Artificial intelligence

ARTIFICIAL INTELLIGENCE: THE KEY TO STANDARDIZING RESPIRATORY DISEASE EVALUATION 2024

ARTIFICIAL INTELLIGENCE: THE KEY TO STANDARDIZING RESPIRATORY DISEASE EVALUATION

KEVIN B.W. GROOT LIPMAN

Doctoral thesis

ARTIFICIAL INTELLIGENCE: THE KEY TO STANDARDIZING RESPIRATORY DISEASE EVALUATION

Kevin B.W. Groot Lipman

2023

ARTIFICIAL INTELLIGENCE: THE KEY TO STANDARDIZING RESPIRATORY DISEASE EVALUATION

Dissertation

To obtain the degree of Doctor 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 January 22nd, 2024, at 16.00 hours

by

Kevin Bernardus Wilhelmus Groot Lipman

Promotor
Prof. Dr. Regina G.H. Beets-Tan

Copromotor
Dr. Jacobus A. Burgers
Dr. Stefano Trebeschi

Assessment Committee
Prof. dr. Frits M.E. Franssen
Prof. dr. André L.A.J. Dekker
Prof. dr. Christiane K. Kuhl
dr. Jonas Teuwen

Cover	K.B.W. Groot Lipman & K.M.L. van Duren with GPT-4/DALL · E 3
Production	K.B.W. Groot Lipman
ISBN	978-94-6469-749-0

Contents

1

General Introduction

Background & Challenges

Respiratory diseases present in a wide spectrum of manifestations, primarily imaged through Computed Tomography (CT) scans [1]. Generally, these diseases can be categorized as focal or diffuse [2]. Focal diseases are localized to a specific area, e.g., lung nodules or pulmonary embolism [2], while diffuse diseases, such as interstitial lung disease (ILD) and most pleural diseases, are more widespread across the lungs and pleura [3]. Diffuse diseases share common challenges in radiological evaluation: they all cover an extended number of CT slices in thoracic CT scans, making their burden notoriously hard to quantify [4, 5, 6, 7]. Standardizing radiological methods to evaluate the extent of the patient's disease is therefore challenging, often in contrast with the clinical practice, which requires evaluation criteria that are easy to use, time-efficient, and cost-effective [8]. These restrictions have led to using approximations as evaluation criteria, most prominently diameters, for example in tumor diameters to estimate the total tumor volume, and visual inspections, for example to estimate the percentage affected lung parenchyma. The subjectivity and estimation error of these methods can lead to variability in the measurement, and therefore diagnosis, which can impact treatment decisions and patient outcomes [8].

This dissertation analyzes two examples of diffuse pulmonary diseases: COVID-19 and asbestosis. COVID-19 has greatly impacted the need for radiological evaluation, with 700 million cases and nearly 7 million deaths as of November 2023 [9]. The pandemic has highlighted the role of radiology in disease management [10]. However, for COVID-19 lesion quantification on CT, there is a lack of consensus on which imaging features are most predictive of disease severity, making it difficult to quantify the severity of the disease accurately [11, 12]. Moreover, cognitive biases make radiologists vulnerable to overestimating the extent of diseases [11], underscoring the necessity for quantitative volumetric assessment. Asbestosis, an ILD caused by long-term occupational exposure to asbestos, has seen an increasing incidence rate since 1990, leading to 9400 diagnoses worldwide in 2017 [13]. Despite efforts to regulate

asbestos use, cases are particularly reported in high-income regions [13]. The visual assessment of asbestosis in CT scans is subject to considerable inter-observer variability, which makes it difficult to assess the true extent of the fibrotic tissue [14]. This visual inspection currently determines the eligibility for state aid: patients with asbestos-induced fibrosis qualify for financial compensation when the fibrosis is assessed to cover more than 5% of the lung parenchyma. Considering the difficulties of determining the percentage of affected lung parenchyma visually [11], there is a growing need for improved assessment methods for this disease.

Similar to asbestosis, the primary cause of diffuse pleural diseases such as pleural plaques and pleural mesothelioma is the inhalation of asbestos fibers [15, 16, 17]. While asbestosis diagnoses quality for state aid, patients with pleural plaques currently do not, partly due to inconclusive evidence about their effect on lung function, exacerbated by difficulties in quantifying pleural plaque volume [17, 18]. Investigating the relationship between pulmonary function parameters and the extent of pleural plaques could clarify the impact of plaques on lung function, potentially justifying state aid for affected patients [14, 19]. However, the visual quantification of pleural plaque extent and volume is challenging [20], highlighting the need for accurate quantification methods to support research on the clinical significance of pleural plaques.

Pleural mesothelioma is an aggressive tumor that is difficult to assess due to its irregular shape and growth patterns. It has an age-standardized rate of 0.30 per 100,000 individuals with the highest incidence in Northern Europe [21]. Despite being a rare disease, its impact is underlined by the poor median survival ranging from 8.7 months [22] to 10.3 months [23], barely increasing over the past 50 years [21]. The lack of curative treatments and the challenges in evaluating new therapies, partly due to the tumor's unique crescent shape, contribute to difficulties in disease management [24]. The current gold standard of evaluating therapeutic response, i.e. assessing the reaction of the tumor to a given treatment, is Response Evaluation Criteria In Solid Tumors (RECIST). Primarily used in clinical trials, it involves diameter mea-

surements at various CT scan slices and subsequent comparisons with follow-up images [25]. Treatment response is classified based on diameter changes, with over 20% increase indicating progressive disease (PD) and over 30% decrease classifying partial response (PR) [25]. Complete response (CR) is rare in mesothelioma cases [26].

The mesothelioma-specific modified RECIST (mRECIST) differs by measuring the diameter perpendicular to the thoracic cavity contour rather than the longest diameter [27]. Typically, the best-observed response (PD, SD, or PR) during a trial is reported. However, significant interobserver variability remains, as demonstrated in a recent phase II trial where experts agreed on the primary endpoint in only 53% of the patients for the best-observed response [28], which could have happened due to differences in defining the tumor's location and measurement angle [24]. These challenges underscore the need for more precise methods to quantify the entire disease burden at one single time (CT scan) and assess treatment effects over time (difference over multiple CT scans). Such advancements are essential for enhancing the evaluation of both new and existing treatments for pleural mesothelioma.

Proposed Solution

We hypothesize that Artificial Intelligence (AI) has the potential to aid in increasing standardization by automating the quantification process and reducing inter-observer variability. Automatic solutions leveraging AI have the potential to accurately quantify pleural and lung anomalies such as the volumetry of pleural plaques and mesothelioma, and the detection of asbestosis. These automated solutions could offer a standardized approach, reducing inter- and intra-observer variability in the diagnostic process. The development of these AI-based solutions involves the use of complex computational models that are trained on extensive datasets. Convolutional neural networks (CNN) have demonstrated remarkable success in the biomedical imaging field due to their ability to process imaging data with varying degrees of abstraction,

and to learn imaging features automatically [29]. These properties enable navigation and exploration of massive datasets to discover complex structures and patterns that can be employed for prediction, classification, and segmentation (labeling of each voxel). Presently, the state-of-the-art technologies in this domain are based on these CNNs [30]. The models typically involve subsequent filtering operations that down-sample the input image size, effectively reducing the image to a lower dimensional space that contains highly informative quantitative features for classification or regression tasks [29]. For image segmentation, a decoder is typically added to the model, which applies the inverse operation to map the low dimensional space back to the full-resolution image, thereby producing a label map. The utilization of AI-based solutions could enable a more streamlined and standardized diagnostic process while removing the approximate measure. However, AI has its own challenges, such as generalizability, convergence, explainability, and data quality dependence [31]. To ensure the reliability of AI-based approaches, it is important to rigorously validate their performance on external datasets and prospectively evaluate them in clinical practice [32]. Such validation can help establish the robustness and generalizability of AI-based methods and demonstrate their added value in improving the accuracy and consistency of disease quantification [31, 32].

Research Aim and Outline of Thesis

Aim

The primary objective of this thesis is to enhance the standardization of diffuse respiratory disease evaluations. We focus on developing and validating innovative methodologies for the automated and quantitative analysis of abnormalities in chest CT scans, focusing on case studies for asbestos-related diseases and COVID-19. By leveraging the potential of machine learning and advanced image processing techniques, we

intend to improve the accuracy and reproducibility of disease quantification in these scans. Ultimately, we envision these advancements contributing to improved patient outcomes and better-informed clinical decision-making.

Thesis Outline

Part I: Enhancing Disease Quantification at Baseline

Study Question: How can we refine the process of disease quantification at baseline for various respiratory conditions using AI models?

Part II: Evaluating Therapeutic Response in Pleural Mesothelioma

Study Question: What AI-based imaging techniques can be developed to accurately assess and classify the therapeutic response in patients with Pleural Mesothelioma?

Part I: Enhancing Disease Quantification at Baseline

Chapter 2 focuses on AI model development for classification of asbestosis and the eligibility of patients for government support based on clinical examinations.

Chapter 3 entails the prospective validation of the AI system developed in *Chapter 2* for the evaluation of eligibility for state-aid.

Chapter 4 explores the relationship between pleural plaque volume and pulmonary function tests, and the development of automatic AI-driven segmentation of pleural plaques to automatize the task of segmentation.

Chapter 5 evaluates an externally developed AI model for the segmentation of COVID-19 affected tissue in CT scans, and the corresponding CO-RADS score.

Part II: Evaluating Therapeutic Response in Pleural Mesothelioma

Chapter 6 quantifies response to treatment in Pleural Mesothelioma using an AI algorithm for automatic volume quantification in CT scans. It proposes novel volumetric cutoffs for response evaluation and performs external validations.

Part I

Enhancing Disease Quantification at Baseline

European Radiology 2022, doi: 10.1007/s00330-022-09304-2

2

Artificial Intelligence-based diagnosis of asbestosis: analysis of a database with applicants for asbestosis state-aid

Kevin B. W. Groot Lipman, Cornedine J. de Gooijer, Thierry N. Boellaard, Ferdi van der Heijden, Regina G. H. Beets-Tan, Zuhir Bodalal, Stefano Trebeschi*, Jacobus A. Burgers*

*Shared Last Author

Abstract

Objectives

In many countries, workers who developed asbestosis due to their occupation are eligible for government support. Based on the results of clinical examination, a team of pulmonologists determine the eligibility of patients to these programs. In this Dutch cohort study, we aim to demonstrate the potential role of an artificial intelligence (AI)-based-system for automated, standardized, and cost-effective evaluation of applications for asbestosis patients.

Methods

A dataset of n=523 suspected asbestosis cases/applications from across the Netherlands was retrospectively collected. Each case/application was reviewed, and based on the criteria, a panel of three pulmonologists would determine eligibility for government support. An AI-system is proposed, which uses thoracic CT images as input, and predicts the assessment of the clinical panel. Alongside imaging, we evaluated the added value of lung function parameters.

Results

The proposed AI-algorithm reached an AUC of 0.87 ($p<0.001$) in the prediction of accepted versus rejected applications. Diffusion capacity (DLCO) also showed comparable predictive value (AUC=0.85, $p<0.001$); with little correlation between the two parameters (r-squared=0.22, $p<0.001$). The combination of the imaging AI-score and DLCO achieved superior performance (AUC=0.95, $p<0.001$). Interobserver variability between pulmonologists on the panel was estimated at α=0.65 (Krippendorff's alpha).

Conclusion

We developed an AI-system to support the clinical decision-making process for the application to the government support for asbestosis. A multicenter prospective validation study is currently ongoing to examine the added value and reliability of this system alongside the clinic panel.

Keywords

Asbestos, Asbestosis, Tomography, X-Ray Computed, Respiratory Function Tests, Artificial Intelligence

Key Points

Artificial Intelligence can detect imaging patterns of asbestosis in CT scans in a cohort of patients applying for state-aid

Combining the AI prediction with the diffusing lung function parameter reaches the highest diagnostic performance

Specific cases with fibrosis but no asbestosis were correctly classified, suggesting robustness of the AI system, which is currently under prospective validation.

Introduction

Asbestosis is diffuse pulmonary fibrosis emerging after prolonged, mainly occupational, exposure to asbestos [33]. Many countries have banned asbestos in construction and manufacturing [34]. However, due to the long incubation time, many (former) exposed workers now present with asbestosis[35].

Asbestosis patients with occupational asbestos exposure might be eligible for financial compensation [14]. The criteria for obtaining it varies by country, although international attempts have been undertaken for standardization [14, 36, 37]. Standardization is hard to achieve, with disagreement among experts [38, 39] on the Helsinki criteria [37] for asbestosis hindered this process. In the Netherlands, the following criteria are legally set for financial reimbursement: (1) Computed Tomography (CT) imaging, preferably High-Resolution CT (HRCT) with fibrosis covering >5% of the lung area, (2) lung function loss should be present, and (3) occupational asbestos exposure of at least five fiber years (product of the intensity of asbestos exposure times the occupational years [40, 41]). Three independent and experienced pulmonologists review the clinical case and state whether the most likely diagnosis is asbestosis. The majority of votes set the diagnosis for reimbursement. Similar procedures are followed in other countries [42, 19, 43].

This law-driven diagnosis does not coincide entirely with the clinical, multidisciplinary board meeting-driven diagnosis. Additionally, a shared limitation is the unknown inter-rater variability, leaving the quality and reproducibility of the final verdict unknown. This could lead to the same patient receiving different diagnoses for unclear cases. Alternatively, obvious cases are still processed by three experts, where their effort could have had more impact analyzing the unclear cases.

We hypothesize that a system based on artificial intelligence (AI) could replicate the assessments of the three experts. AI is a method to automatically extract data patterns from raw data (e.g. CT scans) to predict outcomes of interest (e.g. decision of the pulmonologists' panel). In this

study, we aim to develop and test an AI-system to assess applications of subjects with recorded exposure to asbestos, and determine whether they are eligible for financial support. If the AI is certain about its prediction, one pulmonologist could be sufficient to verify the AI assessment. More pulmonologists can be assigned to process the unclear case when the AI is uncertain. The resulting AI-algorithm evaluates eligibility and can be implemented uniformly in multiple centers, allowing for increased consistency in handling financial support requests.

Material and Methods

Datasets

We performed a retrospective analysis on a dataset of prospectively included applicants for financial support, collected by the Dutch Institute of Asbestos Victims (IAS) and the Netherlands Cancer Institute (NKI; Amsterdam/NL) between 05/2014 and 11/2019 [14]. Applicants gave informed consent for the use of their data. CT scans with 3mm slice thickness/increment were preferred over 1mm due to hardware constraints. While $\leq$1mm slices are preferred in clinics for diagnosing ILDs, the GPU cannot process hundreds of $\leq$1mm slices as one volume.

Exclusion criteria for the current analysis were: no chest-CT scan available, lungs not fully present in the scan, slice thickness >5mm, or absence of panel verdict. The dataset was divided between training, validation, and test sets based on a random, reproducible split. Each set consisted of an equal ratio of positive/negative cases. For evaluation, four contradictory cases (interstitial lung disease (ILD) but no asbestosis, or asbestosis but little to no ILD) were held out.

When available, the following lung function tests were retrieved: vitalcapacity (VC, cm^3), forced vital capacity (FVC, cm^3), and diffusing capacity of the lung for carbon monoxide (DLCO, mL/min/mmHg). To

compensate for differences in body type, the lung function parameters are denoted in percentage (%) of expected value. To quantify the loss of lung function, Hagmolen Of Ten Have et al. adapted the American Medical Association (AMA) classes described by Rondinelli et al. [44]. The worst-recorded parameter between FVC and DLCO was converted to an impairment class (AMA class, Table 2.1). AMA $\geq$2 is regarded as sufficient for financial support (see Supplement).

Class	**0**	**1**	**2**	**3**	**4**
FVC	$\geq$80%	70-79%	60-69%	50-59%	<50%
DLCO	$\geq$75%	65-74%	55-64%	45-54%	<45%

Table 2.1: Table for converting loss of lung function to a specific AMA class (Guides to the Evaluation of Permanent Impairment Sixth Edition). FVC and DLCO values are the corrected percentages for age, length, and sex of the predicted normal value. The parameter with the highest loss determined the AMA class, which in turn was correlated to the extent of the financial reimbursement.

Design of the Artificial Intelligence

We designed an AI-system for the assessment of eligibility of asbestosis financial support applications following subsequent steps: 1. identification of the lungs and surrounding tissue in chest-CT scans through a localizer, 2. detection of anomalies within the lungs with a detector, and 3. automatic assessment of eligibility through a classifier, based on the CT scan and the anomalies found in 2. These modules function synchronously within the overall AI-diagnostic-system (Figure 2.1). The code is publicly available on the AI-repository of our department[1], enabling other researchers to redo a similar study.

[1] `https://github.com/nki-radiology/asbestosis`

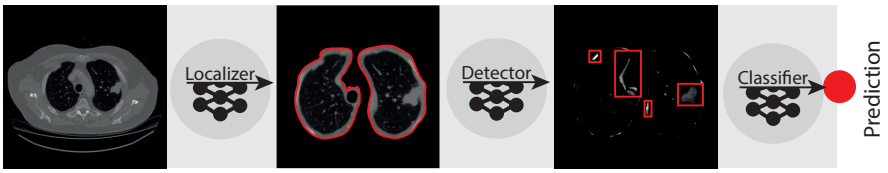

Figure 2.1: An overview of the AI-system with the Localizer, Detector, and Classifier modules. The red outline indicates the areas of interest for each module.

Localizer

This module aims to detect and segment the lungs. We reused an AI-model[2] of LaLonde et al. [45]. Once it identified the lungs, we automatically removed all non-lung pixels from the image. This facilitates the subsequent analysis, ensuring that they will only be performed on lung tissue.

Detector

The goal of this module is to highlight anomalies in the lungs. We based the design on a set of algorithms for anomaly detection, called variational autoencoders (VAE) [46]. In our case, we trained a VAE on a dataset of chest-CTs (see Supplement). By training this network on healthy CT slices, the VAE learns to synthesize healthy lung structures. When a CT with lung anomalies is presented to the network, the VAE will reconstruct those abnormal regions of the lungs poorly, since they are not learned during training. This phenomenon allows us to highlight the anomalies, effectively creating an anomaly heatmap (details in Supplement).

[2] `https://github.com/lalonderodney/SegCaps`

Classifier

This module aims to identify patients who received a positive assessment for asbestosis financial support. We based our design on the ResNet architecture [47], which is commonly employed for image classification tasks. We trained the network using the CT+anomaly heatmap as input and the asbestosis panel verdict as training objective, where the cross-entropy loss function quantified the difference between AI prediction and panel verdict. Once trained, the network made predictions between 0 and 1, to be interpreted as a probability, with 0 being no evidence to support positive assessment and 1 being the opposite.

