S. & Others. Does nutrition labeling lead to healthier eating? *Journal of Food Distribution Research* **28**, 18–27 (1997).
15. Drichoutis, A. C., Lazaridis, P. & Nayga, R. M., Jr. Consumers' use of nutritional labels: a review of research studies and issues. *Academy of marketing science review* **2006**, 1 (2006).

16. Shopping for health. *Food Marketing Institute and Prevention Magazine, Washington, DC and Emmaus, Penn* (1995).

17. Borra, S. Consumer perspectives on food labels--. *Am. J. Clin. Nutr.* **83**, 1235S–1235S (2006).

18. Burton, S., Biswas, A. & Netemeyer, R. Effects of Alternative Nutrition Label Formats and Nutrition Reference Information on Consumer Perceptions, Comprehension, and Product Evaluations. *Journal of Public Policy & Marketing* **13**, 36–47 (1994).

19. Ciocchetti, C. A. The future of privacy policies: A privacy nutrition label filled with fair information practices. *John Marshall J. Comput. Inf. Law* **26**, 1 (2008).

20. Kelley, P. G., Bresee, J., Cranor, L. F. & Reeder, R. W. A nutrition label for privacy. in *Proceedings of the 5th Symposium on Usable Privacy and Security* 4 (ACM, 2009).

21. Kelley, P. G., Cesca, L., Bresee, J. & Cranor, L. F. Standardizing Privacy Notices: An Online Study of the Nutrition Label Approach. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 1573–1582 (ACM, 2010).

22. Yang, K. *et al.* A Nutritional Label for Rankings. (2018).

23. Gebru, T. *et al.* Datasheets for Datasets. *arXiv [cs.DB]* (2018).

24. n/a. Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science.

25. Hunt, A. & Thomas, D. *The pragmatic programmer: from journeyman to master*. (Addison-Wesley Professional, 2000).

26. Metadata Vocabulary for Tabular Data. https://www.w3.org/TR/tabular-metadata/.

27. Model for Tabular Data and Metadata on the Web. https://www.w3.org/TR/tabular-data-model/.

28. Chandrakar, R. Digital object identifier system: an overview. *The Electronic Library* **24**, 445–452 (2006).

29. Food, U. S., Administration, D. & Others. A food labeling guide. *Center for Food Safety & Applied Nutrition* (1999).

30. Mansinghka, V., Tibbetts, R., Baxter, J., Shafto, P. & Eaves, B. BayesDB: A probabilistic programming system for querying the probable implications of data. *arXiv [cs.AI]* (2015).

31. Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*. (Springer Science & Business Media, 2013).

32. Surendra H, M. H. S. A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH* **6**, (2017).

33. Select Committee on Artificial Intelligence. *AI in the UK: ready, willing and able?* https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

34. Hanson, D. *CE Marking, Product Standards and World Trade*. (Edward Elgar Publishing, 2005).

11

Chapter 11

The Importance of Transparency and Reproducibility in Artificial Intelligence Research

B Haibe-Kains, GA Adam, *A Hosny*, F Khodakarami, MAQC Society Board, L Waldron, B Wang, C McIntosh, A Kundaje, CS Greene, MM Hoffman, JT Leek, W Huber, A Brazma, J Pineau, R Tibshirani, T Hastie, JPA Ioannidis, J Quackenbush & HJWL Aerts

Nature 2020

In their study, McKinney et al. showed the high potential of artificial intelligence for breast cancer screening. However, the lack of methods' details and computer code undermines its scientific value. We identify obstacles hindering transparent and reproducible artificial intelligence (AI) research as faced by McKinney et al.. and provide solutions with implications for the broader field.

The work by McKinney et al.[1] demonstrates the potential of AI in medical imaging, while highlighting the challenges of making such work reproducible. The authors assert that their system improves the speed and robustness of breast cancer screening, generalizes to populations beyond those used for training, and outperforms radiologists in specific settings. Upon successful prospective clinical validation and approval by regulatory bodies, this new system holds great potential for streamlining clinical workflows, reducing false positives, and improving patient outcomes. However, the absence of sufficiently documented methods and computer code underlying the study effectively undermines its scientific value. This shortcoming limits the evidence required for others to prospectively validate and clinically implement such technologies. Here, we identify obstacles hindering transparent and reproducible AI research as faced by McKinney et al. and provide potential solutions with implications for the broader field.

Scientific progress depends upon the ability of independent researchers to (1) scrutinize the results of a research study, (2) reproduce the study's main results using its materials, and (3) build upon them in future studies[2]. Publication of insufficiently documented research does not meet the core requirements underlying scientific discovery[3,4]. Merely textual descriptions of deep learning models can hide their high level of complexity. Nuances in the computer code may have dramatic effects on the training and evaluation of results[5], potentially leading to unintended consequences[6]. Therefore, transparency in the form of the actual computer code used to train a model and arrive at its final set of parameters is essential for research reproducibility. The authors state *"The code used for training the models has a large number of dependencies on internal tooling, infrastructure and hardware, and its release is therefore not feasible"*. Computational reproducibility is indispensable for high-quality AI applications[7,8]; more complex methods demand greater transparency[9]. In the absence of code, reproducibility falls back on replicating methods from textual description. Although, the authors claim that *"all experiments and implementation details are described in sufficient detail in the Supplementary Methods section to support replication with non-proprietary libraries"*, key details about their analysis are lacking. Even with extensive description, reproducing complex computational pipelines based purely on text is a subjective and challenging task[10].

Table 1. Essential hyperparameters for reproducing the study for each of the three models (Lesion, Breast, and Case), including those missing from the description in Mckinney et al.

	Lesion	Breast	Case
Learning rate	Missing	0.0001	Missing
Learning rate schedule	Missing	Stated	Missing
Optimizer	Stochastic gradient descent with momentum	Adam	Missing
Momentum	Missing	Not applicable	Not applicable
Batch size	4	Unclear	2
Epochs	Missing	120,000	Missing

In addition to the reproducibility challenges inherent to purely textual descriptions of methods, the authors' description of the model development as well as data processing and training pipelines lacks critical details. The definitions of multiple hyperparameters for the model's architecture (composed of three networks referred to as the Breast, Lesion, and Case models) are missing (Table 1). In their original publication, the authors did not disclose the settings for the augmentation pipeline; the transformations used are stochastic and can significantly affect model performance[11]. Details of the training pipeline were also missing. Without this key information, independent reproduction of the training pipeline is not possible.

There exist numerous frameworks and platforms to make artificial intelligence research more transparent and reproducible (Table 2). For the sharing of code, these include Bitbucket, GitHub, and GitLab among others. The multiple software dependencies of large-scale machine learning applications require appropriate control of the software environment, which can be achieved through package managers including Conda, as well as container and virtualization systems, including Code Ocean, Gigantum, Colaboratory, and Docker. If virtualization of the McKinney et al. internal tooling proved to be difficult, they could have released the computer code and documentation. The authors could have also created small artificial examples or used small public datasets[12] to show how new data must be processed to train the model and generate predictions. Sharing the fitted model (architecture along with learned parameters) should be simple aside from privacy concerns that the model may reveal sensitive information about the set of patients used to train it. Nevertheless, techniques for achieving differential privacy exist to alleviate such concerns. Many platforms allow sharing of deep learning models, including TensorFlow Hub, ModelHub.ai, ModelDepot, and Model Zoo with support for multiple frameworks such as PyTorch and Caffe, as well as the TensorFlow library used by the authors. In addition to improving accessibility and transparency, such resources can significantly accelerate model development, validation, and transition into production and clinical implementation.

Table 2. Frameworks and platforms to share code, software dependencies and deep learning models to make artificial intelligence research more transparent and reproducible.

Resource	URL
Code	
BitBucket	https://bitbucket.org
GitHub	https://github.com
GitLab	https://about.gitlab.com
Software dependencies	
Conda	https://conda.io
Code Ocean	https://codeocean.com
Gigantum	https://gigantum.com
Colaboratory	https://colab.research.google.com
Docker	https://www.docker.com
Deep learning models	
TensorFlow Hub	https://www.tensorflow.org/hub
ModelHub	http://modelhub.ai
ModelDepot	https://modeldepot.io
Model Zoo	https://modelzoo.co
Deep learning frameworks	
TensorFlow	https://www.tensorflow.org/
Caffe	https://caffe.berkeleyvision.org/
PyTorch	https://pytorch.org/

Another crucial aspect of ensuring reproducibility lies in access to the data the models were derived from. In their study, McKinney et al. used two large datasets under license, properly disclosing this limitation in their publication. Sharing of patient health information is highly regulated due to privacy concerns. Despite these challenges, sharing of raw data has become more common in biomedical literature, increasing from under 1% in the early 2000s to 20% today[13]. However, if the data cannot be shared, the model predictions and data labels themselves should be released, allowing further statistical analyses. Above all, concerns about data privacy should not be used as a smokescreen to distract from the requirement to release code.

Although sharing of code and data is widely seen as a crucial part of scientific research, the adoption varies across fields. In fields such as genomics, complex computational pipelines and sensitive datasets have been shared for decades[14]. Guidelines related to genomic data are clear, detailed, and most importantly, enforced. It is generally accepted that all code and data are released alongside a publication. In other fields of medicine and science as a whole, this is much less common, and data and code are rarely made available. For scientific efforts where a clinical application is envisioned and human lives would be at stake, we argue that the bar of transparency should be set even higher. If a dataset cannot be shared with the entire scientific community, because of licensing or other insurmountable issues, at a minimum a mechanism should be set so that some highly-trained, independent investigators can access the data and verify the analyses.

The lack of access to code and data in prominent scientific publications may lead to unwarranted and even potentially harmful clinical trials[15]. These unfortunate lessons have not been lost on journal editors and their readers. Journals have an obligation to hold authors to the standards of reproducibility that benefit not only other researchers, but also the authors themselves. Making one's methods reproducible may surface biases or shortcomings to authors before publication[6]. Preventing external validation of a model will likely reduce its impact, as it also prevents other researchers from using and building upon it in future studies. The failure of McKinney et al. to share key materials and information transforms their work from a scientific publication open to verification and adoption by the scientific community into a promotion of a closed technology.

We have high hopes for the utility of AI methods in medicine. Ensuring that these methods meet their potential, however, requires that these studies be scientifically reproducible. The recent advances in computational virtualization and AI frameworks are greatly facilitating the implementations of complex deep neural networks in a more structured, transparent, and reproducible way. Adoption of these technologies will increase the impact of published deep learning algorithms and accelerate the translation of these methods into clinical settings.

References

1. McKinney, S. M., Sieniek, M., Godbole, V. & Godwin, J. International evaluation of an AI system for breast cancer screening. *Nature* (2020).
2. Nature Research Editorial Policies. Reporting standards and availability of data, materials, code and protocols. *Springer Nature* https://www.nature.com/nature-research/editorial-policies/reporting-standards.
3. Bluemke, D. A. *et al.* Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers—From the Radiology Editorial Board. *Radiology* 192515 (2019) doi:10.1148/radiol.2019192515.
4. Gundersen, O. E., Gil, Y. & Aha, D. W. On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications. *AI Magazine* **39**, 56–68 (2018).
5. Crane, M. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. *Transactions of the Association for Computational Linguistics* **6**, 241–252 (2018).
6. Sculley, D. *et al.* Hidden Technical Debt in Machine Learning Systems. in *Advances in Neural Information Processing Systems 28* (eds. Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R.) 2503–2511 (Curran Associates, Inc., 2015).
7. Stodden, V. *et al.* Enhancing reproducibility for computational methods. *Science* **354**, 1240–1241 (2016).
8. Hutson, M. Artificial intelligence faces reproducibility crisis. *Science* **359**, 725–726 (2018).
9. Bzdok, D. & Ioannidis, J. P. A. Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends Neurosci.* **42**, 251–262 (2019).
10. Gundersen, O. E. & Kjensmo, S. State of the art: Reproducibility in artificial intelligence. in *Thirty-second AAAI conference on artificial intelligence* (2018).
11. Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* **6**, 60 (2019).
12. Lee, R. S. *et al.* A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data* **4**, 170177 (2017).
13. Wallach, J. D., Boyack, K. W. & Ioannidis, J. P. A. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015-2017. *PLoS Biol.* **16**, e2006930 (2018).
14. Amann, R. I. *et al.* Toward unrestricted use of public genomic data. *Science* **363**, 350–352 (2019).
15. Carlson, B. Putting oncology patients at risk. *Biotechnol. Healthc.* **9**, 17–21 (2012).

IV

PART IV

Beyond Cancer Imaging

12

Chapter 12

Artificial Intelligence for Global Health

A Hosny & HJWL Aerts

Science 2019

Artificial intelligence (AI) has demonstrated remarkable progress in the detection, diagnosis, and treatment of diseases. Deep learning, a subset of machine learning based on artificial neural networks, has enabled applications with performance levels approaching those of trained professionals in tasks including the interpretation of medical images and discovery of drug compounds[1]. Not surprisingly, most AI developments in healthcare cater to the needs of high-income countries (HICs) where the majority of research is conducted. Conversely, very little is discussed about what AI can bring to medical practice in low- and middle-income countries (LMICs) where workforce shortages and limited resources constrain the access to and delivery of care. AI could play an important role in addressing global healthcare inequities at the individual patient, health system, and population levels. However, challenges in developing and implementing AI applications must be addressed ahead of widespread adoption and measurable impact.

Health conditions in LMICs and HICs are rapidly converging, as indicated by the recent shift of the global disease burden from infectious diseases to chronic non-communicable diseases (NCDs, including cancer, cardiovascular disease, and diabetes)[2]. Both contexts also face similar challenges, such as physician burnout due to work-related stress[3], inefficiencies in clinical workflows, inaccuracies in diagnostic tests, and increases in hospital-acquired infections. Despite these similarities, more basic needs remain unmet in LMICs. These include healthcare workforce shortages, particularly specialist medical professionals such as surgical oncologists and cardiac care nurses. Patients often face limited access to drugs, diagnostic imaging hardware (e.g., ultrasound, X-ray), and surgical infrastructures (operating theatres, devices, anesthesia). When equipment is available, LMICs often lack the technical expertise needed to operate, maintain, and repair it. As a result, 40% of medical equipment in LMICs is out of service[4]. Conditions are exacerbated in fields that require both specialized workforce and equipment. For example, radiotherapy requires a team of radiation oncologists, medical physicists, dosimetrists, and radiation therapists - together with sophisticated particle accelerator equipment. Consequently, more than 50% of cancer patients requiring radiotherapy in LMICs lack access to this relatively affordable and effective treatment modality, with this number reaching 90% in some low income countries[5].

LMICs have undertaken substantial healthcare spending increases, saving millions of lives by improving access to clean water, vaccinations, and HIV treatments. However, changes in healthcare needs owing to increased mortality from complex NCDs require high-quality, longitudinal, and integrated care[6]. These emerging challenges have been central to the United Nations' Sustainable Development Goals, including the aim to reduce by one third premature mortality from NCDs by 2030. AI has the potential to fuel and sustain efforts towards these ambitious goals.