Data Curation and Labels

To minimize differences between imaging protocols and artifacts from foreign metal bodies, e.g. pacemakers, all Hounsfield Units (HU) were clipped between -1024 HU (air) and 3072 HU (dense bone) [48] and scaled on 0—1 interval. To include adjacent tissue of the thoracic wall (such as pleural plaques/thickening), we dilated the segmentations through morphological operators with a kernel of 13x13x5 voxels. Subsequently, they were visually inspected and adjusted in 3DSlicer (v4.10) [49] if the pleura was not present in the dilation.

The segmented lungs in the CT were cropped to 192x192x96 (sagittal, coronal, axial) and rescaled where needed due to hardware constraints. The ground-truth, i.e. the *label*, was implemented in two configurations: hard and soft. The hard labels were binary (i.e. asbestosis or not), whereas the soft labels reflected the panel's agreement (i.e. ratio of positive assessments). These soft labels were implemented to investigate whether the AI could replicate that level of agreement. When pulmonologists disagree, soft labels penalize uncertain AI predictions less than hard labels. In specific, an AI prediction during training of 0.5 (uncertain) is closer to the fraction of pulmonologists positive: 0.33

(1/3 pulmonologist, 0.17 off target, soft label) than the final verdict of the panel: 0 (1/3 pulmonologist, 0.5 off target, hard label).

Statistical Analysis

To evaluate the panel's inter-observer variability, we calculated Krippendorff's alpha, where $\alpha = 1$ reflects perfect agreement and $\alpha = 0$ disagreement [50]. The performance of the models was evaluated using the ROC-AUC and standard measures of accuracy, sensitivity, specificity, and positive and negative predictive value. We performed McNemar's test to test for significant differences in performance between different methods. The correlation between lung functions and AI predictions was estimated via r-squared (r^2). To visualize the areas where the model focused on in the CT scan, we traced the activations back to the input, creating so-called *saliency maps* [51]. These saliency maps can be interpreted as overlays, which contain higher values on areas of the CT that contribute more towards the final prediction. Furthermore, we aimed to develop models that did not produce significant outliers, i.e. incorrect predictions close to 0 and 1. This improves the explainability of predictions to both applicants and physicians.

Results

Study cohort

In total, we retrospectively collected n=523 applications for financial support. Median age was 75 years (IQR 69—80). The dataset contained two female applicants (0.4%). The pool of pulmonologists consisted of n=23 experts, with 20 years of experience at the median (16—27). For each application, n=3 pulmonologists were assigned to process the application, with each pulmonologist having the same probability of getting assigned.

At the database lock of November 2019, n=16 did not receive an assessment of the panel. Of the n=507 remaining cases, n=233 applicants received a positive assessment (46.0%), with n=166 (71.2%) unanimously. The remaining n=274 applications did not meet the criteria, with n=219 (79.9%) unanimous assessments. Inter-observer variability between pulmonologists was estimated at *alpha*=0.65 (Krippendorff's alpha), with 75.9% unanimously (n=385).

For AI development, n=88 additional applications were excluded: n=78 for absence of fully imaged lungs in the CT and n=10 for CT slice thickness >5mm. A total of n=419 formed the study dataset. The excluded cases did not differ significantly by age or lung function. CT scans protocol were heterogeneous due to the multicenter origin of the data (median, CI): Voltage (120 kVp, 118.7—121.3), Tube current (194 mA, 177.4—210.6), Slice Thickness (3 mm, 2.85—3.14).

AI Training

We split the dataset into a training (n=263), validation (n=64), and test set (n=88), based on a train-test split of 80/20 [52], with a reproducible pseudo-randomization [sklearn v0.24.1]. We ran experiments with different label formats (i.e. *soft*, which reflects the agreement, and *hard*, which is binary), and with and without anomaly heatmap.

AI Predictive Performance

Soft labels combined with the anomaly heatmap yielded the best performance in all metrics (Table 2.2). Soft labels yielded a more uniform prediction distribution between 0 and 1 compared to hard labels (soft std=0.33, hard std=0.40, p<0.001). Following the McNemar test comparing the predictions, the soft label model yields higher performance (AUC=0.87, CI: 0.78—0.94, p<0.001) than the hard label model

(p=0.017) Moreover, soft labels with anomaly heatmap performed significantly better than soft labels without heatmap (p=0.042), indicating that both the soft labels and anomaly heatmap were required for increased performance. While the setup without anomaly heatmap and with hard labels scored best on sensitivity and negative predictive value, the overall performance of the soft label with anomaly heatmap was significantly better as well (p=0.017, McNemar test) (Table 2.2). The model yields an accuracy of 0.82 (0.74—0.90), with a sensitivity of 0.76 (0.62—0.88), and a specificity of 0.87 (0.77—0.96). Positive and negative predictive values were 0.84 (0.71—0.95) and 0.81 (0.69—0.91), respectively.

Label	**Anomaly Heatmap**	**ACC**	**SENS**	**SPEC**	**PPV**	**NPV**	**p-value**
Hard	No	0.65	**0.93**	0.40	0.58	**0.86**	0.017
Soft	No	0.66	0.78	0.55	0.60	0.74	0.042
Hard	Yes	0.65	0.46	0.81	0.68	0.63	0.043
Soft	Yes	**0.82**	0.76	**0.87**	**0.84**	0.80	**-**

Table 2.2: The results of the different tested setups of AI-models. The bold number shows the maximal performance in terms of the metric of that column. Hard labels are binary, while soft labels reflect the agreement of the pulmonologists on the panel. The p-values were calculated with the McNemar test compared to the best performing model (soft labels with anomaly heatmap).

The distribution of the predicted scores is shown in Figure 2.2A, stratified according to the agreement of pulmonologists in the panel (i.e. number of pulmonologists that gave a positive assessment). We performed visual analysis of outliers where all three pulmonologists were positive, but the model prediction was negative (n=5). Most of those applicants (n=4) had a severe reduction in lung function, but the fibrosis in the CT scans did not reflect this.

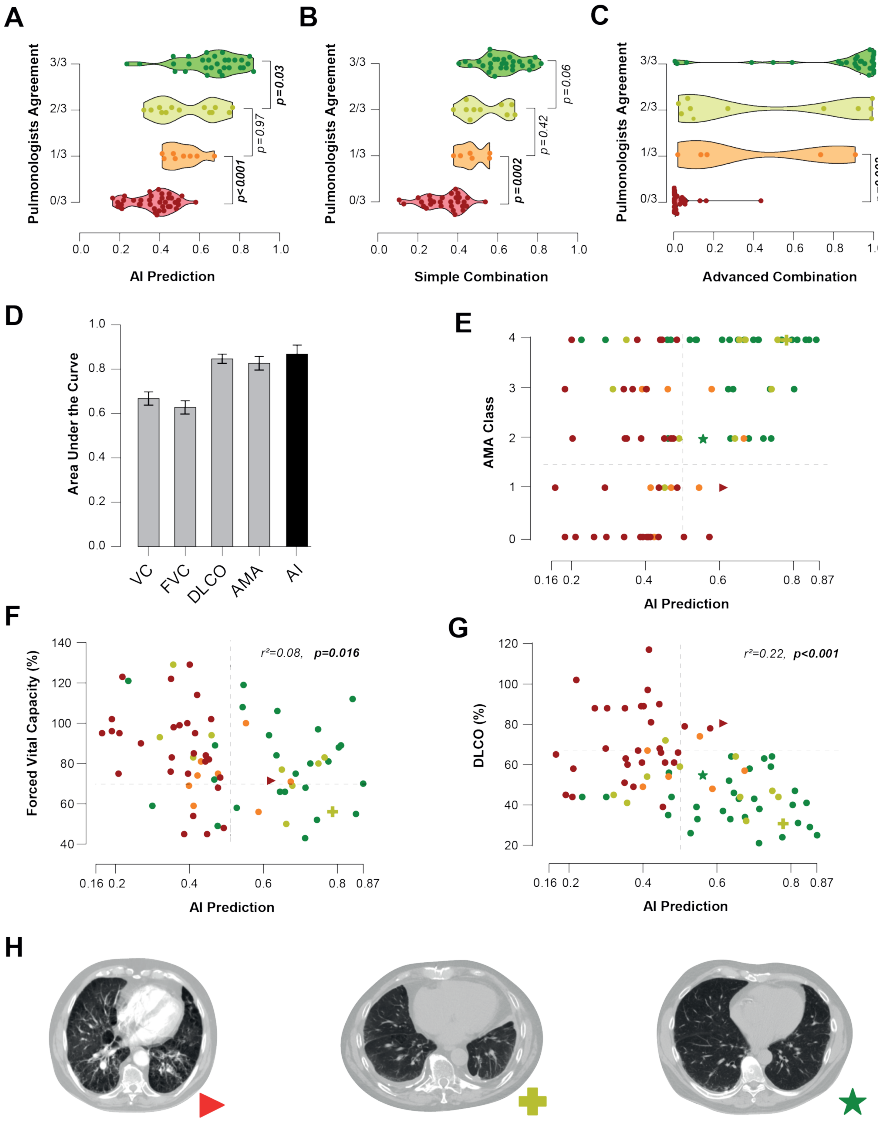
A
B
C
Pulmonologists Agreement
3/3
2/3
1/3
0/3
p=0.03
p=0.97
p<0.0001
p=0.06
p=0.42
p=0.002
p=0.67
p=0.06
p=0.002
0.0 0.2 0.4 0.6 0.8 1.0
AI Prediction
Simple Combination
Advanced Combination
D
Area Under the Curve
1.0 0.8 0.6 0.4 0.2 0.0
VC
FVC
DLCO
AMA
AI
E
AMA Class
4 3 2 1 0
0.16 0.2 0.4 0.6 0.8 0.87
AI Prediction
F
Forced Vital Capacity (%)
140 120 100 80 60 40
r²=0.08, p=0.016
0.16 0.2 0.4 0.6 0.8 0.87
AI Prediction
G
DLCO (%)
120 100 80 60 40 20
r²=0.22, p<0.001
0.16 0.2 0.4 0.6 0.8 0.87
AI Prediction
H

Figure 2.2: The colors reflect the agreement of the panel of pulmonologists: asbestosis negative (red dots), one out of three positive (orange), two out of three positive (light green), asbestosis positive (green dots). (A-C) Violin plots on different setups of prediction. The y-axis shows the agreement of the panel of pulmonologists. The x-axis shows the predicted probability of asbestosis. $p<0.001$ between the predictions in class 0 and 3 for all setups. (A) The prediction of the AI-model. (B) The score of the AI-model linearly weighted with the DLCO. (C) The prediction of the AI-model that took both the CT and the DLCO as input. (D) Bar plot of the diagnostic value (expressed as AUC) of the different lung function parameters, the AMA class, and the AI-model. (E-G) Probability of asbestosis predicted by the AI-model versus (E) AMA, (F) FVC, and (G) DLCO. The horizontal dotted line indicates the cut-off value for lung function loss, the vertical dotted line indicates the cut-off of the AI prediction. (H) Shows several cases where the amount of fibrotic tissue does not reflect the diagnosis of the pulmonologists. The symbols of each example are visualized in E-G when the respective lung function parameter of the patient is known.

Integration of Lung Function Tests

DLCO yielded predictive performance close to the AI-model (AUC=0.85, CI: 0.80—0.89, $p<0.001$). The remaining lung function parameters yielded lower results: AUC=0.67 for VC (CI: 0.60—0.72, $p<0.001$), AUC=0.63 for FVC (CI: 0.56—0.68, $p<0.001$), and AUC=0.83 for AMA (CI: 0.78—0.87, $p<0.001$).

Interestingly, the DLCO showed a weak correlation with the AI prediction ($r^2=0.22$, $p<0.001$), suggesting they can be independent predictors of asbestosis. We tested a simple combination, formalized as the average between AI-score and DLCO-value — $((1-DLCO)+AI)/2$; and an advanced combination strategy, where the DLCO is added as additional input to the AI-system.

The simple combination yielded an AUC of 0.95 (0.89—0.98, $p<0.001$), with an accuracy of 0.84 (0.76—0.92), a sensitivity of 0.77 (0.63—0.89), and a specificity of 0.91 (0.81—1.00). Positive and negative predictive

values were 0.91 (0.80—1.00) and 0.78 (0.65—0.90), respectively. Furthermore, from the distribution of the scores, this setup reported no false negative or false positive under 0.35 and above 0.60, respectively (Figure 2.2B).

The advanced combination strategy yielded an AUC of 0.92 (0.86—0.97, $p<0.001$), an accuracy of 0.84 (0.76—0.92), with a sensitivity of 0.74 (0.60—0.87), and a specificity of 0.94 (0.85—1.0). Positive and negative predictive values were 0.94 (0.83—1.0) and 0.77 (0.64—0.89), respectively. The spread of predictions in agreement with the pulmonologists was wider, as shown in Figure 2.2C. More specifically, this model predicts more CT scans closer to either zero or one than the AI and the simple combination do.

Following the AUC (Figure 2.2D) and outliers, the simple combination of AI+DLCO was considered the best model. Compared to the advanced combination, it yielded a distribution of predictions with lower standard deviation and was more interpretable due to DLCO weighting apart from the AI-model.

AMA Class Decomposition

To meet the requirement of lung function loss for financial reimbursement, AMA$\geq$2 is needed. Figure 2.2E shows how the AI prediction distributes over the AMA classes. When decomposing the AMA class in FVC and DLCO, the differences in predictive values become noticeable. FVC values (Figure 2.2F, AUC=0.63) were more scattered than the DLCO values (Figure 2.2G, AUC=0.87). The held out cases with FVC or DLCO reported (Figure 2.2H) show how the specific cases interact with the AI prediction and the lung function parameters in Figure 2.2E-G.

Visual Interpretation

To enhance interpretability, we generated saliency maps showing that the asbestosis-positive cases show more activations than the asbestosis-negative cases (Figure 2.3). From visual inspection, we can see that the CT scan with visible ILD yields more activations, indicating that the AI-system learned to identify ILD. In the CT scan where ILD is barely visible, there were hardly activations of the AI-system.

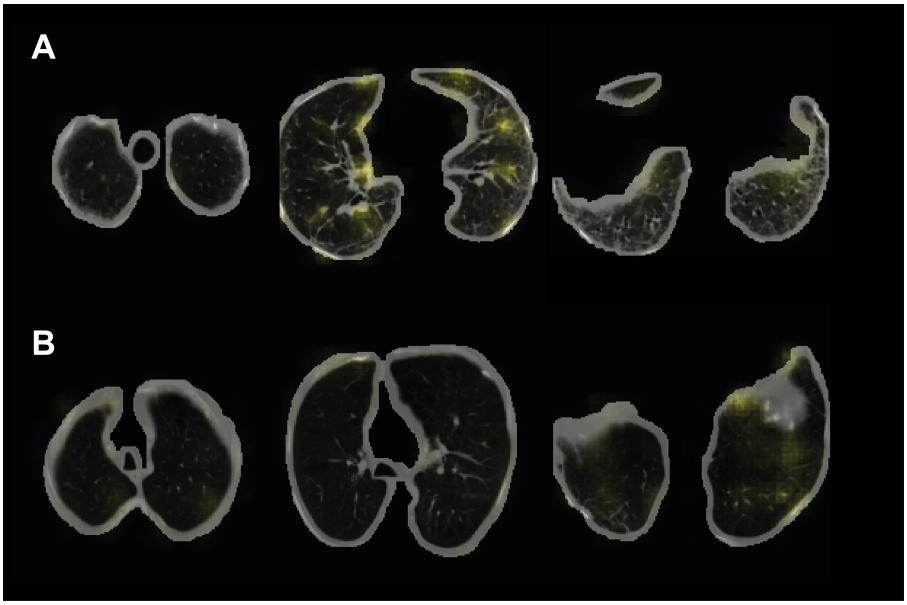

Figure 2.3: Saliency map yielded by the AI-model of two CT scans in the test set. The areas in yellow represent the attention of the model. The left side shows a slice from the top of the lungs, the middle a slice in the middle of the lungs, and the right side a slice from the bottom of the lungs. (A) CT scan where 3/3 pulmonologists were positive and the model yielded a high probability of asbestosis (0.81). (B) CT scan where 0/3 pulmonologists were positive and the model yielded a low probability of asbestosis (0.19).

Discussion

The current process for the assessment to determine the eligibility for financial support of workers who had been in contact with asbestos is laborious, costly, and has high intra-observer variability [14]. This study aimed to automate and standardize this process via artificial intelligence (AI) [53]. To do this, we have implemented an AI-system [47] that uses thoracic CT scans to replicate the assessment of a panel of three pulmonologists (as required by national law). Our AI-model to automatically classify the eligibility of applicants for state-aid for people with asbestosis yielded significant results and classified eligible applications with high accuracy. The best performing lung function parameter DLCO showed comparable results [54]. The combination of the AI+DLCO yielded a superior predictive performance than either AI or DLCO alone.

By accounting for the uncertainty in the pulmonologists' assessment (i.e. soft labels), our model reached higher accuracy than the same model that ignores it (i.e. hard labels). We hypothesize that the soft labels enable the AI to learn the uncertainty in specific cases, while the hard labels promote predicting either 0 or 100% probability. This is supported by the difference in standard deviation in the predictions of the soft/hard label AI-models. The level of agreement observed in the panel of pulmonologists is lower than the cut-off considered sufficient for reliable results [50]. AI-systems are notorious for their susceptibility to uncertainty in the provided labels [55]. The implementation of soft labels is based on the assumption that the levels of agreement between pulmonologists reflect a true, underlying level of uncertainty, which is also present in multidisciplinary meetings of interstitial expert teams [56]. The ability of the model to replicate the uncertainty suggested that it is not random but rather dependent on clinical or biological characteristics. Pure binary models would allow only for two outcomes: accept or reject. Due to the ability to replicate the uncertainty, we accepted a third outcome: process further. We envision the unsure cases (AI probability between 0.35-0.6) getting extra attention from the

panel, while one pulmonologist handles the clear cases (<0.35, >0.6). Therefore, there will always be a need for a (multidisciplinary) panel.

Lung function tests played a significant role in identifying false-negative cases where all pulmonologists returned a positive assessment, and the AI-model returned a negative one (Figure 2A-B: difference in top rows). This suggested that the lung function tests largely drove the verdict for these applicants. In other words, the AI-model could not detect a loss of lung function based on the CT scan of these applicants. This was supported by the weak-moderate correlation observed between DLCO and AI-model.