Healthcare-related AI interventions in LMICs can be broadly divided into three application areas (Figure 1). The first includes AI-powered low-cost tools running on smartphones or portable instruments. These mainly address common diseases and are operated by non-specialist community health workers (CHWs) in off-site locations including local centers and households. CHWs may use AI recommendations to triage patients and identify those requiring close follow-up. Such AI applications include

diagnosing skin cancer from photographic images and analyzing peripheral blood samples to diagnose malaria[7]; more are expected given the emergence of pocket diagnostic hardware, including ultrasound probes and microscopes. With increasing smartphone penetration, patient-facing AI applications may guide lifestyle and nutrition, allow symptom self-assessment, and provide advice during pregnancy or recovery periods. Such applications may allow patients to take control of their own health and thereby reduce the burden on limited health systems.

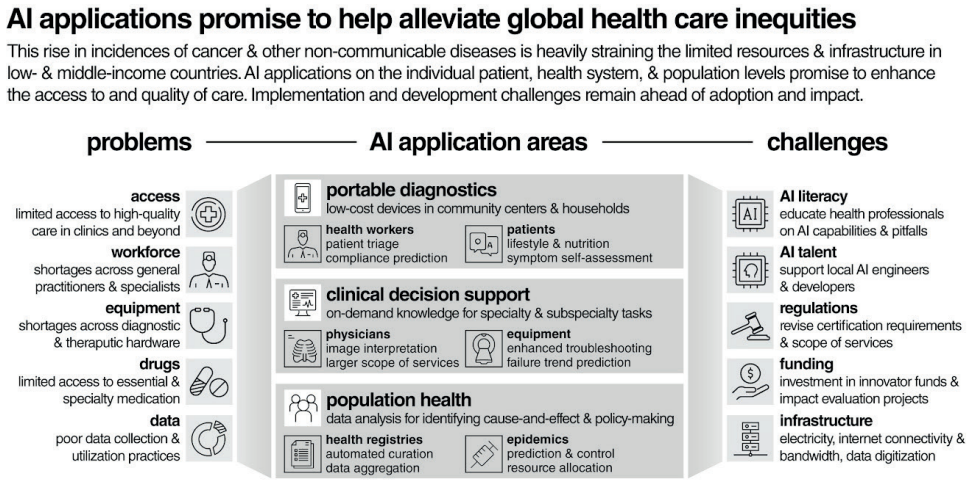

Figure 1. Figure depicting health care problems in low- and middle-income countries, artificial intelligence application areas, and implementation challenges.

The second application area focuses on more specialized medical needs, with the goal of supporting clinical decision-making. AI may allow non-specialized primary care physicians to perform specialized tasks including reading diagnostic radiology and pathology images, only referring to specialists if necessary. As such, the availability of equipment for image capture is a prerequisite for these applications. AI tools may also help provide specialists with expert knowledge across multiple subspecialties. This is particularly important in oncology, for example, oncology, where lack of subspecialists may force an oncologist to manage multiple anatomical tumor sites, and thus deliver care of inferior quality due to the constantly varying scope of services. In radiotherapy, for example, semi-automation of the treatment planning process may speed up treatment delivery, increase patient intake, and allow greater focus on the clinical nuances of patient management - all without requiring additional personnel. Although AI does not directly address diagnostic and therapeutic equipment shortage, AI integration into equipment design may help non-technical operators better troubleshoot and address issues when service technicians are scarce. By analyzing historic maintenance data, AI may also help sustain long-term operations by predicting failures and avoiding lengthy lead times on spare parts and consumables.

The third application area relates to population health and allows public agencies to realize cause-and-effect relationships, appropriately allocate often limited resources, and ultimately mitigate the progression of epidemics[8]. Improving data collection practices in LMICs is central to these applications. For example, AI may help maintain up-to-date national cancer registries. Automated registry curation, by extracting standard data from unstructured free-form text found in radiology and pathology reports, may help reduce labor costs that account for more than 50% of all registry activity expenses[9]. Other applications include identifying hotspots for potential disease outbreaks in unmapped rural areas by utilizing AI-powered analysis of aerial photography and weather patterns, as well as planning and optimizing CHWs' household visiting schedules. Although these applications may prompt immediate actionable interventions, their translation into effective long-term health policies remains unclear.

HIC-based AI applications in healthcare are far from perfect. Most are at the proof-of-concept stage and require further demonstration of utility through clinical validation in prospective trials. The underlying methods are often uninterpretable, making it difficult to predict failures and critically assess results. Data used to train AI models are almost entirely collected within HICs, and models are hence skewed towards certain diseases, demographics, and geographies. With varying degrees of statistical data analysis and quality control, errors and systematic biases are introduced into models thereby limiting their generalizability, especially when deployed in different contexts. Ethical concerns about the use of AI in healthcare include undermining patient data privacy protections, exacerbating the existing tension between providing care and generating profit, as well as introducing a third party into the patient-doctor relationship, which changes expectations of confidentiality and responsibility[10]. From a regulatory perspective, medical malpractice and liabilities in health-related algorithmic decision-making are yet to be formulated. Nearly all AI tools in healthcare are single-task applications, and so they are incapable of fully substituting for health professionals who carry out a wide variety of tasks. Understanding these limitations may help avoid falling prey to hype and inflated expectations.

Introducing AI tools in resource-constrained settings presents additional challenges. The distinct needs, diseases, demographics, and standards of care in LMICs must be acknowledged through identifying specific use cases where AI involvement would have the greatest impact. Data for AI training and validation must be context-specific: Computer vision systems may be required to work with legacy data formats (e.g., film versus digital X-ray), whereas developing chatbots will require compiling corpora in local languages, including medical terminology. Solutions must also be context-specific. For example, an automated system should not recommend treatment options that are unavailable locally or come at prohibitively high costs. Moreover, human factors should be considered: What levels of skill, education, and computer literacy are required of end users? The amount of behavioral change needed to raise awareness and confidence in AI systems should also be addressed, enabling users to recognize limitations and accurately interpret results. Infrastructure constraints should be assessed including the availability of devices for serving AI applications, reliability of internet connectivity and bandwidth,

electrical power availability, amount and quality of existing digital data, as well as future data digitization efforts.

Multiple digital initiatives have been proposed to enhance access to and quality of healthcare in LMICs. These include technologies to support healthcare practices using electronic processes (eHealth) and remote telecommunications (teleHealth), an example of which is mobile health (mHealth) using mobile phones and tablets. Best practices and recommendations for scaling these initiatives in LMICs have been established based on real-life experiences, including the World Health Organization's mHealth Assessment and Planning for Scale (MAPS) Toolkit[11]. These relatively mature efforts could provide learning opportunities for similar digital AI applications. Many of the challenges faced by integrating electronic medical records (EMRs) in LMICs, for example, are likely to also impede AI applications, including limited funding, poor infrastructure for reliably delivering technologies, and discontinuous participation from users[12]. Integration opportunities could also be considered: an existing mHealth application for patient-physician remote communication can be enhanced with an AI chatbot to triage patients prior to the consultation.

There is skepticism about the value of introducing AI in LMICs given the need to prioritize investments in basic infrastructure[13]. AI-driven interventions should not be evaluated in isolation nor should they be regarded as a universal panacea: Although AI development may require sizable initial investments, the marginal cost of providing an existing AI software service to one more user is miniscule, giving it such applications economical scalability. An AI application may also utilize the same deployment channels used by existing digital technologies in LMICs, making it almost readily deployable and reducing infrastructure spending.

Given careful strategic planning of development and implementation efforts fronts, AI solutions could promise to help address major challenges in global health. Ultimately, AI interventions in LMICs should be initiated, owned, and administered by local stakeholders from end users to health agencies - with HICs providing funding, expertise, and advice when needed. AI literacy may be included in existing global health educational programs to raise awareness about its capabilities and pitfalls. Empowering local technical AI talent will also be crucial, and may be accelerated through high-quality free educational resources available online from massive open online courses (MOOC). AI implementation will require rethinking existing regulatory frameworks. For example, the training and scope of practice of CHWs may be expanded to include, for example, screening and diagnosing NCDs[14]. Investment areas critical to bringing AI into LMICs must also be identified, as well as gathering evidence on the impact of AI solutions[15]. Uneven distribution of the access to technologies has created a digital divide between the rich and poor, while contributing to existing global inequities. AI could emerge as a socially responsible technology with inherent equity - benefiting humanity across the globe.

References

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
2. Mayor, S. Non-communicable diseases now cause two thirds of deaths worldwide. *BMJ* **355**, i5456 (2016).
3. Rotenstein, L. S. *et al.* Prevalence of Burnout Among Physicians: A Systematic Review. *JAMA* **320**, 1131–1150 (2018).
4. Perry, L. & Malkin, R. Effectiveness of medical equipment donations to improve health systems: how much medical equipment is broken in the developing world? *Med. Biol. Eng. Comput.* **49**, 719–722 (2011).
5. Zubizarreta, E. H., Fidarova, E., Healy, B. & Rosenblatt, E. Need for radiotherapy in low and middle income countries--the silent crisis continues. *Clin. Oncol.* **27**, 107–114 (2015).
6. Kruk, M. E. *et al.* High-quality health systems in the Sustainable Development Goals era: time for a revolution. *The Lancet Global Health* **6**, e1196–e1252 (2018).
7. Oliveira, A. D. *et al.* The Malaria System MicroApp: A New, Mobile Device-Based Tool for Malaria Diagnosis. *JMIR Res. Protoc.* **6**, e70 (2017).
8. Wahl, B., Cossy-Gantner, A., Germann, S. & Schwalbe, N. R. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health* **3**, e000798 (2018).
9. Tangka, F. K. L. *et al.* Resource requirements for cancer registration in areas with limited resources: Analysis of cost data from four low- and middle-income countries. *Cancer Epidemiology* vol. 45 S50–S58 (2016).
10. Char, D. S., Shah, N. H. & Magnus, D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *New England Journal of Medicine* vol. 378 981–983 (2018).
11. Labrique, A. B. *et al.* Best practices in scaling digital health in low and middle income countries. *Global. Health* **14**, 103 (2018).
12. Jawhari, B., Ludwick, D., Keenan, L., Zakus, D. & Hayward, R. Benefits and challenges of EMR implementations in low resource settings: a state-of-the-art review. *BMC Med. Inform. Decis. Mak.* **16**, 116 (2016).
13. Gyawali, B. Does global oncology need artificial intelligence? *Lancet Oncol.* **19**, 599–600 (2018).
14. Mishra, S. R., Neupane, D., Preen, D., Kallestrup, P. & Perry, H. B. Mitigation of non-communicable diseases in developing countries with community health workers. *Global. Health* **11**, 43 (2015).
15. Lancet, T. & The Lancet. Artificial intelligence in global health: a brave new world. *The Lancet* vol. 393 1478 (2019).

13

Chapter 13

General Discussion and Future Perspectives

The goal of improving cancer care efficacy and efficiency is an ultimate driver of new technologies and innovations into the clinic. The ever increasing amount of healthcare data generated as a result of increasing demand for health services has prioritized the need to optimize and streamline clinical workflows. From the early days of X-ray imaging in the 1890s to more recent advances in CT, MR and PET scanning, medical imaging continues to be a pillar of cancer diagnosis and treatment. Current advances in imaging hardware - in terms of quality, sensitivity and resolution - enable the discrimination of minute differences in tissue densities. AI methods offer the opportunity to transform medical image interpretation from a subjective task to one that is quantitative and reproducible. Moreover, AI may identify imaging features that are difficult to recognize by a trained eye and hence support clinical decision making. Within cancer imaging specifically, the ability to automatically detect, characterize, and monitor tumors in imaging data will have a profound impact on our fight against cancer. This final chapter will provide an overall discussion of results within this thesis, as well as outline challenges that lie ahead of widespread clinical adoption of AI tools.

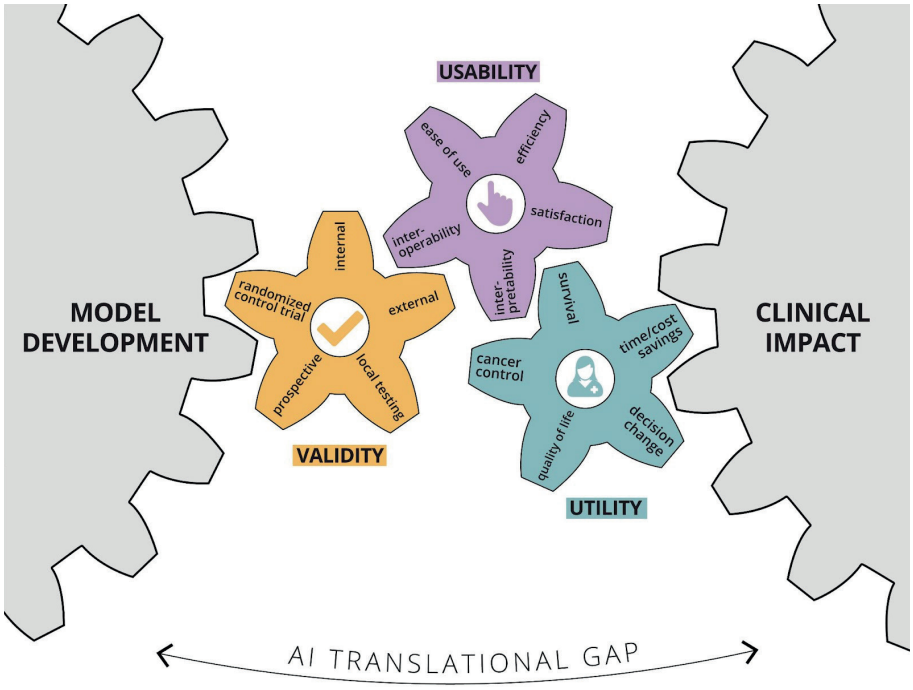

Figure 1. Bridging the AI translational gap between initial model development and routine clinical cancer care by emphasizing and demonstrating three essential concepts: clinical validity, utility, and usability.

PART 1: Artificial Intelligence in Cancer Imaging

Narrow-task AI applications interacting at specific touchpoints along the cancer care path were discussed in **Chapter 2**, together with the increasing cancer datastreams and advances in computational algorithms that have well positioned AI to improve clinical oncology. While there are a number of promising AI applications for clinical oncology in development, substantial challenges remain to bridge the gap to clinical translation. These challenges often relate to the clinical validity, utility, and usability of the AI application **(Figure 1)**. Validity refers to efficacy measures and validation ranging from in silico retrospective experiments to prospective studies and randomized clinical trials. Utility puts the user front and center by examining the ease of use, ergonomics, and interoperability of the proposed solution. Usability outlines the value proposition to all stakeholders, ensuring solutions are improving patient outcomes and quality of life, while also enabling potential time and cost savings for administrators. Incorporation of these concepts into model design and evaluation is easy to overlook, yet is critical to move clinical AI beyond the research and development stage into real-world cancer care.

AI applications at the intersection of radiology and oncology were then explored in **Chapter 3**. In addition to the automation of tumor detection, characterization, and monitoring in imaging data, this chapter also discussed AI interventions in image reconstruction, registration, and report generation **(Figure 2)**.