DLCO contributed to the diagnostic accuracy of our model, whereas the inclusion of FVC only deteriorated the ability to distinguish between positive and negative applications. Two reasons might explain this phenomenon: 1. the pulmonologists (unconsciously) based their verdict mainly on DLCO, while not taking FVC into account, and 2. decreased DLCO correlates with diffuse fibrosis, where the pulmonologists based their verdict mainly on the radiologic features of diffuse fibrosis. These findings align with Nogueira et al., who found that DLCO correlated most to the short-term progression of abnormalities in HRCT [54]. It may be beneficial for the panel to make DLCO measurements obligatory for more consistent, standardized, and objective evaluations.

Although AI-models contain biases on their own [57], they could help overcome human bias [58] and ensure a fairer public health policy in this situation. Our work aligns with current literature that suggests automatic AI-systems for ILD classification could improve patient healthcare [59].

Our study contained several limitations. Because of missing lung function parameters, each lung function performance was computed on slightly different sub-cohorts. Due to hardware limitations in the operational resolution of the AI-algorithm, we had to downsample the CT images, blurring finer-grained structures like fibrosis [60]. While the AI model performs excellently in classifying cases where the panel is anonymous, it lacks explainability. Saliency maps indicate where the

AI model is 'looking at' but are insufficient in explaining why a patient's application got accepted/rejected. There will always be a need for humans in these processes. Another improvement would be to include the cumulative asbestos exposure as input. Furthermore, our analysis was only retrospectively validated. Most AI algorithms are not validated prospectively [61], and the value of commercially available products is often not substantiated by peer-reviewed publications [62]. Therefore, we chose to validate our simple combination of the AI-model and the DLCO in a prospective setting (PROSBEST, Trial NL9064).

Given these results, we can envision an automatic and standardized diagnostic AI-system of the application based on the CT scan and lung function tests [63]. Further research in other clinical settings should reveal whether the method used might be useful in diagnostication of patients with interstitial lung disease in general.

Conclusion

We developed an AI-model to diagnose asbestosis in applicants for financial reimbursement according to parameters set by Dutch law. Classification models based on only the CT scan and a combination of the CT scan and the lung function test were quantitatively and qualitatively assessed. The model based on the CT scan and the DLCO was superior to the other models and reached excellent diagnostic accuracy. Whether this method could be implemented in other diagnostic settings for asbestosis or interstitial lung diseases is under investigation.

Acknowledgements

We are thankful for the data provided by the Dutch Institute for Asbestos Victims (IAS) and the Mesothelioma Working Party (SAGA) of the Netherlands Pulmonologists Organisation (NVALT). We would like to thank NVIDIA, the NVALT, and Maurits en Anna de Kock stichting for sponsoring GPUs. The authors state that this work has not received any funding.

Declarations

Funding

No funding was received for this project.

Guarantor

The scientific guarantor of this publication is Kevin Groot Lipman.

Conflict of interest

The authors of this manuscript declare relationships with the following companies: JAB is on the advisory board of Roche International (payment to institution) and received financial support and free drugs for an investigator-initiated study by MSD.

Statistics and biometry

No complex statistical methods were necessary for this paper.

Informed consent

Written informed consent was waived by the Institutional Review Board. Ethical approval from the Institutional Review Board approval was obtained.

Supplementary Materials

Application Procedure

Currently, the members of the committee give their approval for a positive asbestosis diagnosis if three criteria are met: 1) the patient has a sufficient history of occupational asbestos exposure, 2) the surface of the lung parenchyma in the CT scan of the patient is at least 5% covered with fibrosis, and 3) the patient has a reduced lung function. This is the legal diagnosis for asbestosis, rather than the clinical one. A committee of three pulmonologists evaluates whether the applicant fulfills the three criteria for a positive assessment. They are blinded to their respective diagnoses, and a unanimous decision is not required [14]. For the first criterion, a risk matrix was developed to state the intensity of asbestos of the most common occupations per decade, for the period of 1945-1995. More specifically, the years of work are multiplied by the corresponding intensity factor for the patient's occupations during that time, leading to an overall grade of the intensity of total asbestos exposure, which can be converted to fiber years. This value has to be higher than five fiber years to meet the criterion of sufficient history of occupational asbestos exposure. The second criterion of lung parenchyma fibrosis is evaluated through visual radiological inspection, where an experienced reader estimates the 3D volume of fibrosis, from the 2D slices of the CT scan. The fibrosis has to cover at least 5% of the pleural surface. The third criterion is lung function loss, which is estimated on a 5-point scale based on the criteria by the American Medical Association (AMA) and "Guides to the evolution of permanent impairment," 6th edition 2008. These guidelines describe the three most indicative parameters of lung function loss of applicants with asbestosis: (1) forced vital capacity (FVC), (2) diffusing capacity for carbon monoxide (DLCO), and (3) the maximal oxygen consumption (VO2 max). FVC is the total amount of air the patient can exhale by force after a full inhalation in liters. The DLCO describes the ability of carbon monoxide (as a substitute for oxygen) to transfer into the blood

in ml/min/mmHg. VO2 max is the maximal uptake of oxygen during incremental exercise in ml/min/kg. The lowest-scoring one determines the lung function loss category (Table 2.1). AMA class >1 is required to meet the third criterion. Besides the AMA classification and their corresponding lung function tests, the vital capacity (VC) is often given to assist the pulmonologists in their assessment of the lung function of the patient.

Network Design & Implementation

The 3D ResNet-18 architecture was implemented (Figure 2.4). It learned features from the CT scan (and corresponding anomaly heatmap) from 192 x 192 x 96 x 2 through multiple convolutions with striding operations to 6 x 6 x 3 x 512. The global average pooling layer compresses the feature maps to a vector representation. These 512 features are subsequently fed to the logistic classifier, which results in a corresponding probability of each class (e.g. asbestosis or no asbestosis). For the advanced combination, where the AI-system included the DLCO, an additional layer was implemented before the classification layer with four fully connected nodes to summarize the 512 pooled features of the CT image input. We implemented the lung function parameter value parallel to this layer and connected it to the classification layer. Each setup of the 3D ResNet network was trained using Tensorflow (v1.15.0) and Keras (v2.3.1) libraries on two NVIDIA GeForce RTX 2080Tis. The batch size was set at sixteen total, eight per GPU. Adam was used as optimizer, with an initial learning rate of 1e-3. The AI was trained for a maximum of 200 epochs, where early stopping was used to stop the training if the validation loss did not improve over 30 epochs. The best model checkpoint at the end of every epoch was performed. Data augmentation with rotation (up to 10°) around the longitudinal axis, and flipping over the sagittal plane of the image was implemented at runtime during training.

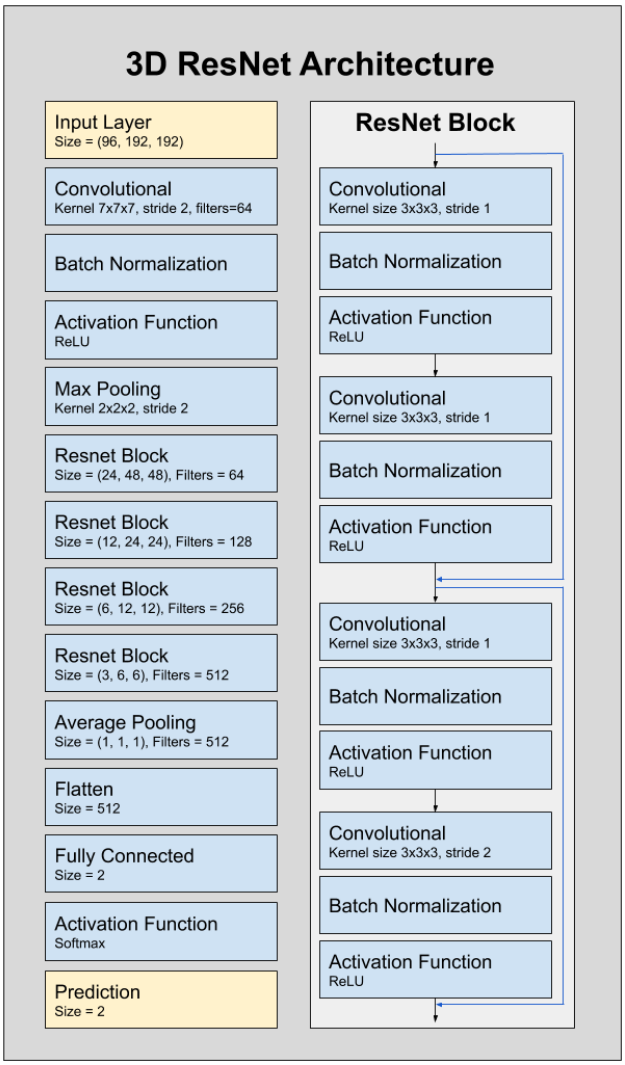

Figure 2.4: The architecture of the implemented 3D ResNet. The left column shows the encoder, where the image is downsampled through subsequent ResNet blocks to generate a prediction. The right column shows the ResNet block architecture. The black arrows represent the connections of the blocks. The blue arrows represent the identity connections, where the output of an activation layer is added to the input of another convolutional layer.

Variational Auto-Encoder - Anomaly Heatmap

Variational autoencoders (VAE) are types of networks that learn to identify common features, or characteristics, of a reference "normal" population. When presented with "abnormal" cases (i.e. cases that fall outside of this reference population), the algorithm will not be able to correctly estimate these features, resulting in a deviation between the algorithm-measured value and the actual value, i.e. an anomaly.

Variational Auto-Encoder - Dataset

To train a VAE to model healthy lung tissues, we collected a publicly-available CT dataset of lymphadenopathy patients [64]. CT slices containing labeled enlarged lymph nodes were removed, since the dataset should only contain healthy CT slices. The dataset contained N=867 patients, corresponding to a total of N=205 519 CT scan slices.

Variational Auto-Encoder - Data curation

To mitigate differences in imaging protocols, all CT density histograms were clipped between -1024 and 3072 Hounsfield Units (HU) and scaled on the interval [0, 1]. Slices were also resampled to 256 x 256 due to hardware constraints. To focus the attention of the VAE on the lungs, we performed segmentation of the lungs, and we blackened the background region. The segmentation was performed using a publicly-available deep learning segmentation network [45]. Lung segmentations were dilated through morphological operators with a kernel of 20 x 20 x 5 voxels to include adjacent tissue (i.e. thoracic wall) where pleural plaques are commonly found.

Variational Auto-Encoder - Network Design

The proposed network design follows the standard architecture of the variational autoencoder [46], where encoder, latent space, and decoder are placed in subsequent order. The encoder is composed of 6 convolutional blocks. Blocks are composed of repeated layers of convolutions, batch normalization, and the LeakyReLU activation function[65]. Downsampling is implemented through striding. The first block starts with 16 filters. Each subsequent block adds 16 filters. The decoder is composed of the mirrored architecture of the encoder, where the convolutional layer with stride 2 is replaced with a convolutional layer with a single stride and a subpixel upscaling layer [66] at the end of the convolutional block. Sigmoid is used on the last layer of the reconstruction to constrain the image on the interval [0, 1]. While there has been some advancement in the architecture, most notably the usage of fully-convolutional layers in the latent space for medical image reconstruction [67], we kept fully connected nodes in the latent representation. This might seem disadvantageous to spatial representations, but through internal experiments, we observed that the fully connected architecture prevents the VAE from reconstructing anomalies with patches and features learned from healthy tissue. Values in the latent space are reshaped to a 4 x 4 x 96 format and passed forward to the decoder part. The decoder upsamples this vector through convolutional layers and subpixel upscaling to reconstruct the full-size image.

Variational Auto-Encoder - Network Implementation

The VAE network was designed and trained using Tensorflow (v1.15.0) and Keras (v2.3.1) libraries on an NVIDIA GeForce RTX 2080Ti. N=195 519 slices were assigned to the training set and N=10 000 to the validation set for monitoring the training process. The batch size was set to 48. Adam was used as optimizer, with an initial learning rate of 1.5e-3. The VAE trained for 200 epochs, where the weight of the KL term in the loss was increased by 0.05 after each epoch, reaching

a maximum value of 1.0 in total. Best model checkpoint at the end of every epoch was performed. Data augmentation with rotation (up to 20) and horizontal flipping of the image was implemented during training.

Under Review

3

PROSBEST - Prospective evaluation of an Artificial Intelligence model for automatic classification of asbestosis for state-aid.

Illaa Smesseim*, Kevin B.W. Groot Lipman*, Stefano Trebeschi, Martijn M. Stuiver, Renaud Tissier, Jacobus A. Burgers, Cornedine J. de Gooijer

*Shared First Author

Journal of Thoracic Imaging 2023, doi: 10.1097/RTI.0000000000000759

4

Artificial Intelligence-based quantification of pleural plaque volume and association with lung function in asbestos-exposed patients

Kevin B.W. Groot Lipman, Thierry N. Boellaard, Cornedine J. de Gooijer, Nino Bogveradze, Eun Kyoung Hong, Federica Landolfi, Francesca Castagnoli, Nargiza Vakhidova, Illaa Smesseim, Ferdi van der Heijden, Regina G.H. Beets-Tan, Rianne Wittenberg, Zuhir Bodalal, Jacobus A. Burgers*, Stefano Trebeschi*

*Shared Last Author

Abstract

Purpose

Pleural plaques (PP) are morphological manifestations of long-term asbestos exposure. The relationship between PP and lung function is not well-understood, while the time-consuming nature of PP delineation to obtain volume impedes research. To automate the laborious task of delineation, we aimed to develop automatic Artificial Intelligence (AI)-driven segmentation of PP. Moreover, we aimed to explore the relationship between pleural plaque volume and pulmonary function tests (PFT).

Methods

Radiologists manually delineated pleural plaques retrospectively in CT images of patients with occupational exposure to asbestos (May 2014 - November 2019). We trained an AI model with a nnUNet architecture. Dice Similarity Coefficient (DSC) quantified the overlap between AI and radiologist. The Spearman correlation coefficient (r) was used for the correlation between PP volume and PFT metrics. When recorded, these were Vital Capacity (VC), Forced Vital Capacity (FVC), and Diffusing Capacity for Carbon Monoxide (DLCO).

Results

We trained the AI system on 422 CT scans in five folds, each time with a different fold (n=84-85) as a test set. On these independent test sets combined, the correlation between the predicted volumes and the ground truth was r=0.90, and the median overlap was 0.71 DSC. We found weak to moderate correlations with PP volume for VC (n=80, r=-0.40) and FVC (n=82, r=-0.38), but no correlation for DLCO (n=84,

r=-0.09). When the cohort was split on the median PP volume, we observed statistically significantly lower VC (p=0.001) and FVC (p=0.04) values for the higher PP volume patients, but not for DLCO (p=0.19).

Conclusion

We successfully developed an AI algorithm to automatically segment PP in CT images to enable fast volume extraction. Moreover, we have observed that PP volume is associated with loss in VC and FVC.

Introduction

Pleural plaques (PP), a specific manifestation of asbestos exposure, often appear on the parietal pleura as localized hyalinized collagen fibers in calcified or non-calcified forms [84, 85, 35]. The exact mechanism of PP formation remains unclear [85, 86]. However, the likelihood of developing PP is associated with the duration and cumulative exposure to asbestos [87]. Despite this, PP can also form after minimal exposure [88].

Patients with PP are typically asymptomatic [7]. Discrepancies exist between a systematic review indicating no statistically significant association between PP and PFT [17] and a study demonstrating a small, statistically significant impact on lung function [18]. Thoracic computed tomography (CT) enables PP extension measurement with excellent intraobserver reproducibility (ICC: 0.98) and good interobserver variability (ICC:0.93) [89]. However, manual segmentation of volume is time-consuming and impractical for large population studies or clinical workflow integration [90]. Consequently, the impact of PP on lung function remains inconclusive.

Public health policies in many countries provide financial support for patients with mesothelioma or asbestosis following occupational asbestos exposure [14, 19]. However, few policies consider pleural plaques due to the lack of evidence supporting a clinically significant loss in lung function [17]. Even with confirmation, manual volumetric assessment would be incompatible with the current radiological workflow [90].

An alternative method for segmentation and volume quantification is needed to facilitate extensive population studies and clinical implementation of PP volume measurements. This method should be fast, accurate, and reproducible. Artificial Intelligence (AI)-based automated segmentation could provide a potential solution by learning to identify patterns in CT scans and yield a volumetric measurement of PP in seconds.

This study aims to develop an AI algorithm for the automatic segmentation of PP and examine the relationship between PP and lung function impairment. The resulting algorithm will enable researchers to investigate the correlation between PP volume and lung function, providing a proof of concept for a clinically compatible, quantitative PP-volume test.

Material and Methods

Datasets

We performed a retrospective analysis on a dataset of people applying for state financial support, between May 2014 and November 2019 [14]. This dataset is comprised of a cohort of applicants who are required to submit a CT scan acquired from their respective local hospital, along with a PFT. The dataset was collected by the Instituut Asbestslachtoffers (IAS) and Section Asbestos Related Disease (SAGA). Inclusion criteria were fully imaged lungs on CT with slice thickness $\leq$ 5mm. Thoracic CTs were collected from multiple hospitals across the country, resulting in heterogeneous data (median, CI): Voltage (120 kVp, 118.7—121.3), Tube current (194 mA, 177.4—210.6), Slice Thickness (3 mm, 2.85—3.14). Vendors, reconstruction kernels, and contrast usage are listed in Table 4.1. CT scans were acquired with breath-hold at mid-respiratory or inspiratory volume. As part of the financial support compensation procedure, three independent pulmonologists determined whether significant fibrosis was present, defined as >5% of lung parenchyma [14].

Applicants signed a written informed consent that their data could be used for systematic analyses. The project was approved by the institutional scientific board (IRBd19-136) and performed in accordance with the Declaration of Helsinki. The de-identification process for the data was executed in compliance with the DICOM standard, utilizing proprietary software developed within the institution.