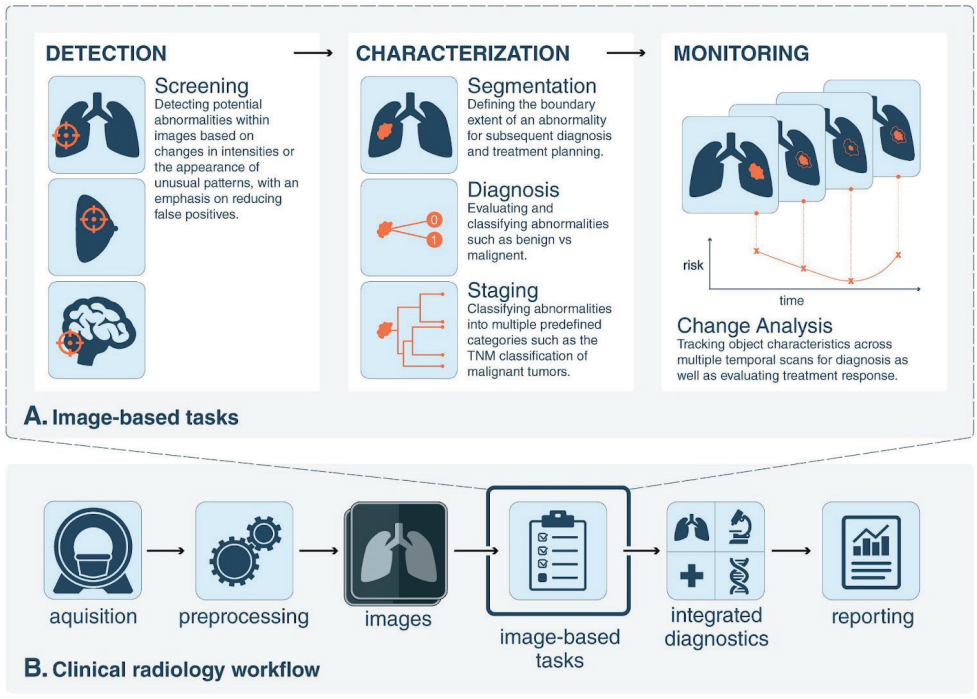

Figure 2. AI impacts areas within oncology imaging. This schematic outlines the various tasks within radiology where AI implementation is likely to have a large impact.

The chapter also highlighted challenges facing clinical implementation. These included the availability of data. Indeed, large amounts of medical data are available, and are stored in such a manner that enables relatively easy access and retrieval. However, such data is rarely curated and this represents a major bottleneck in attempting to develop any AI model. It is imperative that such curation is performed by a trained reader to ensure credibility - making the process expensive and time consuming. Other challenges include the inability for human operators to interpret many deep learning algorithms, often being referred to as 'black-box medicine'. In addition to eroded trust, this makes it difficult to predict failures, isolate the logic for a specific conclusion, or troubleshoot problems. From a regulatory perspective, agencies such as the FDA have been regulating computer-aided detection/diagnosis systems that rely on machine learning and pattern recognition techniques since the earliest days of computing. However, it is the shift to deep learning that now poses new regulatory challenges and requires new guidance for submissions seeking approval. Even after going to market, deep learning methods evolve over time as more data is processed and learned from. Thus, it is crucial to understand the implications of such lifelong learning in these adaptive systems.

Chapter 4 explored the intersection of AI with radiotherapy (RT), an image-based cancer treatment modality. RT plays a critical role in the treatment of cancer, and is indicated in ~50% of cancer patients. RT has become increasingly complex over the past few decades requiring near complete reliance on human-machine interaction including both software and hardware. Beyond gains in accuracy, reproducibility and consistency, partnering human intuition and the capacity of AI to handle large data sets has the potential to drastically improve efficiency and throughput in RT.

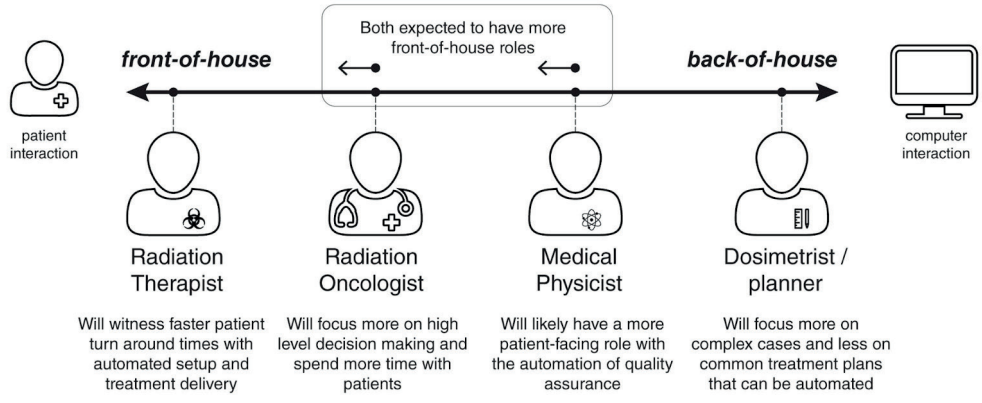

Figure 3. Members of the RT workforce (therapists, radiation oncologists, medical physicists and dosimetrists) are shown along a spectrum of interactions with patients and computers. Our projection of how each profession is expected to evolve with the clinical integration of AI tools is shown and described.

This chapter also highlighted challenges related to AI integration into the RT workflow including data curation as previously noted with radiology. The chapter also touched on the proprietary nature of the RT treatment planning software as another hurdle facing the development of AI solutions. Finally, the chapter noted future directions including the move towards AI tools for predicting patient outcomes, including radiation-specific outcomes (e.g. tumor control, radiation toxicities) as an RT precision medicine approach.

Another important aspect discussed in chapter 4 is the evolution of roles of existing medical RT staff as the shift toward AI integration unfolds over the next few decades **(Figure 3)**. AI will predominantly impact staff members that perform "back-of-house" activities, including the technical aspects of RT such as image segmentation, plan design, and quality assurace, with less of an impact on "front-of-house" activities, that have direct interaction with the patient, typically carried out by physicians, therapists and nurses.

PART 2: Prognostic and Therapeutic Deep Learning Applications

Chapter 5 explored deep learning applications in lung cancer imaging for the automated quantification of radiographic characteristics and potentially improving patient prognostication. We found that deep learning features significantly outperform existing prognostication methods in surgery patients, hinting at their utility in patient stratification and potentially sparing low mortality risk groups from adjuvant chemotherapy. We also demonstrated that areas within and beyond the tumor - especially the tumor-stroma interfaces - had the largest contributions to the prognostic signature, highlighting the importance of tumor-surrounding tissue in patient stratification. Our preliminary genomic associations in this study suggest correlations between the deep learning feature representations and cell cycle and transcriptional processes. Despite their obscure inner workings and lack of a strong theoretical backing, deep learning networks demonstrated a prognostic signal and robustness against specific noise artifacts. This motivates further prospective studies validating their utility in patient stratification and the development of personalized cancer treatment plans.

Next, deep learning was leveraged to develop models for enhancing pathologist accuracy and productivity in **Chapter 6**. Building on data collected through the Boston Lung Cancer Survival cohort, we created deep learning models that can act as non-invasive pathological biomarkers for NSCLC. We trained a convolutional neural network (CNN) to stratify patients into 2 groups based on lung cancer histology. We also found that the CNN-derived CT-radiomics features represented distinct biologic and diagnostic patterns in this cohort, and were associated with underlying tumor microanatomy. This preliminary work has the potential to enhance the human-based decision tree for NSCLC histologic classification, and non-invasive elucidation of tumor biology using radiographic CT data.

In **chapter 7**, we demonstrated that deep learning can perform automated quantification of radiographic characteristics of tumor phenotypes as well as monitor changes in

tumors, before, during, and after treatment in a quantitative manner. More specifically, we illustrated the ability of deep learning networks to predict prognostic endpoints of patients treated with radiation therapy using serial CT imaging routinely obtained during follow-up. We also highlighted their potential in accounting for and utilizing the available serial images to extract the relevant time-point and image features pertinent to the prediction of survival and response to treatment. This provides further insight into applications including the detection of gross residual disease without surgical intervention, as well as other personalized medicine practices.

We then presented a clinical validation framework for therapeutic AI algorithms and demonstrated its application in RT targeting in **chapter 8**. We performed an integrative analysis on eight independent datasets (2208 patients) across four focus areas: benchmarking, primary and secondary validation, as well as human subject experiments. Utilizing a discovery cohort of 787 patients, we developed multiple DL models for localizing and segmenting primary NSCLC tumors and involved lymph nodes in CT images. We then established an interobserver benchmark across six radiation oncologists, followed by an intraobserver benchmark across images segmented by the same radiation oncologist. Primary validation was carried out across 1421 patients including both internal and external cohorts, RT clinical trial data, as well as diagnostic radiology images. Secondary validation was conducted across multiple datasets including test-retest and thorax phantom images. Therein, we assessed the dosimetric and metabolic impact of AI segmentations, as well as measured their stability and accuracy. Finally, in order to gauge the clinical utility of AI segmentations, we carried out a human subject experiment. In a simulated clinical setting, eight radiation oncologists from our institution were asked to perform the segmentation task *de novo* as well as rate and edit a provided AI segmentation.

PART 3: AI Methods and Best Practices

Chapter 9 drew parallels between traditional radiomic methods and their deep learning-based counterparts. We posited that deep learning can emerge as an independent methodology that does not need to rely on handcrafted radiomics to move forward. We outlined how combining traditional radiomic features into deep learning models risks incorporating known human biases into the modelling efforts. We also argued that such a combined approach does not address interpretability issues since even most mathematically-derived handcrafted features capture uninterpretable imaging characteristics that cannot be discerned by the human eye.

Chapter 10 introduced the Dataset Nutrition Label, a diagnostic framework that provides a concise yet robust and standardized view of the core components of a dataset. A growing body of research points to AI systems deployed in a wide range of use cases, where algorithms trained on biased, incomplete, or ill-fitting data produce problematic results. Despite the increased critical attention, data interrogation continues to be a challenging task with many issues being difficult to identify and rectify. Algorithms often come under scrutiny only after they are developed and deployed, which exacerbates

this problem and underscores the need for better data vetting practices earlier in the development pipeline. With the Label, data specialists can efficiently compare, select, and interrogate datasets. We also identified some challenges of the Label, including generalizing across diverse datasets, as well as discussed research and public policy agendas to further advocate its adoption and ultimately improve the AI development ecosystem.

Chapter 11 highlighted the consequences of unpublished code and data in AI publications. In addition to hindering the verification and adoption of scientific findings by the community, lack of reproducibility may lead to unwarranted and potentially harmful clinical trials. The chapter identified various methods for authors to disseminate their works in a transparent manner while abiding by data privacy laws - especially when dealing with sensitive medical data.

PART 4: Beyond Cancer Imaging

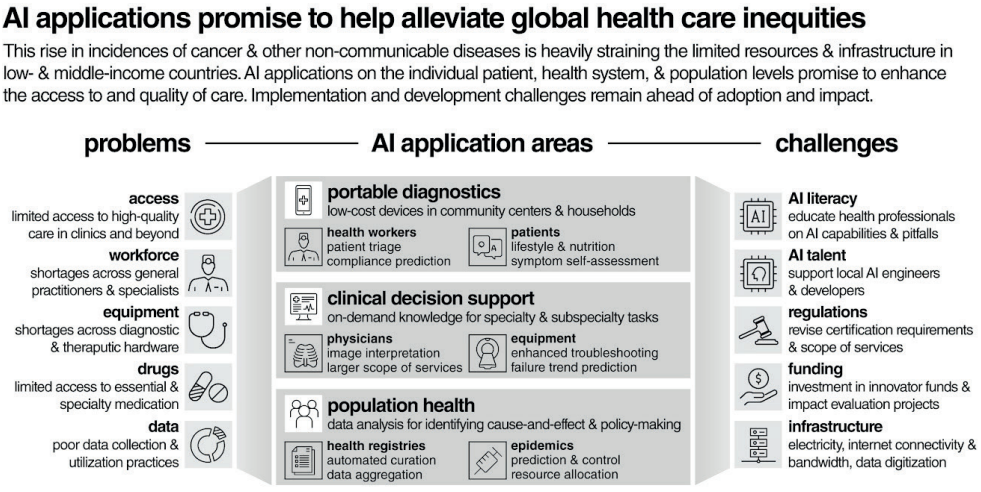

Figure 4. Figure depicting health care problems in low- and middle-income countries, artificial intelligence application areas, and implementation challenges.

Finally, **Chapter 12** explored the framing of AI as a socially responsible technology that promises to address healthcare inequalities. The chapter identified three AI application areas in low- and middle-income countries, namely in portable diagnostics, clinical decision support, and population health. The chapter also outlined challenges in developing and implementing global health AI applications, those that must be addressed ahead of widespread adoption and measurable impact **(Figure 4)**.

Aligning research methodologies is crucial in accurately assessing the impact of AI on patient outcome. In addition to ensuring reproducibility and generalizability, utilizing agreed-upon benchmarking datasets, performance metrics, standard imaging protocols and reporting formats will level the experimentation field and enable unbiased comparison indicators. AI is unlike human intelligence in many ways; excelling in one task does not imply excellence in others. As a result, the promise of up and coming AI methods should not be overstated. Almost all state-of-the-art AI advances fall under the narrow category, where AI is trained for one task and one task only - with only a handful exceeding human intelligence. While such advances may excel in interpreting sensory perceptual information in a bottom-up fashion, they lack higher level, top-down knowledge of contexts as well as fail to make associations in the same way a human brain does. It is therefore evident that the field is still in its infancy and hype should be replaced with rational thinking and mindful planning. AI is also unlikely to replace oncologists within the near or even distant future. The roles of radiologists will expand as they become more connected to technology and have access to better tools. Radiologists are also likely to emerge as critical elements in AI development, contributing knowledge and overseeing efficacy. As AI exceeds human performance, we expect it to evolve into a valuable educational resource. Human operators will oversee outcomes and also seek to interpret the reasoning behind them. This can serve as a means of validation, as well as discovering hidden information that might have been overlooked.

In contrast to prior state-of-the-art AI often locked within proprietary commercial packages, we find that virtually all deep learning software tools available today are open-source. This continues to foster experimentation on a massive scale given the lower barriers to entry and utilization. In terms of data, AI development is expected to shift from processed medical images to raw acquisition data. As raw data is downsampled and optimized for human viewers, loss of information is inevitable and may be avoidable when analysis is run by machines. Some caveats here include reduced interpretability and impeded human validation. As we generate more data, more signal and noise are present. The process of discerning signal from noise is expected to become more challenging over time. Given the difficulties in curating and labelling data, we foresee a push towards unsupervised learning techniques to fully utilize the vast archives of unlabeled medical data.

Open questions include the ambiguity of who controls AI and is ultimately responsible for its decisions, the nature of the interface between AI and healthcare professionals, and whether implementation of an early regulatory policy will cripple AI innovations. Enabling interoperability amongst the vast array of AI applications currently scattered across healthcare will result in a powerful clinical decision-making network. This AI web will not only function at the tool deployment level, but also at the life-long training level. We advocate for creating an interconnected network of deidentified patient data from across the world. Using data on such a scale to train AI models will enable robust and generalizable AI across different patient demographics, geographic regions, diseases, and standards of care. Only then will we see a socially responsible AI benefiting the many and not the few.

Summary

This thesis explored imaging-based applications of deep learning methods in diagnosing and treating non-small cell lung cancer (NSCLC) patients. It comprises a mixture of perspective review articles as well as experimental studies on routine patient imaging data.