Manufacturer	Convolution Kernel	Count
GE MEDICAL SYSTEMS	BONE	1
GE MEDICAL SYSTEMS	BONEPLUS	10
GE MEDICAL SYSTEMS	CHST	7
GE MEDICAL SYSTEMS	LUNG	27
GE MEDICAL SYSTEMS	SOFT	1
GE MEDICAL SYSTEMS	STANDARD	7
Philips	A	4
Philips	B	41
Philips	C	15
Philips	E	3
Philips	IMR1,SharpPlus	3
Philips	L	34
Philips	YA	3
Philips	YC	10
Philips Medical Systems	5	1
SIEMENS	B30 - B45, I30 - I45	99
SIEMENS	B60 - B80, I50 - I70	51
SIEMENS	Bl57	3
SIEMENS	Bl64	1
SIEMENS	Br40	2
SIEMENS	Br69	1
SIEMENS	Ub44u	1
TOSHIBA	BODY	1
TOSHIBA	FC02 - FC35	42
TOSHIBA	FC51 - FC86	30
TOSHIBA	LUNG	1
Unknown	Unknown	23

Table 4.1: Technical parameters of the included CT scans, filtered by manufacturer and reconstruction kernel.

PFT for patients in the study dataset were retrospectively retrieved. When recorded, the parameters were: Vital Capacity (VC), Forced Vi-

tal Capacity (FVC), and Diffusing Capacity of Lung for Carbon Monoxide (DLCO). The PFT data was acquired from spirometry tests in upright position with expiratory measurements and converted to percent predicted values following Global Lung Function Initiative 2012 reference equations for spirometry [91, 92].

Segmentation Procedure

A team of five board-certified radiologists (TB, NB, EKH, FL, FC) manually segmented the pleural plaques using 3D Slicer v4.11 [49], with the workload was split equally among them. The time per segmentation was not recorded. However, readers mentioned that segmentation took 30-60 minutes per scan. Calcified and non-calcified portions of the plaques were both segmented as one single segmentation. One technical physician [93] with two years of experience in thoracic CT imaging (KGL) reviewed all CT scans and segmentations and forwarded inconsistent segmentations to another team of radiologists (NV, IS, RW, TB). They adjusted the segmentations of suboptimal quality (segmentation artifacts, missing plaques, etc.). To analyze the AI segmentation performance of calcified versus non-calcified plaques, we set an empirical threshold of 120 HU to differentiate between them as a postprocessing step.

Design of the AI algorithm

We implemented the AI algorithm following the design of the no-new-UNet (nnUNet) [94]. This represents the state-of-the-art in image segmentation, with the algorithm leveraging several preprocessing techniques and training procedures. The nnUNet system automatically determines the optimal Convolutional Neural Network (CNN) architecture and other hyperparameters based on the characteristics of the dataset (i.e. the thoracic CT scans). During training, a 'patch' equal to the input size of the model is retrieved from the CT scan, and the algorithm

iterates over these patches until the entire CT scan is analyzed. The configuration chosen for the architecture was 3D full resolution with training procedure (trainer) nnUNetTrainerV2. A schematic overview of the architecture is shown in Figure 4.1.

Data Preprocessing and Training Procedure

We split the dataset into a training (n=337, 80%) and a test set (n=85, 20%), based on a random reproducible split. All CT scans were resampled to [0.71, 0.71, 1] (x, y, z) spacing with a patch size of [160, 160, 96] (x, y, z). The training procedure consisted of 1000 epochs with a batch size of 4. The loss function was a combination of the dice loss and cross-entropy. To test whether the ensemble five-fold cross validation outperformed the single model trained on all data, we ran experiments with and without internal cross-validation.

AI Model Evaluation

The segmentation performance of the trained AI model was evaluated using the Dice Coefficient Score (DSC), which is a quantitative measure to determine the overlap between the predicted segmentation by the AI, and the ground truth. The higher the overlap between the two, the better the performance of the AI. In addition to the DSC, we calculated the correlation between the volume predicted by the AI model and the ground-truth volume derived from the segmentation of the expert readers. This allowed us to identify possible systematic errors of the model, and the presence of outliers. To test whether the AI can measure the PP volume in the CT scans correctly, we used the different percentiles to convert the segmentation task to a classification problem. Here, we monitor whether the AI classified the CT scans as containing a higher or lower PP volume than the cut-off, compared to the radiologist's segmentation.

Lung Volume Assessment

The lung volume was quantified using an external AI model for lung segmentation [45]. The output generated by the model was then manually reviewed and corrected, if needed, by K.G.L. using the 3D Slicer software. This particular AI model was selected because it demonstrated adequate accuracy and robustness during internal evaluation, generalizing reasonably well to fibrotic tissue. This attribute made corrections for fibrotic lungs more feasible compared to other models, establishing it as a reliable choice for our research study.

Association between PP and decreased lung function

Due to the slow growth rate of pleural plaques [85], we do not expect significant volume differences over distinct periods of months. As a result, we performed the analysis using PFT data from patients within one year, measured from the date of the CT scan. To determine whether an increase in pleural plaque volume is associated with decreased lung function, we calculated the correlation between PP volume and lung function parameters, and tested for significant differences in lung function for groups at different cut-offs of PP volume, namely the 25th, 50th, and 75th percentiles. Given that the lung function parameters are normalized in percent predicted values, we normalized the PP volume as well through the total lung volume of the patient. The normalized PP volume consisted of the PP volume divided by the lung volume of the patient. Differences between FVC and VC may indicate air trapping or small airway collapse. Therefore, we tested this difference versus the PP volume. Patients with diffuse fibrosis were excluded for correlation between the lung function and the volume of pleural plaques since fibrosis is a confounding variable [95].

Statistical Analysis

Since the PP volumes and PFT results were not normally distributed, the Spearman r was calculated. We applied the Mann-Whitney U test to test the differences between the cross-validated model and the single model on the same test set. Differences in PFT between groups with different PP volumes were assessed via the Wilcoxon signed-rank test. The 95% confidence intervals (CI) were calculated via bootstrapping with replacement. Bonferroni correction was applied when multiple tests were conducted. Bonferroni corrections were applied to account for the three distinct tests conducted across various quartiles of PP volume in relation to PFT. This adjustment resulted in a significance level (α) of $p < 0.05/3 = 0.017$.

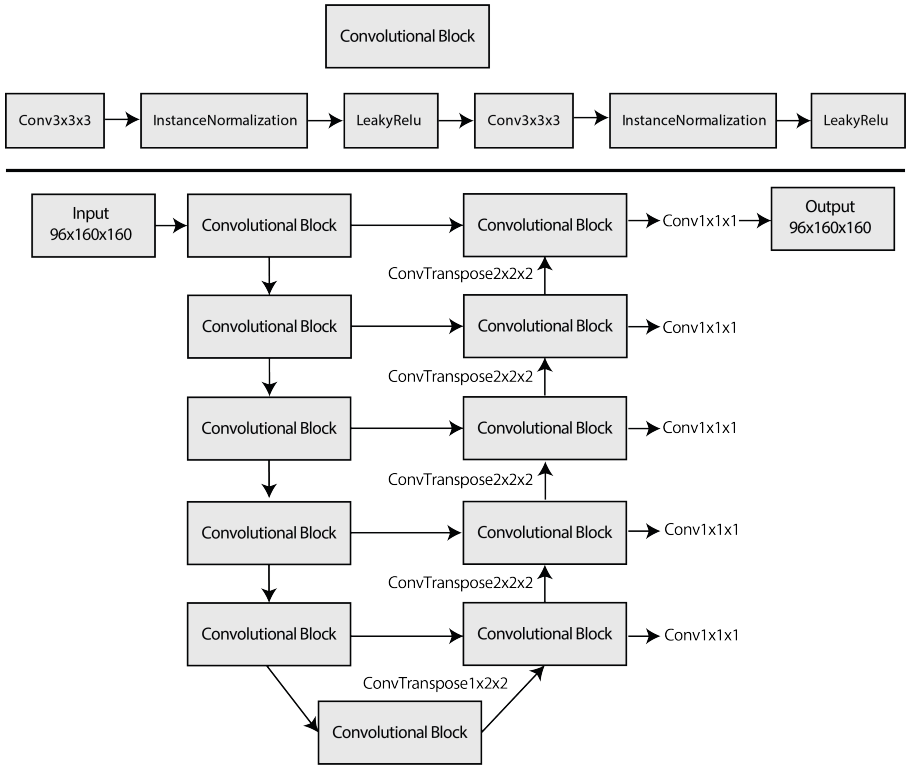

Figure 4.1: Schematic overview of the nnUNet architecture based on the characteristics of the pleural plaque dataset. Top: Details of the convolutional block used throughout the model. Bottom: Overview of the total architecture. The input size (slices, y, x) is equal to the output size, referred to as the patch size.

Results

Study Cohort

We retrospectively collected n=523 applications for asbestosis government support. The median age of the applicants was 75 years (IQR 69 — 80), and applicants were almost exclusively male (2 females, 0.4%). Applications were excluded due to the absence of CT scans (n=74) and any PP (n=27), yielding a total dataset of n=422 CT scans (n=303 with contrast). All scans were segmented by radiologists and reviewed. Three PFTs were collected when available: VC (n=393, median 79, IQR 64 — 96), FVC (n=398, median 78, IQR 64 — 95), DLCO (n=408, median 57, IQR 44 — 71). There was no statistically significant difference observed between the total cohort and the cohort after exclusion in terms of age (median: 73 years versus 74 years, p=0.39) or PFTs (VC: 74% versus 74%, p=0.47; FVC: 79% versus 79%, p=0.39; DLCO: 55% versus 55%, p=0.43).

Interrater variability of adjusted cases

In total, n=68 segmentations were admitted for review due to inconsistencies, and all of them were adjusted after inspection by the radiologist. For these adjustments, we observed the following medians: DSC of 0.61, sensitivity of 0.48, initial volume of 37.4 cm^3, and a corrected volume of 81.8 cm^3. Figure 2 shows several examples of adjusted annotations, with reasons such as partly segmented pleural plaques, using lung window during segmentation, and missing pleural plaques.

Cross-validation versus Single Model

The first experiment consisted of five-fold cross-validation (standard nnUNet procedure, ensemble model) and another experiment of a single

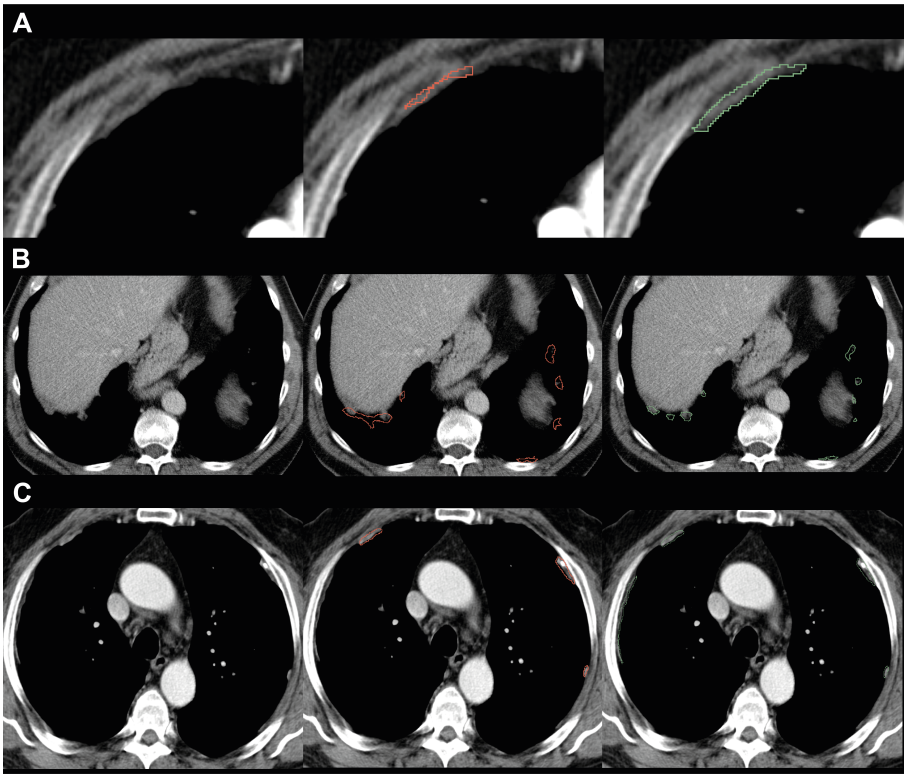

Figure 4.2: Several examples of adjusted segmentation after inconsistencies were noted. The first column are the CT scans; second column the segmentation of the first radiologist; third column the revised segmentation. (Row A) CT scans in axial plane with contrast; pleural plaque only partially segmented. (Row B) CT scans in axial plane without contrast; pleural plaques segmented on lung window, leading to overestimation of the volume. (Row C) CT scans in axial plane with contrast; missed pleural plaque.

training procedure with all training data (single model). The ensemble model reached a median DSC of 0.70 (0.66—0.73), on par with the single model with a median DSC of 0.70 (0.69—0.74), p=0.60, evaluated on the same independent test set.

Single Models on Different Test Sets

We trained multiple models to study the influence of the chosen test set, with each patient in the independent test set once. Therefore, the nnUNet architecture was trained five times, each with a different, random, reproducible split. We made all trained algorithms available.[1]

All n=5 AI models yielded similar performances over the individual scans in the test set with a median DSC of of 0.70 (0.69—0.74), 0.72 (0.67—0.73), 0.71 (0.67—0.73), 0.72 (0.69—0.75), and 0.71 (0.68—0.75) (Figure 4.3A, Table 4.2). No statistically significant differences existed between the test set results ($p>0.05$). Combining the predictions on all test sets, the median DSC is 0.71 (0.70—0.73). Overall median sensitivity on the combined test set is 0.78 (0.74—0.80). In terms of PP volume, the difference between the volume segmented by experts (median: 104.0 cm^3, CI: 86.9—119.9 cm^3) versus the AI models (median: 121.8 cm^3, CI: 101.6—136.1 cm^3) did not reach the level of statistical significance ($p=0.09$). The mean absolute error was 29.7 cm^3 (CI: 23.5—35.7). AI-predicted volume and the segmented volume showed a strong correlation (spearman $r = 0.90$, CI: 0.88—0.92, $p<0.001$) (Figure 4.3B). The difference between radiologists and AI segmentation increased as the segmented volume of the radiologists increased (Figure 4.3C).

We visualize several cases with different quality of segmentation in Figure 4.4A-C. Segmentation of the calcified pleural plaques yielded a DSC of 0.92 (0.91—0.93), sensitivity of 0.96 (0.95 - 0.97), with a significant difference ($p<0.0001$) between AI predicted volume of 27.38 cm^3 (21.13 - 32.40 cm^3) and the expert derived volume of 23.72 cm^3 (19.62 - 29.75 cm^3). The non-calcified part of PP yielded a DSC of 0.62 (0.60 - 0.64), sensitivity of 0.69 (0.66 - 0.72), where the difference between AI predicted volume of 81.58 cm^3 (70.54 - 93.81 cm^3) and the expert volume of 74.34 cm^3 (61.15 - 89.31 cm^3) was not statistically significant ($p=0.28$). The AI was able to classify all percentiles with excellent performance (25th percentile: AUC=0.94 (0.92—0.97), $p<0.0001$, 50th percentile:

[1] `https://github.com/nki-radiology/pleural-plaques`

AUC=0.95 (0.93—0.97), p<0.0001, 75th percentile: AUC=0.95 (0.93—0.97) p<0.0001).

#	DSC	SENS	R-Vol	AI-Vol	AE-Vol	p-value
1	0.70 0.69–0.74	0.78 0.71–0.83	102.1 75.5–185.5	104.7 80.5–136.5	23.0 17.4–30.4	<0.0001
2	0.71 0.67–0.73	0.75 0.71–0.81	102.4 75.3–125.9	98.4 80.7–153.1	41.4 22.6–53.6	<0.0001
3	0.72 0.69–0.75	0.78 0.70–0.84	87.58 63.9–112.6	104.9 70.1–144.5	29.7 21.9–37.7	<0.0001
4	0.72 0.67–0.73	0.77 0.69–0.81	87.43 67.3–140.0	119.6 72.3–152.5	28.1 17.5–46.1	<0.0001
5	0.71 0.68–0.75	0.77 0.74–0.83	143.7 107.7–179.9	161.8 121.8–186.7	32.0 22.0–37.9	<0.0001

Table 4.2: Metrics of each of the individual trained models reported in median and 95% confidence interval. # = Model number, DSC = Dice Similarity Coefficient, SENS = Sensitivity, R-Vol = PP volume segmented by the radiologists, AI-Vol = Volume segmented by the AI model, AE-vol the absolute volume difference between radiologist and AI segmentation, p-value is calculated with Wilcoxon paired test between R-Vol and AI-Vol. Volume is in cm^3.

Comparison with Pulmonary function tests

The dataset contained n=188/423 patients without diffuse fibrosis, of which n=106 patients had a PFT within a year of the CT date. We collected the VC (n=80), FVC (n=82), and DLCO (n=84), where n=50 patients had complete data for all three measurements. Figure 4.5A-C shows the relation of each parameter to the PP volume segmented by the radiologists, whereas Figure 4.5D-F shows the relation with AI segmented volume. PP volume segmented by the radiologists was moderately negatively correlated with VC (r=-0.40, CI: -0.54—-

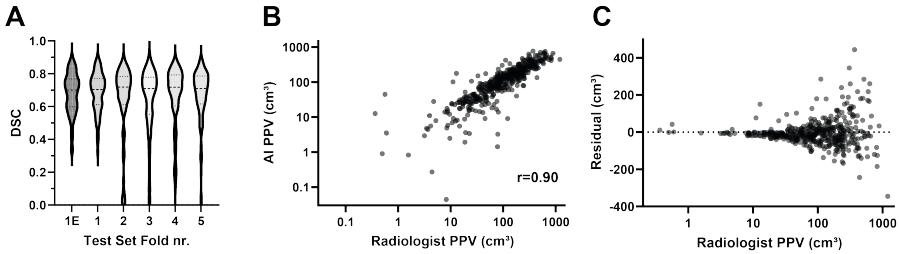

Figure 4.3: (A) Dice Similarity Coefficient (DSC) distribution over the AI models, each with a different 20% as test set, including the ensemble method (1E). (B) Correlation between the combined test sets of AI-predicted pleural plaque volume (PPV) and the radiologist segmented PPV. (C) The x-axis denotes the radiologist segmented PPV, the y-axis represents the difference between the radiologist and the AI. The higher the radiologists' segmented volume, the larger the difference.

0.22, p=0.0003) and FVC (r=-0.38, CI: -0.52—-0.21, p=0.0005), but not correlated with DLCO (r=-0.09, CI: -0.25—0.08, p=0.39). All panels show a non-linear relation between the PP volume and lung function, where a high PP volume suggests an association with low lung function values. The normalized PP volumes by lung volume were moderately negatively correlated for VC (r=-0.45, CI: -0.60—-0.28, p<0.0001) and FVC (r=-0.42, CI: -0.57—-0.25, p<0.0001), but no statistically significantly correlation was observed for DLCO (r=-0.11, CI: -0.26—0.05, p=0.30). No correlation was found between normalized PP volume and the difference between VC and FVC (r=0.00, CI: -0.20—0.21, p=0.60). By splitting the PP volume distribution on the different quartiles (45, 106 and 229 cm^3), we observed a statistically significant lower VC and FVC for the higher PP volume group (Table 4.3). DLCO did not yield any statistically significant difference.

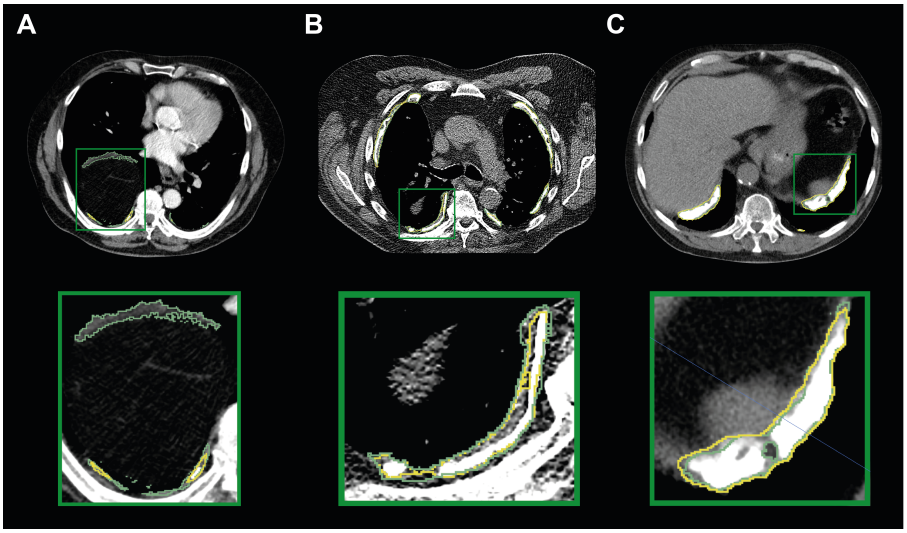

Figure 4.4: CT scans in axial plane with contrast; example of the lower value of Dice Similarity Coefficient (DSC) between AI in yellow outline, and radiologist in green outline (DSC=0.43). (B) CT scans in axial plane without contrast; average segmentation performance (DSC=0.75). (C) CT scans in axial plane without contrast;well-segmented plaques on the diaphragm (DSC=0.85).

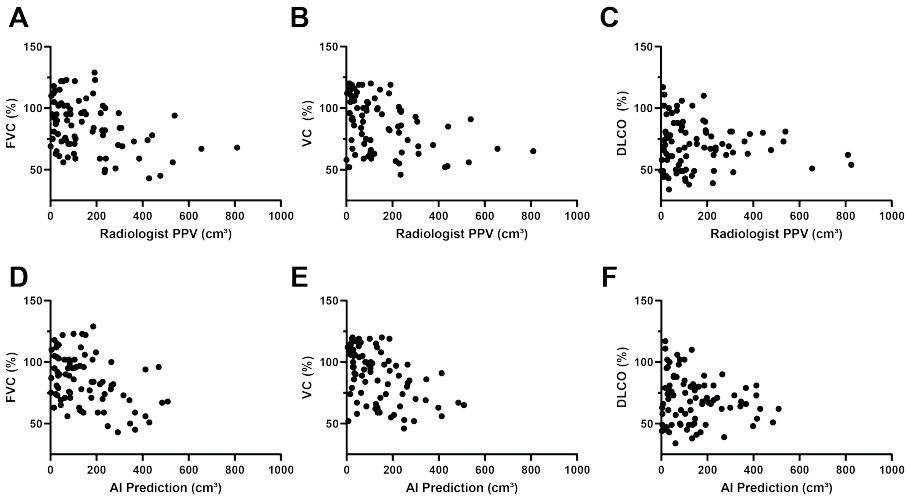

Figure 4.5: (A) Scatterplot of the Forced Vital Capacity (FVC) versus the Pleural Plaque Volume (PPV) by the radiologist. (B) Vital Capacity (VC) versus PPV. (C) Diffusing Capacity for Carbon Monoxide (DLCO) versus PPV. (D) Scatterplot of the FVC versus the PPV by the AI models. (E) VC versus PPV by AI. (F) DLCO versus PPV by AI.

PFT	Perc	PPV	n upper	n lower	Mean upper	Mean lower	p
VC	25	34.7	60	20	84.6	96.3	**0.016**
VC	50	98.8	40	40	80.3	94.6	**0.001**
VC	75	228.2	20	60	73.6	92.1	**<0.001**
DLCO	25	37.3	63	21	69.5	71.0	0.49
DLCO	50	103.5	42	42	67.4	72.4	0.19
DLCO	75	225.5	21	63	66.8	70.9	0.27
FVC	25	41.3	61	21	83.2	91.4	0.055
FVC	50	104.8	41	41	81.0	89.6	0.037
FVC	75	233.3	21	61	69.0	90.6	**<0.001**

Table 4.3: Difference in pulmonary function test (PFT) based on several cut-offs of pleural plaque volume (PPV) in cm^3. VC = Vital Capacity, FVC = Forced Vital Capacity, DLCO = Diffusing Capacity of Lung for Carbon Monoxide. Bonferonni correction has been applied (significance level=0.017).

Discussion

In this study, we proposed an AI algorithm for fast, automatic assessment of PP volume. Our goal was to design an automated segmentation model for pleural plaques to enable further research on the impact of PP volume on patients. The segmentation results suggested an adequate ability of the algorithm to replicate the expert reader's segmentation and estimate the PP total volume. We showcased how the algorithm can enable researchers to test the relation between PP volume and PFTs in a dataset of former asbestos workers applying for government support. A non-linear association between vital capacity, forced vital capacity, and both PP volume and PP volume corrected for lung volume was observed, exceeding the relation found in current literature [17, 18, 90, 96].

To the best of our knowledge, we are the first study providing an automatic segmentation tool for future research in PP and asbestos exposure. In our study, we use state-of-the-art 3D segmentation and a larger dataset to obtain higher accuracy and precision, and share it freely online for the scientific and medical community to use. We showed that the ensemble method did not outperform a single model training procedure for this dataset. Interestingly, the other folds yielded a non-significant higher median DSC than the first fold (both ensemble and single model) but also produced outliers with DSC scores between [0, 0.2]. No outliers would have been reported if only the first fold results were published. However, by running multiple experiments on different test sets, we showed that four out of five folds yielded outliers, leading to a moderate median DSC over all test sets. A potential reason for the difference between AI and expert segmentation are the different acquisition and reconstruction protocols in the dataset (Table 4.1), where the AI model does not generalize sufficiently. Having multiple radiologists independently delineate PP without consensus could be another reason, where each radiologist would segment PP differently. While the AI outperformed the interobserver variability of the worst annotated cases

(the revisions), the study could not analyze the overall interobserver variability, nor the intraobserver variability.

In related work, another study investigated pleural plaque segmentation on 5 mm slices with deep learning [20] but didn't investigate association with PFT. A study that investigated the association included 26 patients, who were divided into three groups of <10 mL, 10-20 mL, and >20 mL PP volume, where no statistically significant differences were found between the groups in terms of lung function values [90]. We defined different cut-offs to determine the high and low PP volume groups. Our lowest cut-off of the 25th percentile was 39.6 cm^3 (or mL) for VC, where a similar study defined the highest volume group of plaques as >20 cm^3 [11], representing a substantial shift in our understanding of the extent of the disease. Another study measured PP volume of 75 patients on three axes and could not correlate this volume to lung function, exercise capacity, and cumulative asbestos exposure [96]. The full volumetric measure, instead of a surrogate measure of the three longest diameters [96], seems to be a keypoint in the understanding of the relation between pleural plaques and lung function.

To demonstrate the usability of the algorithm to study lung function, we presented an example with FVC, VC, and DLCO. In our showcase, significant differences were observed in PP volume in relation to both FVC and VC. The difference between FVC and VC is the forcefulness of exhalation. When discrepancies occur, it could indicate airway resistance, for example. The results suggest that PP volume does not lead to differences between FVC and VC, since we observed no correlation. The total lung capacity (TLC) is the volume of gas in the lung at the end of full inspiration. A decreased TLC reflects a restrictive lung disorder. It is the sum of the inspiratory reserve volume (IRV), tidal volume (Vt), expiratory reserve volume (ERV), and residual volume (RV). The (F)VC is the volume of exhaled air after maximal inspiration, consisting of the Vt, ERV, and IRV. A reduction in (F)VC can indicate restrictive lung disease, which can be categorized as an intrapulmonary (parenchymal) disease, such as lung fibrosis. Therefore, a possible cause of the observed decreased (F)VC in our patient group is that PP volume reduced the

expansion of the lungs, which decreases the inspiratory reserve volume. DLCO (gas exchange) is less affected by total air inhalation [97], which may explain the non-significant relation.

If, by means of our algorithm, further studies are able to unveil the relation between quality of life and extent of PP, and to confirm a decrease in quality of life, financial compensation programs for patients with pleural plaques in more countries might arise. For example, the United Kingdom canceled its compensation for PP in 2007 due to the absence of evidence that pleural plaques impeded lung function [98]. Moreover, if, in the future, governments might decide that from a certain PP volume lung function loss endorsed for compensation, our model might be used to detect whether that PP volume threshold is reached. This avoids the labor-intensive and time-consuming work of the radiologists that would otherwise have to segment the plaques. In such a workflow, a radiologist should evaluate the segmentation of the AI model and approve or adjust it for finalization. An intuitive graphical user interface to interact with the AI segmentation should therefore be developed.

From a clinical perspective, a completely automated and precise model has the potential to monitor alterations in PP volume over time. Notably, if specific areas of PP demonstrate accelerated growth, it may be suggestive of pleural mesothelioma [99]. Given the typical late-stage detection of mesothelioma [100], this method might provide an active surveillance approach for patients with PP who have been exposed to asbestos.

There are limitations to this study. First, a substantial portion of the PP segmentations was revised, which indicates a high interobserver variability among the radiologists, unlike the findings of another study that found minimal interobserver variability [89]. Radiologists reported that in CT scans with suboptimal quality, it was hard to distinguish pleural plaques from other structures, leading to interobserver variance. CT scans at mid-respiratory and inspiratory volume were included, which may bias the lung volume measurement, and therefore subsequently the PP volume versus lung volume ratio. The suboptimal DSC of the

otherwise excellent performing nnUNet architecture could have several reasons: poor generalizability over different reconstruction kernels, vendors, or resolutions; high interrater variability among readers, leading to inconsistent segmentations as ground truth; or suboptimal segmentation guidelines (e.g., window selection, decision-making in uncertain cases regarding whether to segment or not). Furthermore, in our showcase, we could not correct for confounders in the correlation between the lung function parameters and the volume. Moreover, lung function parameters were already in percentage of predicted value, resulting in a complex comparison with an absolute measure such as PP volume. Therefore, an additional analysis was performed where the PP volume was corrected for lung volume. While we did exclude patients with substantial pulmonary fibrosis, lung function parameters for other confounders (e.g. asbestos exposure, BMI, and smoking [17]) could not be corrected since that information was unknown. An extensive analysis of the correlation between lung function and PP extension is beyond the scope of this study. The algorithm is available online, for other researchers to use, replicate our results, and study the influence of confounders.

Conclusion

In this study, we trained an AI model for the automatic segmentation of the pleural plaques in CT scans to estimate the volume. The segmentations were quantitatively and qualitatively adequate and showed a high correlation to the segmentation of expert readers. Moreover, we showed that higher pleural plaque volumes are significantly associated with a decreased FVC and VC, but not with DLCO. The AI model is publicly available and can be used to decrease or eliminate the workload for the expert readers, and to study the relation between pleural plaques and lung function more extensively.

Acknowledgements

We express our gratitude to the Dutch Institute for Asbestos Victims (IAS) and the Section Asbestos Related Disease (SAGA) of the Netherlands Pulmonologists Organisation (NVALT) for supplying the data. Our appreciation extends to NVIDIA and the NVALT for generously providing GPU sponsorship.

Declarations

Funding

No funding was received for this project.

Guarantor

The scientific guarantor of this publication is Kevin Groot Lipman.

Conflict of interest

The authors of this manuscript declare relationships with the following companies: JAB is on the advisory board of Roche International (payment to institution) and received financial support and free drugs for an investigator-initiated study by MSD.

Statistics and biometry

No complex statistical methods were necessary for this paper.

Informed consent

Written informed consent was waived by the Institutional Review Board. Ethical approval from the Institutional Review Board approval was obtained.

European Radiology 2023, doi: 10.1007/s00330-022-09303-3

5

Is the generalizability of a developed Artificial Intelligence algorithm for COVID-19 on chest CT sufficient for clinical use?

Laurens Topff*, Kevin B. W. Groot Lipman*, Frederic Guffens, Rianne Wittenberg, Annemarieke Bartels-Rutten, Gerben van Veenendaal, Mirco Hess, Kay Lamerigts, Joris Wakkie, Erik Ranschaert, Stefano Trebeschi, Jacob J. Visser, Regina G. H. Beets-Tan; ICOVAI

* Shared First Author

Abstract

Objectives

Only few published artificial intelligence (AI) studies for COVID-19 imaging have been externally validated. Assessing the generalizability of developed models is essential, especially when considering clinical implementation. We report the development of the International Consortium for COVID-19 Imaging AI (ICOVAI) model and perform independent external validation.

Methods

The ICOVAI model was developed using multicenter data (n=1286 CT scans) to quantify disease extent and assess COVID-19 likelihood using the COVID-19 Reporting and Data System (CO-RADS). A ResUNet model was modified to automatically delineate lung contours and infectious lung opacities on CT scans, after which a random forest predicted the CO-RADS score. After internal testing, the model was externally validated on a multicenter dataset (n=400) by independent researchers. CO-RADS classification performance was calculated using linearly weighted Cohen's kappa and segmentation performance using Dice Similarity Coefficient (DSC).

Results

Regarding internal versus external testing, segmentation performance of lung contours was equally excellent (DSC=0.97 vs. DSC=0.97, p=0.97). Lung opacities segmentation performance was adequate internally (DSC=0.76), but significantly worse on external validation (DSC=0.59, p<0.0001). For CO-RADS classification, agreement with

radiologists on the internal set was substantial (kappa=0.78), but significantly lower on the external set (kappa=0.62, p<0.0001).

Conclusion

In this multicenter study, a model developed for CO-RADS score prediction and quantification of COVID-19 disease extent was found to have a significant reduction in performance on independent external validation versus internal testing. The limited reproducibility of the model restricted its potential for clinical use. The study demonstrates the importance of independent external validation of AI models.

Key words

Artificial Intelligence, COVID-19, Tomography, X-Ray Computed, Reproducibility of Results, Validation Study

Key Points

- The ICOVAI model for prediction of CO-RADS and quantification of disease extent on chest CT of COVID-19 patients was developed using a large sample of multicenter data.
- There was substantial performance on internal testing, however, performance was significantly reduced on external validation, performed by independent researchers. The limited generalizability of the model restricts its potential for clinical use.
- Results of AI models for COVID-19 imaging on internal tests may not generalise well to external data, demonstrating the importance of independent external validation.

Introduction

Artificial intelligence (AI)-based analysis of imaging performed for coronavirus disease 2019 (COVID-19) evaluation has been extensively researched [101]. During the pandemic, several deep learning models have been developed, aiming to assist radiologists in interpreting and reporting chest CT scans in COVID-19 patients.

Volume quantification of affected lung tissue on chest CT scans has been shown to correlate with disease severity in COVID-19 [102, 103, 104, 105, 106]. Manual delineation of lung abnormalities by radiologists is labour-intensive and time-consuming, and therefore not routinely conducted in clinical practice. Automated segmentation of affected lung tissue can be made readily available, thereby allowing clinical adoption of quantitative analysis.

To standardise reporting of chest CT scans, the COVID-19 Reporting and Data System (CO-RADS) was introduced [107]. The grading system includes five categories of increasing disease probability, ranging from negative (CO-RADS 1) to typical imaging findings of COVID-19 (CO-RADS 5). CO-RADS has shown reasonable to very good diagnostic performance and interobserver agreement [107, 108, 109, 110]. Applying machine learning techniques to automate CO-RADS classification could potentially improve the interobserver agreement, especially for less experienced readers. Moreover, such an automated analysis can be performed before clinicians have the opportunity to read the CT scan, ensuring the CO-RADS classification and volume quantification are present at the time of interpretation. This could potentially result in a more efficient clinical workflow if the automated assessment is sufficiently accurate.

Before any AI application is considered for widespread clinical use, external validation of the model should be performed [111]. In the systematic review by Roberts et al., only 8 of 37 (22%) deep learning papers on COVID-19 imaging analysis that passed their quality check, had completed external validation [112]. This might especially be worrisome

for AI applications in COVID-19 imaging since several methodological flaws and biases in these studies were reported [112]. The authors stressed the importance of performing an external validation on a well-curated dataset of appropriate size to evaluate the generalizability of an AI model, ensuring it translates well to unseen, independent data.

This study aimed to develop and independently validate an AI model consisting of COVID-19 segmentation and likelihood estimation (CO-RADS) on chest CT using multicenter data.

Material and Methods

International Consortium for COVID-19 Imaging AI (ICOVAI) During the initial phase of the COVID-19 pandemic, there was a need for accurate and efficient analysis of chest CT scans. ICOVAI was formed to address this need. The collaboration consisted of multiple hospitals and industry partners across Europe. The consortium aimed to develop an AI-based quantification and CO-RADS classification tool for clinical use, following good-practice guidelines. These principles included high-quality diverse data and multiple expert readers to perform data annotation.

Data collection

The ICOVAI consortium included a multicenter, international cohort of patients suspected of COVID-19 pneumonia undergoing chest CT. The dataset for model creation consisted of n=1092 CT scans of patients with available reverse transcriptase-polymerase chain reaction (RT-PCR) test results for COVID-19 (n=580 positive, n=512 negative), shown in Figure 5.1. The data was collected between December 2019 and May 2020 through ten participating institutions (Table 5.1). The male (n=545) to female (n=547) ratio was 1:1. To balance the dataset, n=194 CT scans from the National Lung Screening Trial were

added as negative control samples. Combined, the total dataset yielded n=1286 CT scans from n=1266 unique patients.