PART 1: Artificial Intelligence in Cancer Imaging

Part 1 provided an overarching introduction into the topic through a collection of three perspective articles. **Chapter 2** reviewed a selection of AI applications in oncology from the lens of a patient moving through clinical touch points along the cancer care path. It also mapped the challenges faced in clinical translation across clinical validity, usability, and utility. **Chapter 3** established a general understanding of AI methods particularly pertaining to image-based tasks in oncology. It also explored how up-and-coming AI methods will impact multiple radiograph-based practices within oncology. Finally, it discussed the challenges and hurdles facing the clinical implementation of these methods. **Chapter 4** shifted to discussing AI applications in cancer therapeutics, namely radiotherapy. It provided an overview of the potential for AI to transform radiotherapy by walking through each step of the workflow. It also highlighted examples where AI may increase efficiency, accuracy and quality of radiotherapy, thereby enhancing value-based cancer care delivery in today's resource-limited healthcare environment.

PART 2: Prognostic and Therapeutic Deep Learning Applications

Part 2 explored the development and validation of deep learning applications for the prognostication and treatment of NSCLC patients. **Chapter 5** provided evidence that deep learning networks may be used for mortality risk stratification based on standard-of-care CT images of NSCLC patients. This evidence motivates future research into better deciphering the clinical and biological basis of deep learning networks as well as validation in prospective data. **Chapter 6** investigated the utility of convolutional neural networks (CNN) in non-invasively predicting histology in early-stage NSCLC patients, using routinely acquired noninvasive radiologic images. The association of CNN-derived quantitative radiographic image feature maps with histologic phenotype in this cohort was also assessed. **Chapter 7** demonstrated that deep learning can analyse CT imaging scans at multiple time-points to improve clinical cancer outcome prediction, namely progression, distant metastases and local-regional recurrence. This highlights the impact AI-based non-invasive biomarkers can have in the clinic, given their low cost and minimal requirements for human input. **Chapter 8** validated DL models for localizing and segmenting primary NSCLC tumors and involved lymph nodes in CT images for RT targeting. Beyond establishing inter and intraobserver benchmarks, we performed multi-tiered validation on external datasets including clinical trial and diagnostic radiology data. We also carried out additional dosimetric and metabolic validation,

and measured the models' stability and accuracy. Finally, we conducted human subject experiments to measure clinical utility and physician acceptance.

PART 3: AI Methods and Best Practices

Part 3 highlighted best practices in conducting experimental studies, both on the data science and computational methodology fronts. **Chapter 9** investigated two radiomics methodologies for the prediction of response to cancer therapy: handcrafted feature-based and deep learning-based. It also underscored the importance of model interpretability efforts in understanding the biology of the cancer-normal tissue interface, and ultimately targeting localised cancer therapies such as RT and surgery. **Chapter 10** explored building labels that highlight the key ingredients in a dataset such as meta-data as well as unique or anomalous features regarding distributions, missing data, and comparisons to other 'ground truth' datasets. Such a label may afford data specialists a better and more efficient process of data interrogation, which will produce higher quality AI models. **Chapter 11** underscored the importance of transparency in AI research through the sharing of reproducible computational methods as well as data whenever possible. Such practices may increase the impact of published AI algorithms and accelerate their translation into clinical settings.

PART 4: Beyond Cancer Imaging

Part 4 and **Chapter 12** identified unique challenges in the global health system that may be addressed through AI applications, while assisting in reaching the United Nations' Sustainable Development Goals.

Societal Impact and Valorizations

This thesis explored deep learning applications in lung cancer imaging. Chapters of the thesis involved the study and development of computational image analysis and machine learning methods to extract meaningful information from routine imaging data beyond that currently captured by trained experts. This section will discuss the potential impact of the research on related fields of study as well as society at large.

The global cancer burden is constantly on the rise. This thesis focused on improving the utilization of a routine and standard of care data type pertaining to cancer patients i.e. medical imaging data. Applications discussed in the thesis focused on lung cancer as the leading cause of cancer-related mortalities worldwide - larger than the next five cancer types combined. The computational approaches described here, however, are broadly applicable to other cancer types that come in the form of solid tumors. For instance, another disease candidate for similar studies would be breast cancer. Breast cancer is a major disease affecting women worldwide, and similar to lung cancer, relies on imaging data for patient management. Finally, it is also worth noting that this thesis may have a translational impact beyond oncology. Other diseases that rely on standard of care imaging techniques e.g. neurodegenerative and cardiovascular diseases, may greatly benefit from some of the methods described herein.

The work presented here broadly falls under precision medicine. This emerging approach allows for early diagnosis and customized patient-specific treatments thus delivering the appropriate medical care to the right patient at the right time. Extracting insights from imaging data represents a single facet of such an approach. Similar methodologies to those described in this thesis are actively being applied to other cancer patient data types including genomics, pathology, and electronic health records among others. As such, this work must be considered within the larger context of a growing body of multi-dimensional cancer data, with AI applications deducing patterns and predicting outcomes to improve decision-making. Finally, it is noteworthy that such decision-making applies to both the clinician and the patient alike: the former to advise on the best treatment pathway, and the latter to choose their desired quality of life.

Much hype surrounds AI applications and their utility in healthcare and other domains. Currently, most scientific literature that studies the clinical impact of these technologies tend to be at the proof-of-concept stage i.e. confined to *in silico* validation in small internal data cohorts, and lacking data on real-world clinical utility. Experiments in this thesis aimed at moving studies beyond this preliminary stage by validating models in large external data cohorts, as well as performing clinical validation through human subject experiments wherever possible. As such, this thesis attempts to close the translational gap between early *in silico* validation and larger scale prospective clinical trials. Closing this gap may provide the high levels of confidence needed to pursue AI clinical trials in medicine, uncover model weaknesses that would have been otherwise overlooked, generate preliminary data on human factors given our incomplete understanding of this

area, as well as help quantify the time and effort needed to bring AI outputs to clinically acceptable levels.

To translate the research outputs described herein into clinical tools with direct impact on patients, further validation must be performed. To this end, maintaining the transparency and reproducibility of datasets and methods is crucial. Multiple datasets utilized in this thesis come from open-access online repositories, one of which is The Cancer Imaging Archive[3]. This allows for future improvements on the same data together with developing performance benchmarks. Virtually all computational methods used in this thesis were based on widely available open-source tools, a testament to the great value brought along by open-source software. This ranged from computational languages e.g. R[4] and Python[5], to deep learning libraries e.g. Tensorflow[6], as well as medical imaging software e.g. 3DSlicer[7]. Additionally, multiple trained AI models as a result of work described in chapter 5 have been shared publicly in well-documented and reproducible formats[8].

This research was made possible in large part due to an existing scientific infrastructure, relative abundance of research funding, as well as a network of experts and collaborators - all traits of high-income countries. As a result, most AI developments in healthcare cater to the needs of these countries where the majority of research is conducted. Conversely, little is discussed about what AI can bring to medical practice in low- and middle-income countries, where workforce shortages and limited resources constrain the access to and quality of care. It is therefore crucial to view the work presented here from a global health perspective. Health conditions between these two contexts are rapidly converging, as indicated by the recent shift of the global disease burden from infectious diseases to chronic noncommunicable diseases including cancer. Some of the methods described here may have global oncology applications through allowing non-specialized primary care physicians to perform specialized tasks including interpreting imaging data. Other methods may provide specialists with expert knowledge across multiple subspecialties. This is particularly important in oncology where lack of subspecialists may force an oncologist to manage tumors across multiple anatomical sites, and thus deliver care of inferior quality owing to the constantly varying scope of services.

Uneven distribution of the access to technologies has created a digital divide between the rich and poor, while contributing to existing global inequities. AI could emerge as a socially responsible technology with inherent equity.

3 https://www.cancerimagingarchive.net/
4 https://www.r-project.org/
5 https://www.python.org/
6 https://www.tensorflow.org/
7 https://www.slicer.org/
8 https://github.com/modelhub-ai/deep-prognosis

Some awards and media coverage include:

- Chapter 3 "Artificial Intelligence in Radiology" was featured on the cover of Nature Reviews Cancer journal in August 2018[9], and was discussed in popular media outlets including Wired[10] and The New York Times[11].