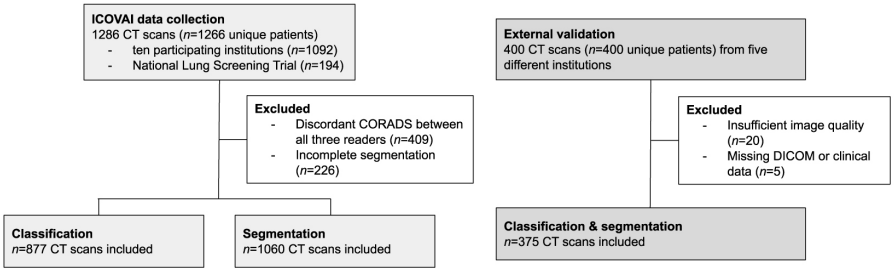

Figure 5.1: Data flowchart for the ICOVAI model development and external validation.

An independent test dataset for external validation was retrospectively collected from five different hospitals in Europe (Table 5.2). The cohort included n=400 adult patients undergoing chest CT for suspected COVID-19 pneumonia or triage between February 2020 and May 2020. Twenty-five patients were excluded due to severe breathing or motion artefacts (n=9), insufficient inspiration (n=9), low resolution (n=2), or missing DICOM data or clinical information (n=5). After exclusion, n=375 CT scans of unique patients remained, with a mean age of 61.1 years (SD 16.8), and male-to-female ratio of 1.1:1. The majority of patients showed symptoms of respiratory infection at the time of imaging (n=332, 88.5%). RT-PCR tests performed within seven days of imaging were used as a reference standard and available for n=363 patients (96.8%). Available RT-PCR test results were positive for n=181 patients and negative for n=182 patients.

Data annotation

ICOVAI model

Multiple radiologists independently classified all CT scans (n=1286) using the CO-RADS scheme (n=1058 by three readers, n=228 by two readers). A total of 409 cases were excluded due to discordance, i.e. all readers yielded different CO-RADS scores, resulting in 877 CT scans. The distribution of classification labels for both the training (n=805) and internal test set (n=72) is shown in Table 5.3. The total lung volume and lung opacities were manually segmented by medical students in n=1060 CT scans and reviewed by a team of n=15 radiologists (2-23 years of experience). For n=905, more than two readers segmented each CT scan, after which both segmentation masks were averaged and rounded. Segmentations were performed using Veye Annotator (Aidence BV).

External validation

The external test dataset (n=400) was classified by two readers using CO-RADS. Each case was read twice; first by a radiology resident (F.G., fourth year of training) or radiologist (L.T., 5 years of experience), and thereafter by a certified thoracic radiologist (A.B., 8 years of experience or R.W., 6 years of experience). In cases of discordance or uncertainty, a consensus reading was performed by a third radiologist (A.B., R.W. or L.T.). The distribution of CO-RADS scores for the external test dataset is shown in Table 5.3. Segmentations of total lung volumes were performed by a technical physician (K.G.L.) and reviewed by a radiologist (L.T.). In addition, manual segmentations of infectious lung opacities were performed by a radiology resident or radiologist (F.G., L.T.), and reviewed by a certified thoracic radiologist (A.B., R.W.). Segmentations were performed using RVAI (Robovision BV).

Data preprocessing

To prepare the pixel data from the DICOM series as input for the AI model, quintic interpolation was performed on all slices, yielding a voxel spacing of 1.25 mm x 0.5 mm x 0.5 mm. Subsequently, voxel values were scaled such that the "lung window", i.e. -1000 HU to 300 HU, corresponded to the range of -1.0 to 1.0, for numeric stability. Axial slices were extracted from the generated volume and scaled to a fixed size of 256 x 256 pixels.

Design of the artificial intelligence system

The AI system was designed to delineate COVID-19 infected areas and yield a CO-RADS score through two separate AI models that function in synchrony. First, a convolutional neural network (CNN) with ResUNet-a architecture [113] takes the CT as input and returns two segmentation masks, labelling every voxel in the CT scan as infectious/non-infectious and lung/no-lung. The ResUNet-a architecture for segmentation contained several adjustments (see Supplements). Subsequently, a tree-based ensemble model was used to predict the CO-RADS score. The input features were constructed based on the segmentation masks of the CNN and the corresponding CT image voxel values. The tree-based ensemble model was constructed through a random forest classifier (RandomForestClassifier, scikit-learn v.0.24.1), with the following settings: n-estimators=300, max-depth=48, min-samples-split=12, max-features=32, and random oversampling with 'no majority' strategy (RandomOverSampler, imblearn v0.8.1). All other parameters were at default.

Statistical analysis

The performance of the AI model's CO-RADS predictions was evaluated through the weighted Cohen's kappa score (Equation 5.1) since it

considers how far the prediction is off.

$$\kappa = 1 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} m_{ij}} \tag{5.1}$$

Equ 5.1: With w the confusion matrix weights (Supplementary Table 5.7 for linear), x the observed confusion matrix values, m the expected confusion matrix values based on chance agreement, and n the number of categories.

We implemented the Dice Similarity Coefficient (DSC) to quantify the overlap between the ground-truth label and the AI segmentation in two ways. First, we calculated the DSC (Equation 5.2) based on the true positives (TP), false positives (FP), and false negatives (FN) on each individual CT scan. Here, we reported the median DSC and its 95% confidence interval (CI). However, since the negative RT-PCR cases in the test set have no segmented volume, the DSC is not defined (dividing by zero). Therefore, the DSC was only calculated on CT scans of patients with a positive RT-PCR. Secondly, to include false-positive segmentations returned by the AI model for RT-PCR negative CT scans, we included the 'micro Dice Similarity Coefficient' (mDSC) as well. Here, the TP, FP, and FN are multiplied by the voxel size (mm^3) of the respective CT scan. The resulting values over the CT scans are summed, and the mDSC is calculated via Equation 5.2. This method yields one value, where larger segmented volumes will have an increased impact on the total score. To analyse the correlation between segmented volumes, we implemented Spearman's correlation. For statistical tests, $p<0.05$ was considered significant. See supplemental material for p-value calculation.

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{5.2}$$

Equ 5.2: The Dice Similarity Coefficient (DSC) equation, where TP, FP, and FN are the numbers of true positive, false positive, and false negative observations, respectively.

Model training & deployment

The resulting dataset was divided into a training (n=971) and an internal test (n=89) set, based on a randomly stratified split. Therefore, the ratios of the different CO-RADS classifications were approximately equal in the two sets. The segmentation model was trained with randomly sampled slices from the training set CT scans. Scaling, rotation, translation, mirroring, and addition of noise were applied to the slices to augment the training data. Stochastic Gradient Descent was used as the optimizer with a learning rate of 0.1 and Nesterov Momentum of 0.9. DSC was implemented as the loss function. The AI model was developed and trained with Tensorflow (v2.3.2). The classification model was trained on 805 CT scans with 10-fold cross-validation. To account for class imbalance, random over-sampling of minority CO-RADS classification scores was performed. To perform external validation, the AI model was deployed within the hospital environment and inference was executed on two NVIDIA Quadro RTX 8000.

Institution	Classification		Segmentation	
	Training	Internal test	Training	Internal test
Albert Schweitzer Hospital, NL	15	1	20	1
AZ Turnhout, BE	82	4	102	7
Catharina Hospital, NL	19	3	26	3
Imapôle Lyon-Villeurbanne, FR	108	17	136	19
Laurentius Hospital, NL	145	9	179	17
Lifetrack, SG	1	0	2	0
NHSX, UK	50	7	59	9
Rijnstate, NL	14	3	22	3
Tergooi MC, NL	13	0	13	0
Franciscus Gasthuis & Vlietland, NL	216	14	252	16

Table 5.1: Dataset of the ICOVAI consortium. Number of CT scans per participating institution for both the classification and segmentation task. The data is split into a training and internal test set for both tasks. NL is the Netherlands, BE is Belgium, FR is France, SG is Singapore, and UK is the United Kingdom.

Institution	External test
Amphia Hospital, NL	56
Antwerp University Hospital, BE	171
Campus Bio-Medico University of Rome, IT	15
University Hospital of Liège, BE	87
OLV Hospital, BE	46
Total	**375**

Table 5.2: Dataset for external validation. Number of CT scans per participating institution. NL is the Netherlands, BE is Belgium, and IT is Italy.

CO-RADS	Training	Internal test	External test
1	362 (45%)	30 (42%)	137 (37%)
2	121 (15%)	12 (17%)	69 (18%)
3	66 (8%)	6 (8%)	48 (13%)
4	60 (7%)	6 (8%)	13 (3%)
5	196 (24%)	18 (25%)	108 (29%)
Total	**805**	**72**	**375**

Table 5.3: CO-RADS COVID-19 Reporting and Data System. Number of CT scans per CO-RADS category in the training, internal test, and external test datasets.

Results

Imaging data

For the ICOVAI dataset, the CT manufacturers were GE (n=424, 33.0%), Siemens (n=499, 38.9%), Philips (n=323, 25.1%), Toshiba (n=37, 2.9%), and unknown (n=3, 0.2%). More detailed acquisition parameters are listed in Supplementary Table 5.8. For the external validation dataset, chest CT scans were acquired without intravenous contrast in 74.1% patients (n=278), and with intravenous contrast in 25.9% patients (n=97). Distribution of CT manufacturers was GE in 55.7% cases (n=209), and Siemens in 44.3% cases (n=166). Slice thickness ranged from 1.0 to 3.0 mm (average 1.5 mm).

Internal test

Inter-reader agreement

To report on inter-reader agreement with respect to classification using CO-RADS, all scans with a score of at least two readers were analysed. This analysis also included scans for which no majority consensus could be found, yielding a total of 1058 CT scans. Between all reader pairs (n=4895 combinations), Cohen's kappa scores were 0.48 (unweighted), 0.72 (linear weighted), and 0.85 (quadratic weighted).

AI performance

The AI model achieved a COVID-19 segmentation DSC of 0.76 and sensitivity of 0.79. The mean true positive, false positive, and false negative volume of COVID were 228.9 mL, 88.3 mL, and 59.1 mL, respectively. The mean absolute error was 117.1 mL. For total lung segmentation, the AI model achieved a DSC of 0.97 and sensitivity of

CO-RADS		Prediction				
		1	**2**	**3**	**4**	**5**
Ground Truth	**1**	29	1	0	0	0
	2	3	5	3	1	0
	3	4	1	1	0	0
	4	0	2	0	1	3
	5	0	0	1	2	15

Table 5.4: Confusion matrix of CO-RADS classification on internal test set.

0.97. The mean true positive, false positive, and false negative volume of COVID were 4433.9 mL, 97.0 mL, and 137.1 mL, respectively. The mean absolute error was 147.9 mL. For CO-RADS classification, the AI model achieved Cohen's kappa scores of 0.58 (not weighted), 0.78 (linearly weighted), and 0.89 (quadratically weighted). The confusion matrix is shown in Table 5.4.

External test

AI performance

The ICOVAI model pipeline excluded n=1 case, leaving n=374 for final analysis. For COVID-19 segmentation, the AI model achieved a performance of 0.59 mDSC and 0.63 sensitivity on the external test dataset, significantly lower than on the internal test set ($p<0.0001$). The mean true positive, false positive, and false negative volumes of COVID were 237 mL, 197 mL, and 138 mL, respectively. The mean absolute error was 142mL (CI: 81– 246 mL). The median DSC over all COVID-19 positive CT scans was 0.48. The distribution of DSC scores is shown in Figure 5.2A. The correlation between the segmented volume by the AI model and the segmentation by the expert reader was strong (spearman r=0.83, $p<0.001$), see Figure 5.2B. The total lung segmentation

achieved 0.97 mDSC and 0.98 sensitivity on the external test dataset. The mean true positive, false positive, and false negative volumes of COVID were 4.1 L, 178 mL, and 80 mL, respectively. The mean absolute error was 148 mL (CI: 135– 156 mL). The median DSC over all COVID-19 positive CT scans was 0.97. Figure 5.2 shows total lung segmentation in two patients with extensive opacities. The CO-RADS classification achieved Cohen's kappa scores of 0.41 (not weighted), 0.62 (linearly weighted), and 0.75 (quadratically weighted). See Table 5.5 for the confusion matrix. Figure 5.3 shows two examples of misclassification.

CO-RADS		**Prediction**				
		1	**2**	**3**	**4**	**5**
Ground Truth	**1**	94	29	9	3	2
	2	17	35	5	8	3
	3	12	9	2	3	22
	4	0	1	4	2	6
	5	4	6	5	15	78

Table 5.5: Confusion matrix of CO-RADS classification on external test set.

Visual interpretation

A radiologist (L.T., 5 years of experience) performed a qualitative visual inspection of the segmentation results on the external test set. The AI delineation of infectious lung opacities was determined adequate to excellent for the majority of cases. When compared to the ground truth labels generated by the radiologists, the ICOVAI model was less sensitive to discrete ground-glass opacities. In several cases, the ICOVAI model generated false-positive segmentations of non-infectious lung opacities such as atelectasis or fibrosis.

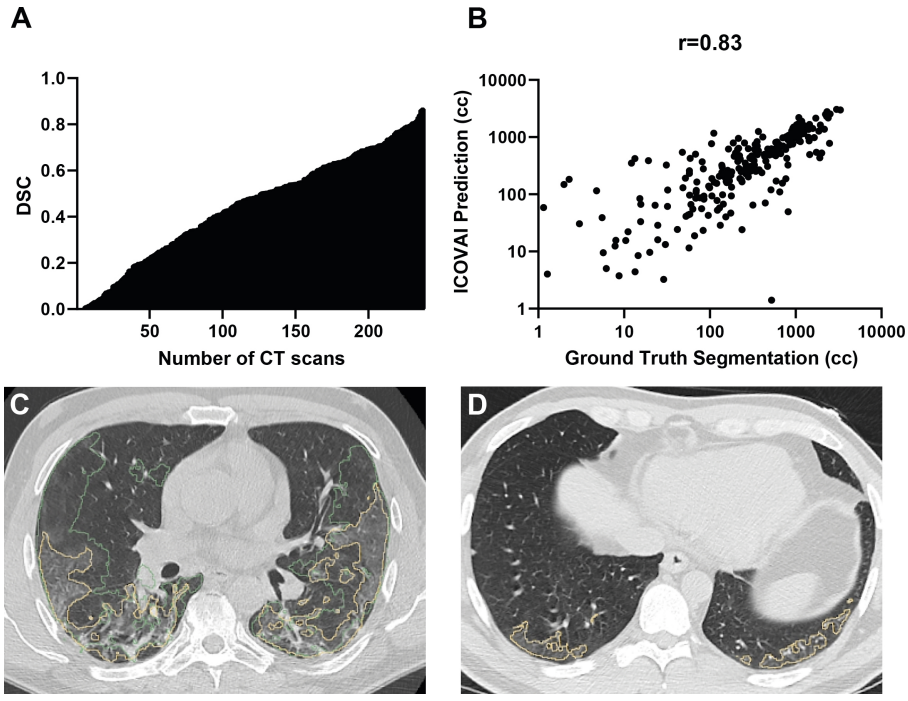

Figure 5.2: Segmentation of infectious lung opacities by the ICOVAI model on external validation. (A) Distribution of DSC in the external test set of patients with RT-PCR confirmed COVID-19. (B) There is a strong correlation between the volume of infectious lung opacities segmented by the experts (ground truth) and the ICOVAI model. (C) Ground truth segmentations (green contours) included a larger area of discrete ground-glass opacity, versus ICOVAI segmentation (yellow contours) which included only marked ground-glass opacities. (D) False-positive segmentation by the ICOVAI model of normal increased attenuation in the posterior lung bases.

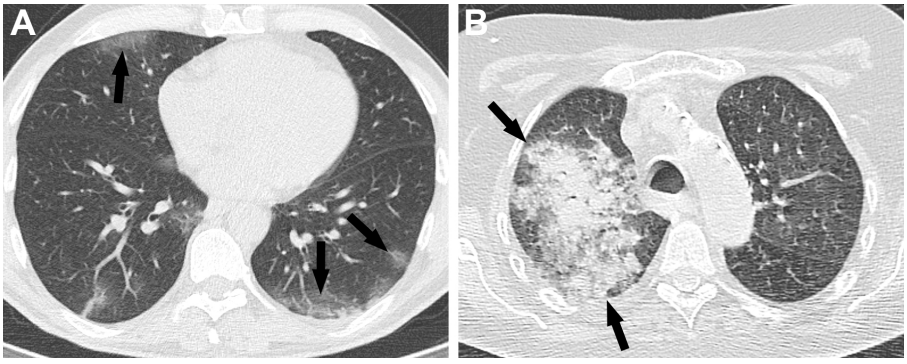

Figure 5.3: CO-RADS misclassification by the ICOVAI model on the external test dataset. (A) A 55-year-old patient with small subpleural ground-glass opacities in both lungs (arrows), consisted with a typical appearance of COVID-19 (CO-RADS 5), later confirmed with RT-PCR. The case was misclassified as negative (CO-RADS 1) by the ICOVAI model. (B) A 70-year-old patient was admitted to the intensive care unit with lobar pneumonia, atypical appearance for COVID-19 (CO-RADS 2). CT showed infectious consolidation in the right upper lobe (arrows), and increased attenuation due to hypoventilation in the other pulmonary lobes. The case was misclassified as CO-RADS 5 by the ICOVAI model.

Discussion

In this multicenter study, we described the development of the ICOVAI model and performed an independent external validation using data from five institutions. We observed a significant reduction in performance on the external test as compared to the internal test for lung opacity segmentation and CO-RADS prediction, but not for lung contour segmentation.

To the best of our knowledge, we have performed the first pre-market external validation study that independently assessed the segmentation performance of a COVID-19 imaging AI solution using large volume multicenter data. Our work shows that published results of COVID-19 segmentation on internal test sets may not generalise well to patient data from other institutions.

External validation can highlight the shortcomings of a predictive model, which were not apparent during internal testing on CT scans sampled out of the same cohort. The importance of external validation is illustrated by the increasing number of high-impact journals requesting it for all predictive models before publication [114]. Moreover, repetitive test set use by slightly different experiments can lead to 'test set overfitting' [115], where the model fits the test data well by chance in one of the experiments. We solved these problems with external validation, where the model is tested once at an external location with an unrelated dataset.

The reported differences in segmentation performance of COVID-19 pneumonia on the internal versus external datasets may partially be explained by interreader variation. The lung areas labelled as abnormal by annotators of the development dataset versus independent annotators of the external dataset may vary because of variations in default window-level settings on the different annotation platforms used to perform the ground truth segmentations, leading to distinct cut-off values to label lung densities (Figure 5.2C).

Variation in CO-RADS scoring between the internal and the external test set could, to some extent, be explained by selection bias. For the internal dataset, CO-RADS scores were excluded when there was no majority consensus between readers, eliminating the 'hardest to evaluate' cases. This is most likely also the cause for the AI model's kappa score being higher than the inter-reader kappa score. When results on the internal test set are better than the ground truth, test set overfitting may be occurring [115]. In this case, external validation can reflect the true, tempered performance of the AI model more accurately.

A prior study by Lessmann et al. trained an AI system with single-centre data to score the likelihood of COVID-19 using CO-RADS [116]. They found a moderate to substantial agreement between observers, reporting a linearly weighted kappa of 0.60 on their internal test set, and 0.69 on their external test set. In our multicenter study, we found a similar level of agreement (kappa values of 0.78 and 0.62, respectively). Previous multicenter studies that included external validation have focused on a binary or ternary classification of COVID-19 versus other types of pneumonia and normal lungs [117, 118, 119, 120, 121, 122]. These studies reported a high to outstanding area under the receiver operating curve (AUC) (0.87-0.98) for identifying COVID-19 on CT. However, the results are difficult to compare with our study that focused on predicting CO-RADS, a more complex multicategorical assessment scheme. Additionally, our external validation was executed by independent researchers. Similarly, Jungmann et al. performed an independent external validation on four commercial AI solutions to differentiate COVID-19 pneumonia from other lung conditions [123]. They found high negative predictive values (82-99%) for the tested models, however, deemed only one solution to have an acceptable sensitivity. The specificity of the four solutions was highly variable (31-80%) and positive predictive values were low (19-25%). Their study was limited to evaluating binary classification and did not assess the segmentation accuracy. Regarding lung opacities segmentation performance on COVID-19 patients, the multicenter study of Zhang et al. reached an mDSC of 0.55-0.58 on internal testing, comparable with our findings on exter-

nal validation [120]. Other published studies have reported higher DSC values for segmentation of lung opacities. However, most studies used single-centre data, datasets of limited size, or did not perform external validation [124, 125, 126, 127].

Our study has several limitations. First, patients were selected by convenience sampling, which may have introduced selection bias. The internal dataset included controls from the National Lung Screening Trial that did not correspond to the target population. This was mitigated by performing an independent validation with a balanced external dataset. Second, CO-RADS is prone to interobserver variability and is therefore an imperfect reference standard. Cases in the internal dataset were excluded when there was a disagreement between all readers on CO-RADS classification, arguably inducing a bias towards less complicated cases. For the external dataset, disagreements were resolved using consensus. Third, interobserver variability of COVID-19 segmentations was not evaluated. Therefore, we cannot determine whether the ICOVAI model was reasonably close to the agreement between radiologists. Future AI developers might benefit from a centralised, high-quality reference image repository to perform external validation of their model, which would also be helpful in setting benchmarks of model performance.

Conclusion

This study evaluated the ICOVAI model performance independently using an external, multicenter test dataset. Segmentation of total lung volumes in both internal and external dataset was excellent, even in patients with severe COVID-19 pneumonia. The performance of the ICOVAI model on segmentation of infectious lung opacities and classification of CO-RADS was significantly worse on the external test dataset compared to the internal test dataset. The results showed limitations in the generalizability of the ICOVAI model, therefore restricting the potential for clinical use. Our study demonstrates the importance of independent external validation of AI models.

Acknowledgements

We would like to thank participating hospitals of the ICOVAI consortium. Moreover, we are thankful to the Imaging COVID19 AI group to provide the dataset to perform the external validation of the ICOVAI model.

The International Consortium for COVID-19 Imaging AI (ICOVAI)
Albert Schweitzer Hospital, Dordrecht, The Netherlands; Department of Radiology, Deventer Hospital, Deventer, the Netherlands; Department of Radiology, Tergooi Hospital, The Netherlands; Amsterdam University Medical Center, Amsterdam, The Netherlands; Julien Guiot, Department of Pneumology, University Hospital of Liège, Liège, Belgium; Annemiek Snoeckx, Antwerp University Hospital, Antwerp, Belgium, and Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium; Peter Kint, Department of Radiology, Amphia Hospital, Breda, The Netherlands; Lieven Van Hoe, Department of Radiology, OLV Hospital, Aalst, Belgium; Carlo Cosimo Quattrocchi, Departmental Faculty of Medicine and Surgery, Diagnostic Imaging and Interventional Radiology, Università Campus Bio-Medico di Roma, Rome, Italy; Dennis Dieckens, Albert Schweitzer Hospital, Dordrecht, The Netherlands; Samir Lounis, Imapole Lyon-Villeurbanne, France; Eric Schulze, Lifetrack Medical Systems, Singapore; Arnout Eric-bart Sjer, Medical Clinic Velsen, The Netherlands; Niels van Vucht, University College London Hospital, United Kingdom; Jeroen A.W. Tielbeek, Department of Radiology, Spaarne Gasthuis Haarlem / Hoofddorp, The Netherlands; Frank Raat, Laurentius Hospital Roermond, The Netherlands; Daniël Eijspaart, Red Cross Hospital, The Netherlands; Ausami Abbas, University Hospital Southampton, United Kingdom.

Declarations

Funding

The study has received funding by the Horizon 2020 framework programme of the European Union under grant agreement no. 961522.

Guarantor

The scientific guarantor of this publication is Kevin Groot Lipman.

Conflict of interest

Laurens Topff, Kevin Groot Lipman, Frederic Guffens, Rianne Wittenberg, Annemarieke Bartels-Rutten, Erik Ranschaert, Stefano Trebeschi, Regina Beets-Tan: no disclosures; **Gerben van Veenendaal, Mirco Hess, Kay Lamerigts, and Joris Wakkie**: employees of Aidence, Amsterdam, The Netherlands; **Jacob J. Visser**: Medical advisor Noaber Foundation, medical advisor NLC, medical advisor Contextflow GmbH, medical advisor Quibim SL.

Statistics and biometry

No complex statistical methods were necessary for this paper.

Informed consent

Written informed consent was not required because of the retrospective nature of this study. Ethical approval Institutional Review Board approval was obtained.

Supplemental Materials

ResU-Net-a architecture changes

Adjustments of ResU-Net-a architecture: 1) The PSP pooling layers were omitted, 2) five stages were implemented instead of six, 3) sixteen filters were used in all five stages instead of exponentially increasing the filters, 4) the dilations per stage in the ResNets were adjusted (Supplementary Table 5.6), 5) Instance Normalization [128] was implemented instead of Batch Normalization, and 6) transpose convolutions were used to upsample.

Calculation of p-values

Significant differences were calculated through bootstrapping, since both the kappa scores and de mDSC are a single value for the entire dataset. The external dataset was bootstrapped 10000 times with replacement, yielding mean and standard deviation of the returned kappa and mDSC scores. Subsequently, the z score from this distribution was calculated by Equation 5.3. The scipy package (v.1.2.3, stats.norm.sf) was used to convert the z-score to the corresponding p-value (two-tailed).

$$z = \frac{x - \mu}{\sigma} \tag{5.3}$$

Equ 5.3: With z the z-score, x the performance measure on the internal test set (kappa/mDSC), μ the mean and σ the standard deviation of the bootstrapped distribution of external test set's performance measures.

Stage	Dilation
1	(1, 9, 17)
2	(1, 5, 9)
3	(1, 3, 5)
4	(1, 2, 3)
5	(1)

Table 5.6: Adjustments to the original Res-Unet architecture. Dilations are the 'field of view' over which the upsampling is performed.

CO-RADS		**ICOVAI model**				
		1	**2**	**3**	**4**	**5**
Radiologists	**1**	0	1	2	3	4
	2	1	0	1	2	3
	3	2	1	0	1	2
	4	3	2	1	0	1
	5	4	3	2	1	0

Table 5.7: The linear weights used to calculate the Cohen's kappa for the CORADS classification task. The larger the difference between the CO-RADS score of the radiologist and the AI prediction, the higher the penalty. AI: Artificial Intelligence; CO-RADS: COVID-19 Reporting and Data System; ICOVAI: International Consortium for COVID-19 Imaging AI

		Classification		Segmentation	
		Train Int.	Test	Train Int.	Test
Slice Thickness	**Min**	0.5	0.625	0.5	0.625
	Median	1.0	1.0	1.0	1.0
	Max	3.2	3.2	5.0	3.2
	Mean	1.2	1.2	1.2	1.2
	<2 mm	657	57	800	74
	>= 2 mm	148	15	171	15
	Total	805	72	971	89
Slice Spacing	**Min**	0.45	0.45	0.45	0.45
	Median	1.0	1.0	1.0	1.0
	Max	3.0	3.0	3.0	3.0
	Mean	1.1	1.1	1.1	1.0
	Total	805	72	971	89
X-ray tube current	**Min**	40	50	40	50
	Median	100	90	100	85
	Max	499	160	499	160
	Mean	135.2	98.6	132.3	90.7
	Total	160	14	179	14
Kilo voltage peak	**Min**	80	80	80	80
	Median	120	120	120	120
	Max	140	140	140	140
	Mean	115.3	116.7	114.1	115.9
	Total	805	72	971	89

Table 5.8: Descriptive statistics of the CT scans for the classification and segmentation tasks.

Part II

Evaluating Therapeutic Response in Pleural Mesothelioma

In preparation for submission to peer review journal

6

ARTIMES: Automated Response evaluation to Treatment In Mesothelioma based on Artificial Intelligence

Kevin B.W. Groot Lipman, Rianne Wittenberg, Mateus de Oliveira Taveira, Illaa Smesseim, Alexander Schmitz, Thierry N. Boellaard, Kalina Chupetlovska, Mohamed A. Abdelatty, Liliana Petrychenko, Neeraj Jain, Francesco Arico, Valerio Pugliese, Caroline Zellweger, Alessandra Curioni, Thomas Frauenfelder, Thi Dan Linh Nguyen-Kim, Renaud Tissier, Regina G.H. Beets-Tan, Jacobus A. Burgers, Cornedine J. de Gooijer*, Stefano Trebeschi*, *Shared Last Author

7

General Discussion

Optimizing AI Models for Clinical Application: Insights and Lessons from Various Strategies

Fundamental versus applied AI research in medical imaging

When developing AI models for medical imaging, the approaches can be broadly divided into two categories: fundamental/technical development and applied/clinical development. Neither category surpasses the other in importance; rather, they serve different end goals. This doctoral thesis is focused on applied/clinical models: identifying the most appropriate AI solution for the given clinical problem, applying it to a representative and curated dataset, and investigating its utilization, results, and implications for the clinics [165]. In other words, the objective of this thesis is to study and optimize the orchestration of the components that will enable AI-powered tools to have a meaningful impact on the methods and guidelines of the future in the field of respiratory diseases.

Due to the complexity of the components involved [166], the resulting challenges are difficult to anticipate. For example, in the asbestosis studies, presented in *Chapters 2 and 3*, the AI model was trained end-to-end using the experts' panel's verdict - a noisy output label, when considering the high disagreement between experts. This approach, although coherent and in line with the current deep learning literature [167], revealed less so from an applied clinical perspective: while it was possible to demonstrate that the panel's variability could be replicated, by establishing a relationship between model uncertainty and panel disagreement, the model did not enhance the current clinical procedure, possibly due to the lack of reproducibility of the current criteria [14].

Reflecting on the approach, it might have been more effective to develop new criteria, that would critically analyze current ones, and use AI models to standardize their application [168]. For instance, creating an accurate lung segmentation model with a separate module for fibrosis segmentation to accurately determine the percentage of fibrosis

present in the lungs. Such a solution would enable pulmonologists, after having confirmed the accuracy of the segmentation, to use the percentage of lung volume affected by fibrosis, thereby increasing agreement on the 5% fibrosis rule set for financial compensation [14]. Furthermore, the precise quantification of lung parenchyma and fibrosis could have helped in studying the disease further, possibly establishing new, more informed cut-offs. With the current end-to-end method, the AI model seems to be learning to reproduce the inconsistent and somewhat imprecise assessment of the panel, rather than adding to the clinical knowledge.

Another approach, although invasive for participants in the training set, could have been the count of asbestos bodies from multiple biopsies, a more objective assessment of the disease [169]. By training on this endpoint, an AI model could potentially enhance the eligibility classification based on a more biologically accurate label, possibly surpassing the accuracy of the panel's verdict. In the absence of this, it may have been more beneficial to visualize the fibrosis and lung segmentation, enabling pulmonologists to quickly verify and trust the calculated percentage score. In conclusion, while the chosen approach was in line with the current technical and deep learning literature, it was less effective from a clinical standpoint.

External validations

The current literature often expresses the merits of external validations [32, 170], and while the author aligns with this perspective, external validations should also be approached with care. Simply utilizing any external dataset available for validation of AI performance may prove erroneous, particularly if the AI model is assessed outside its intended use. For example, in the validation of the asbestosis model, an external validation on a general external population would likely lead the model to misclassify CT scans displaying substantial fibrosis, such as silicosis or any other interstitial lung disease, as 'asbestosis'. It is imperative

that the model is trained against a similar population, respecting all the assumptions and requirements formulated during training. Given the absence of a comparable cohort and the model's training to differentiate eligible Dutch patients from non-eligible ones, we opted for prospective validation over external validation for the asbestosis model.

An exemplary case of sub-optimal external validation is shown in *Chapter 5* of this thesis. Here, the AI segmentation model's differentiation threshold was based on the Hounsfield units (HU) representing COVID-affected lung tissue. A challenge arises because the HU scale in CT cannot be visualized in its entirety, as it would exceed the spectrum of shades of grey perceptible by the human eye, thus necessitating windowing [171]. Windowing restricts the focus on a sub-section of the HU spectrum, making it easier to visualize small differences in density [171]. As a consequence, there is a non-linear relationship between what we observe, visually, to the actual HU [172]. In the context of a CT thorax slice, without knowledge of the window width and center, extracting the HU based on visual interpretation of the grey-level appearance alone results in erroneous associations.

Discrepancy may arise when radiologists involved in the internal validation employ e.g. different window levels than the independent radiologists who segment the external validation set. The AI segmentation model could demonstrate remarkable performance in the internal test set and appear visually satisfactory in the external validation set, yet achieve minimal overlap with the radiologists' segmentation in that external validation, as we observed in Chapter 5. Complications may also emerge on a more technical level, for example when datasets are converted to JPEG/PNG formats, where the window's minimum and maximum values are set, eliminating potentially important values outside the set window [172]. For this reason, popular frameworks like nnUNet [94] have adopted a strategy of clipping the HU based on HU percentiles in the segmented dataset. This illustrates the need for a nuanced approach to external validations, awareness of the characteristics of each dataset, and the inherent limitations of certain labeling (and/or clinical) methodologies.

Reconstruction kernels

In the context of the projects in this dissertation, the reconstruction kernel in thoracic imaging was shown to be a significant parameter. CT thorax scans usually undergo both lung reconstruction and soft reconstruction, utilizing the same acquired attenuation data from the CT scanner [173]. Despite imaging identical anatomical structures, the lung reconstructed CT scan offers greater resolution or detail compared to the soft reconstruction, at the cost of increased noise in the image [174]. Radiologists leverage this difference, examining the lung reconstruction for finer details such as small nodules or interstitial lung diseases, while the soft reconstruction is utilized for analyzing lymph nodes and other soft tissue components [175].

This dissertation illustrates that the selection between lung and soft reconstructed CT scans requires careful consideration from researchers working on respiratory diseases. In the projects related to asbestosis and pleural plaques detailed in *Chapters 2, 3 and 4*, a blend of both reconstructions was used, aiming to enhance the generalizability of the AI model through training on both types [176]. However, upon the observations made throughout different studies, it would potentially have been better to utilize the lung reconstruction specifically for the asbestosis model and the soft reconstruction for the pleural plaque segmentation project.

This choice would have likely not resulted in a lower inclusion rate, as CT thorax scans are typically subject to both lung and soft reconstruction in their PACS (Picture Archiving and Communication System). The experience underscores the need for selecting the correct reconstruction types for a specific anatomical structure of interest, acknowledging the properties and applications of lung and soft reconstructions [175]. By aligning the choice of reconstruction with the specific requirements of each project, the efficacy and accuracy of the resulting models can be optimized [176, 177].

Psychological Factors

A final, yet often under-emphasized aspect of the projects detailed in this thesis concerns the psychological factors at play in AI model development [178]. Acknowledging and nurturing the morale of annotators is crucial, as it can influence the accuracy of segmentations, given that the human element remains a part of the process [178, 179]. We empirically observed in chapter 4 that the same radiologist, regardless of their years of experience, is more likely to produce lower quality segmentations when working on a noisy 450-slice, 1 mm, lung reconstruction filter, as opposed to a smoother 150-slice, 3 mm soft reconstruction in a fraction of the time.

Another approach is to challenge the radiologists in the segmentation procedure. Radiologists are engaged in an exercise where they are tasked with identifying errors in the AI-proposed segmentation and making corrections. This method, although introducing a certain bias associated with modifying existing segmentation [180], offers efficiency and consistency across all segmented cases. The knowledge that the AI model is being retrained based on their segmentations creates an additional incentive for radiologists to deliver high-quality adjustments. This, in turn, contributes to a vicious cycle where successive iterations of the AI model continually reduce the annotators' workload. The radiologists are able to create better quality segmentations in less and less time, translating to a faster reward loop mechanism [181].

The final consideration relates to the optimization of workflow. Annotators should be focused solely on their areas of expertise: interpreting images and translating their expert opinions into labels [178]. Thus, the responsibility falls onto researchers to create a streamlined workflow where annotators can seamlessly move from one case to the next with minimal administrative burden. Concerns such as saving files with the correct extension or ensuring the dimensions of segmentation and CT align should be automated by the researcher and feel seamless to the radiologist.

In conclusion, the success of these projects is rooted in a multidisciplinary approach where each team member's expertise is leveraged. Whether it's the nuanced understanding of image reconstruction or the human-centered considerations of annotator engagement and workflow design, each aspect plays a vital role in the overall performance of the AI models developed [178, 181].