- Chapter 5 "Deep Learning for Lung Cancer Prognostication: A Retrospective Multi-Cohort Radiomics Study" was selected by the International Medical Informatics Association (IMIA) as one of the best articles published in 2018 in the 'Cancer Informatics' subfield[12].

- Chapter 7 "Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging" was discussed in an Auntminnie article[13].

- Chapter 11 "The Importance of Transparency and Reproducibility in Artificial Intelligence Research" was discussed in popular media outlets including MIT Technology review[14] and Scientific American[15].

9 https://www.nature.com/nrc/volumes/18/issues/8
10 https://www.wired.co.uk/article/artificial-intelligence-2019-predictions
11 https://www.nytimes.com/2019/05/22/opinion/health-care-privacy-hipaa.html
12 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6697504/
13 https://www.auntminnie.com/index.aspx?sec=log&URL=https%3a%2f%2fwww.auntminnie.com%2findex.aspx%3fsec%3dsup%26sub%3dcto%26pag%3ddis%26ItemID%3d125244%26wf%3d1
14 https://www.technologyreview.com/2020/11/12/1011944/artificial-intelligence-replication-crisis-science-big-tech-google-deepmind-facebook-openai/
15 https://www.scientificamerican.com/article/will-artificial-intelligence-ever-live-up-to-its-hype/

Acknowledgments

I would like to express my sincere gratitude to everyone who supported me during my dissertation, both personally and professionally. I dedicate this thesis to my mother, father, sister - Thank you for your continued love and support during this period. I also dedicate this thesis to my fiancé and partner for her patience, motivation, enthusiasm, and spiritual support - Words can not express how grateful I am.

My deepest gratitude also goes to my thesis advisors Prof. Hugo Aerts and Prof. Raymond Mak for their guidance, academic support, and immense knowledge that formed a treasured part of this thesis - Thank you for encouraging my scientific growth. I would also like to thank my mentor Steve Pieper, PhD for introducing me to the world of medical imaging and always advocating on my behalf.

I dedicate this thesis to Steven Keating[16] (1988-2019). Steven, a cancer patient and collaborator, openly shared his medical data and immediately became my patient zero. His curiosity-driven life has been a real inspiration. I have been very fortunate to be associated with excellent academic institutions throughout my scientific journey, for which I am very grateful. I also appreciate the patients whose data were used in this thesis, as well as the funding agencies that made all this work possible.

I take this opportunity to also thank all my co-authors and collaborators over the past five years - It was an absolute pleasure working with you all. Thank you for all the stimulating discussions and learning opportunities. I also acknowledge my thesis committee members as well as all reviewers for their encouragement, valuable feedback, insightful comments, and hard questions.

Onward and upward.

16 https://news.mit.edu/2019/celebrating-curious-mind-steven-keating-0722

Curriculum Vitae

Ahmed Hosny was born on July 1st, 1987 in Cairo, Egypt. He completed elementary school in Riyadh, middle school in Cairo, and high school in Dubai where he graduated from Cambridge High School in 2004. He went on to study architecture at the American University of Sharjah, where he graduated with a Bachelor of Architecture in 2009. Ahmed started his career as an architect in China. He joined Playze, a boutique architecture firm in Shanghai where he worked on converting shipping containers into habitable spaces. He then joined Foster + Partners in their Beijing office, a world-renowned British architecture firm. There, he worked on computational optimization workflows for facade panel fabrication and structure clash detection simulations.

In 2013, Ahmed moved to Boston where he joined Harvard University's Graduate School of Design working towards a Master degree in Design Technology. For his Master thesis, Ahmed used machine learning to predict the behaviour of sensor-embedded physical objects as a function of human interaction. After graduating in 2015, Ahmed joined the Wyss Institute for Biologically Inspired Engineering as a research fellow. Utilizing his architectural experience in geometry and parametric modeling, he worked with marine biologists to study geometry-function relationships in biological composites including chiton scales & stingray jaws and teeth in order to design their synthetic analogs. At the time, Steven Keating - a research collaborator - was diagnosed with a baseball-sized brain tumor. He openly shared his imaging data and introduced Ahmed to his medical team. This allowed Ahmed to explore a wide array of biomedical problems, and work closely with clinicians to develop medical image processing pipelines, as well as design and fabricate medical devices that have been clinically tested in patients and cadavers.

In 2016, Ahmed started his PhD degree program with Prof. Dr. Hugo Aerts at Maastricht University, The Netherlands, while concurrently conducting machine learning research at Dana-Farber Cancer institute in Boston. Ahmed's research focused on exploring prognostic and therapeutic applications of deep learning in lung cancer imaging. He developed methods to stratify non-small cell lung cancer patients from single and longitudinal radiographic images, including predicting prognostic endpoints such as survival and distant metastasis, as well as response to main and adjuvant therapy. He also explored therapeutic applications including the automated segmentation of target tumors and involved lymph nodes in images of lung cancer patients receiving radiotherapy.

Beyond this journey thus far, Ahmed hopes to continue working with medical domain experts and using technology to positively impact patients' lives.

Scientific Publications

1. A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, H. J. W. L. Aerts, Artificial intelligence in radiology. *Nat. Rev. Cancer* (2018), doi:10.1038/s41568-018-0016-5.

2. E. Huynh, A. Hosny, C. Guthier, D. S. Bitterman, S. F. Petit, D. A. Haas-Kogan, B. Kann, H. J. W. L. Aerts, R. H. Mak, Artificial intelligence in radiation oncology. *Nat. Rev. Clin. Oncol.* **17**, 771–781 (2020).

3. A. Hosny, C. Parmar, T. P. Coroller, P. Grossmann, R. Zeleznik, A. Kumar, J. Bussink, R. J. Gillies, R. H. Mak, H. J. W. L. Aerts, Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **15**, e1002711 (2018).

4. T. L. Chaunzwa, A. Hosny, Y. Xu, A. Shafer, N. Diao, M. Lanuti, D. C. Christiani, R. H. Mak, H. J. W. L. Aerts, Deep learning classification of lung cancer histology using CT images. *Sci. Rep.* **11**, 5471 (2021).

5. Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R. H. Mak, H. J. W. L. Aerts, Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clin. Cancer Res.* **25**, 3266–3275 (2019).

6. A. Hosny, H. J. Aerts, R. H. Mak, Handcrafted versus deep learning radiomics for prediction of cancer therapy response. *The Lancet Digital Health*. **1**, e106–e107 (2019).

7. S. Holland, A. Hosny, S. Newman, J. Joseph, K. Chmielinski, in *Data Protection and Privacy* (Hart Publishing, 2020).

8. B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, Massive Analysis Quality Control (MAQC) Society Board of Directors, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, C. S. Greene, T. Broderick, M. M. Hoffman, J. T. Leek, K. Korthauer, W. Huber, A. Brazma, J. Pineau, R. Tibshirani, T. Hastie, J. P. A. Ioannidis, J. Quackenbush, H. J. W. L. Aerts, Transparency and reproducibility in artificial intelligence. *Nature*. **586**, E14–E16 (2020).

9. A. Hosny, Hugo J W, Artificial intelligence for global health. *Science*. **366**, 955–956 (2019).

10. B. H. Kann, A. Hosny, H. J. W. L. Aerts, Artificial intelligence for clinical oncology. *Cancer Cell* (2021), doi:10.1016/j.ccell.2021.04.002.

11. A. Hosny, D. S. Bitterman, C. V. Guthier, H. Roberts, S. Perni, A. Saraf, J. M. Qian, L. C. Peng, I. M. Pashtan, B. H. Kann, D. Kozono, P. Catalano, H. J. W. L. Aerts, R. H. Mak, Clinical Validation of Deep Learning Algorithms for Lung Cancer Radiotherapy Targeting. [Submitted]