Determinants of Success in medical AI (Segmentation) Projects

The *Lessons Learned* stem primarily from various projects in Part 1 of this thesis, which were applied in the ARTIMES project of Part 2. By not opting for an end-to-end approach in predicting pleural mesothelioma progression, and focusing instead on segmenting tumor volume, our research yielded valuable AI output that gained trust and verification from clinicians. The implementation of a blinded study in external validation provided a nuanced comparison, not merely considering the manual segmentation as presumed ground truth, but assessing radiologists' preferences regarding clinical utility. Standardized segmentation guidelines for target structures, windowing, and threshold masking were applied to ensure consistency. A detailed scan selection criterion focused on soft reconstructions with a 3 mm slice increment and thickness, and active learning was utilized to identify uncertain CT scan segmentations, thereby promoting model convergence with fewer examples. Knowing the limitations of the current clinical standards in response evaluation [143, 182, 183], and in-depth knowledge of the clinical presentation of the disease allowed us to craft new response evaluation criteria. Additionally, the project benefited from optimized segmentation pipelines with custom interfaces in 3D Slicer [49, 160], easing radiologists' interactions with essential tools.

Current and Future Perspectives

Similar approaches will also be investigated to expand the study of asbestosis: quantifying fibrosis, and further investigating the role of pleural plaques. The ARTIMES project's success is evidenced by its incorporation into the radiology department workflow of the Netherlands Cancer Institute. An automated process ensures that CT scans of patients with known pleural mesothelioma are forwarded to the cloud, where the ARTIMES AI model segments the CT scan. The resulting segmentations are accessible for inspection in a web-based viewer. Current deployment at the first external site signals the model's broader applicability, and preparations are underway to assess the ARTIMES model and criteria as exploratory endpoints in upcoming clinical trials. Pending the approval of MDR for in-house developed models, the AI-segmented tumor volumes may soon become available for clinical decision-making, enhancing the role of AI in medical practice.

Bibliography

[1] Nicholas D Weatherley et al. “Experimental and quantitative imaging techniques in interstitial lung disease”. en. In: *Thorax* 74.6 (June 2019), pp. 611–619.

[2] Mario Silva et al. “Pulmonary quantitative CT imaging in focal and diffuse disease: current research and clinical applications”. en. In: *Br. J. Radiol.* 91.1083 (Feb. 2018), p. 20170644.

[3] Melissa Rosado de Christensen et al. *Chest Imaging*. en. Oxford University Press, Aug. 2019, pp. 155–157.

[4] Bhavin G Jankharia and Bhoomi A Angirish. “Computer-Aided quantitative analysis in interstitial lung diseases - A pictorial review using CALIPER”. en. In: *Lung India* 38.2 (2021), pp. 161–167.

[5] Alicia Chen et al. “Quantitative CT Analysis of Diffuse Lung Disease”. en. In: *Radiographics* 40.1 (2020), pp. 28–43.

[6] Sharyn I Katz et al. “Considerations for Imaging of Malignant Pleural Mesothelioma: A Consensus Statement from the International Mesothelioma Interest Group”. en. In: *J. Thorac. Oncol.* 18.3 (Mar. 2023), pp. 278–298.

[7] L Daniel Maxim, Ronald Niebo, and Mark J Utell. “Are pleural plaques an appropriate endpoint for risk analyses?” en. In: *Inhal. Toxicol.* 27.7 (June 2015), pp. 321–334.

[8] Alampady Krishna Prasad Shanbhogue, Anand B Karnad, and Srinivasa R Prasad. “Tumor Response Evaluation in Oncology: Current Update”. In: *J. Comput. Assist. Tomogr.* 34.4 (July 2010), p. 479.

[9] *WHO Coronavirus (COVID-19) dashboard*. en. `https://covid19.who.int/`. Accessed: 2023-11-17.

[10] Alasdair Taylor and Craig Williams. “COVID-19: Impact on radiology departments and implications for future service design, service delivery, and radiology education”. en. In: *Br. J. Radiol.* 94.1127 (Nov. 2021), p. 20210632.

[11] Bogdan A Bercean et al. “Evidence of a cognitive bias in the quantification of COVID-19 with CT: an artificial intelligence randomised clinical trial”. en. In: *Sci. Rep.* 13.1 (Mar. 2023), p. 4887.

[12] Di Dong et al. “The Role of Imaging in the Detection and Management of COVID-19: A Review”. en. In: *IEEE Rev. Biomed. Eng.* 14 (Jan. 2021), pp. 16–29.

[13] Meng Yang et al. “Increasing incidence of asbestosis worldwide, 1990-2017: results from the Global Burden of Disease study 2017”. en. In: *Thorax* 75.9 (Sept. 2020), pp. 798–800.

[14] W Hagmolen Of Ten Have, J M Rooijackers, and J A Burgers. “[Financial compensation for asbestosis patients]”. nl. In: *Ned. Tijdschr. Geneeskd.* 160 (2016), p. D544.

[15] Job P van Kooten et al. “Incidence, treatment and survival of malignant pleural and peritoneal mesothelioma: a population-based study”. en. In: *Thorax* 77.12 (Dec. 2022), pp. 1260–1267.

[16] Samuel G Armato et al. “Imaging in pleural mesothelioma: A review of the 15th International Conference of the International Mesothelioma Interest Group”. en. In: *Lung Cancer* 164 (Feb. 2022), pp. 76–83.

[17] Laura E Kerper et al. “Systematic review of pleural plaques and lung function”. en. In: *Inhal. Toxicol.* 27.1 (Jan. 2015), pp. 15–44.

[18] Leonid Kopylev et al. “A systematic review of the association between pleural plaques and changes in lung function”. en. In: *Occup. Environ. Med.* 72.8 (Aug. 2015), pp. 606–614.

[19] Kwang Min Lee et al. “Comparison of Asbestos Victim Relief Available Outside of Conventional Occupational Compensation Schemes”. en. In: *Int. J. Environ. Res. Public Health* 18.10 (May 2021).

[20] Ilyes Benlala et al. “Deep Learning for the Automatic Quantification of Pleural Plaques in Asbestos-Exposed Subjects”. en. In: *Int. J. Environ. Res. Public Health* 19.3 (Jan. 2022).

[21] Junjie Huang et al. “Global Incidence, Risk Factors, and Temporal Trends of Mesothelioma: A Population-Based Study”. en. In: *J. Thorac. Oncol.* 18.6 (June 2023), pp. 792–802.

[22] J. P. van Kooten et al. “Incidence, treatment and survival of malignant pleural and peritoneal mesothelioma: a population-based study”. In: *Thorax* 77.12 (2022), pp. 1260–1267. DOI: 10.1136/thoraxjnl-2021-217709.

[23] Patrick Bou-Samra et al. “Epidemiological, therapeutic, and survival trends in malignant pleural mesothelioma: A review of the National Cancer Database”. In: *Cancer Medicine* 12.11 (2023), pp. 12208–12220. DOI: https://doi.org/10.1002/cam4.5915.

[24] Samuel G Armato 3rd et al. “Observer variability in mesothelioma tumor thickness measurements: defining minimally measurable lesions”. en. In: *J. Thorac. Oncol.* 9.8 (Aug. 2014), pp. 1187–1194.

[25] E A Eisenhauer et al. “New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)”. en. In: *Eur. J. Cancer* 45.2 (Jan. 2009), pp. 228–247.

[26] C. Bech and J. B. Sørensen. “Chemotherapy induced pathologic complete response in malignant pleural mesothelioma: a review and case report”. In: *J Thorac Oncol* 5.5 (2010), pp. 735–40. DOI: 10.1097/jto.0b013e3181d86ea9.

[27] M J Byrne and A K Nowak. “Modified RECIST criteria for assessment of response in malignant pleural mesothelioma”. en. In: *Ann. Oncol.* 15.2 (Feb. 2004), pp. 257–260.

[28] Li-Anne H. Douma et al. “Pembrolizumab plus lenvatinib in second-line and third-line patients with pleural mesothelioma (PEMMELA): a single-arm phase 2 study”. In: *The Lancet Oncology* 24.11 (2023), pp. 1219–1228. DOI: 10.1016/S1470-2045(23)00446-1.

[29] Keiron O’Shea and Ryan Nash. “An Introduction to Convolutional Neural Networks”. In: (Nov. 2015).

[30] Brendan S Kelly et al. “Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE)”. In: *Eur. Radiol.* 32.11 (Nov. 2022), pp. 7998–8007.

[31] Luciano M Prevedello et al. “Challenges Related to Artificial Intelligence Research in Medical Imaging and the Importance of Image Analysis Competitions”. en. In: *Radiol Artif Intell* 1.1 (Jan. 2019), e180031.

[32] Federico Cabitza et al. “The importance of being external. methodological insights for the external validation of machine learning models in medicine”. en. In: *Comput. Methods Programs Biomed.* 208 (Sept. 2021), p. 106288.

[33] David W Kamp. “Asbestos-induced lung diseases: an update”. en. In: *Transl. Res.* 153.4 (Apr. 2009), pp. 143–152.

[34] European Union. *Commission Directive 1999/77/EC*. Tech. rep. Official Journal of the European Communities, July 1999.

[35] Laurent Greillier and Philippe Astoul. “Mesothelioma and asbestos-related pleural diseases”. en. In: *Respiration* 76.1 (May 2008), pp. 1–15.

[36] M Vujović. “[Standardization of diagnostic criteria for occupational asbestosis of the lungs and lung parenchyma]”. hr. In: *Arh. Hig. Rada Toksikol.* 46.4 (Dec. 1995), pp. 445–449.

[37] Henrik Wolff et al. “Asbestos, asbestosis, and cancer, the Helsinki criteria for diagnosis and attribution 2014: recommendations”. en. In: *Scand. J. Work Environ. Health* 41.1 (Jan. 2015), pp. 5–15.

[38] Philip J Landrigan and Collegium Ramazzini. “Comments on the 2014 Helsinki Consensus Report on Asbestos”. en. In: *Ann Glob Health* 82.1 (Jan. 2016), pp. 217–220.

[39] Xaver Baur et al. “Asbestos, asbestosis, and cancer: The Helsinki criteria for diagnosis and attribution. Critical need for revision of the 2014 update”. en. In: *Am. J. Ind. Med.* 60.5 (May 2017), pp. 411–421.

[40] Michael K Felten et al. “Retrospective exposure assessment to airborne asbestos among power industry workers”. en. In: *J. Occup. Med. Toxicol.* 5 (June 2010), p. 15.

[41] Welzijn en Sport Ministerie van Volksgezondheid. *Protocollen asbestziekten: asbestose*. `https://www.gezondheidsraad.nl/documenten/adviezen/1999/03/29/protocollen-asbestziekten-asbestose`. Accessed: 2021-8-31. Mar. 1999.

[42] E Merler and S Brizzi. “Compensation of occupational diseases and particularly of asbestos-related diseases among the European Community (EEC) countries”. en. In: *Epidemiol. Prev.* 18.60 (Sept. 1994), pp. 170–179.

[43] Safe Work Australia. *Comparison of Workers¿ Compensation Arrangements in Australia and New Zealand*. Australian Government-Safe Work Australia, 2012.

[44] Robert D Rondinelli et al. *AMA Guides to the Evaluation of Permanent Impairment, 6th Edition*. 2008.

[45] Rodney LaLonde and Ulas Bagci. “Capsules for Object Segmentation”. In: (Apr. 2018).

[46] Diederik P Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: (Dec. 2013).

[47] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[48] Joshua Broder. *Diagnostic Imaging for the Emergency Physician E-Book*. en. Elsevier Health Sciences, June 2011.

[49] Andriy Fedorov et al. “3D Slicer as an image computing platform for the Quantitative Imaging Network”. en. In: *Magn. Reson. Imaging* 30.9 (Nov. 2012), pp. 1323–1341.

[50] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. en. SAGE, 2004.

[51] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: (Dec. 2013).

[52] Anita Rácz, Dávid Bajusz, and Károly Héberger. “Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification”. en. In: *Molecules* 26.4 (Feb. 2021).

[53] Ahmed Hosny et al. “Artificial intelligence in radiology”. en. In: *Nat. Rev. Cancer* 18.8 (Aug. 2018), pp. 500–510.

[54] Cristiano Rabelo Nogueira et al. “Lung diffusing capacity relates better to short-term progression on HRCT abnormalities than spirometry in mild asbestosis”. en. In: *Am. J. Ind. Med.* 54.3 (Mar. 2011), pp. 185–193.

[55] Davood Karimi et al. “Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis”. en. In: *Med. Image Anal.* 65 (Oct. 2020), p. 101759.

[56] Simon L F Walsh et al. “Multicentre evaluation of multidisciplinary team meeting agreement on diagnosis in diffuse parenchymal lung disease: a case-cohort study”. en. In: *Lancet Respir Med* 4.7 (July 2016), pp. 557–565.

[57] Trishan Panch, Heather Mattie, and Rifat Atun. “Artificial intelligence and algorithmic bias: implications for health systems”. en. In: *J. Glob. Health* 9.2 (Dec. 2019), p. 010318.

[58] Gustavo Saposnik et al. “Cognitive biases associated with medical decisions: a systematic review”. en. In: *BMC Med. Inform. Decis. Mak.* 16.1 (Nov. 2016), p. 138.

[59] Ana Adriana Trusculescu et al. “Deep learning in interstitial lung disease-how long until daily practice”. en. In: *Eur. Radiol.* 30.11 (Nov. 2020), pp. 6285–6292.

[60] Masanori Akira et al. “High-resolution CT of asbestosis and idiopathic pulmonary fibrosis”. en. In: *AJR Am. J. Roentgenol.* 181.1 (July 2003), pp. 163–169.

[61] Myura Nagendran et al. “Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies”. en. In: *BMJ* 368 (Mar. 2020), p. m689.

[62] Kicky G van Leeuwen et al. “Artificial intelligence in radiology: 100 commercially available products and their scientific evidence”. en. In: *Eur. Radiol.* (Apr. 2021).

[63] Marzyeh Ghassemi et al. “Practical guidance on artificial intelligence for health-care data”. en. In: *Lancet Digit Health* 1.4 (Aug. 2019), e157–e159.

[64] Samuel G Armato 3rd et al. “The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans”. en. In: *Med. Phys.* 38.2 (Feb. 2011), pp. 915–931.

[65] Andrew L. Maas. “Rectifier Nonlinearities Improve Neural Network Acoustic Models”. In: 2013.

[66] Wenzhe Shi et al. “Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network”. In: (Sept. 2016).

[67] Zhengyang Wang, Hao Yuan, and Shuiwang Ji. “Spatial Variational Auto-Encoding via Matrix-Variate Normal Distributions”. In: *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*. Proceedings. Society for Industrial and Applied Mathematics, May 2019, pp. 648–656.

[68] Angeline A Lazarus and Andrew Philip. *Asbestosis*. 2011.

[69] Victor L Roggli et al. *Pathology of Asbestosis—An Update of the Diagnostic Criteria: Report of the Asbestosis Committee of the College of American Pathologists and Pulmonary Pathology Society*. 2010.

[70] Harri Vainio et al. "Helsinki Criteria update 2014: asbestos continues to be a challenge for disease prevention and attribution". en. In: *Epidemiol. Prev.* 40.1 Suppl 1 (2016), pp. 15–19.

[71] Kevin B W Groot Lipman et al. "Artificial intelligence-based diagnosis of asbestosis: analysis of a database with applicants for asbestosis state aid". en. In: *Eur. Radiol.* (Dec. 2022).

[72] Dianne de Gooijer. *Prospective validation of the diagnostic accuracy of an automated asbestosis assessment*. `https://trialsearch.who.int/Trial2.aspx?TrialID=NL9064`. Accessed: 2023-4-3. Nov. 2020.

[73] Patrick M Bossuyt et al. "Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative". en. In: *BMJ* 326.7379 (Jan. 2003), pp. 41–44.

[74] Patrick M Bossuyt et al. "STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies". en. In: *BMJ* 351 (Oct. 2015), h5527.

[75] J A Hanley and B J McNeil. "A method of comparing the areas under receiver operating characteristic curves derived from the same cases". en. In: *Radiology* 148.3 (Sept. 1983), pp. 839–843.

[76] Justus J Randolph. *Free-Marginal Multirater Kappa (multirater K[free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa*. en. ERIC Clearinghouse, 2005.

[77] Julius Adebayo et al. "Sanity checks for saliency maps". In: *Adv. Neural Inf. Process. Syst.* 31 (2018).

[78] Elke Ochsmann et al. “Inter-reader variability in chest radiography and HRCT for the early detection of asbestos-related lung and pleural abnormalities in a cohort of 636 asbestos-exposed subjects”. en. In: *Int. Arch. Occup. Environ. Health* 83.1 (Jan. 2010), pp. 39–46.

[79] Jonas Widell and Mats Lidén. “Interobserver variability in high-resolution CT of the lungs”. en. In: *Eur J Radiol Open* 7 (Mar. 2020), p. 100228.

[80] Christopher J Kelly et al. “Key challenges for delivering clinical impact with artificial intelligence”. en. In: *BMC Med.* 17.1 (Oct. 2019), p. 195.

[81] John Eng. “Sample size estimation: a glimpse beyond simple formulas”. en. In: *Radiology* 230.3 (Mar. 2004), pp. 606–612.

[82] M Benchoufi et al. “Interobserver agreement issues in radiology”. en. In: *Diagn. Interv. Imaging* 101.10 (Oct. 2020), pp. 639–641.

[83] Görkem Algan and İlkay Ulusoy. “Label Noise Types and Their Effects on Deep Learning”. In: (Mar. 2020).

[84] C Peacock, S J Copley, and D M Hansell. “Asbestos-related benign pleural disease”. en. In: *Clin. Radiol.* 55.6 (June 2000), pp. 422–432.

[85] Huw D Roach et al. “Asbestos: when the dust settles an imaging review of asbestos-related disease”. en. In: *Radiographics* 22 Spec No (Oct. 2002), S167–84.

[86] Gang Liu, Paul Cheresh, and David W Kamp. “Molecular basis of asbestos-induced lung disease”. en. In: *Annu. Rev. Pathol.* 8 (Jan. 2013), pp. 161–187.

[87] C Paris et al. “Pleural plaques and asbestosis: dose–and time–response relationships based on HRCT data”. In: *Eur. Respir. J.* 34.1 (2009), pp. 72–79.

[88] Renelle Myers. “Asbestos-related pleural disease”. en. In: *Curr. Opin. Pulm. Med.* 18.4 (July 2012), pp. 377–381.

[89] Gael Dournes et al. “3-Dimensional Quantification of Composite Pleural Plaque Volume in Patients Exposed to Asbestos Using High-resolution Computed Tomography: A Validation Study”. In: *J. Thorac. Imaging* 34.5 (Sept. 2019), p. 320.

[90] Yoon Ki Cha, Jeung Sook Kim, and Jae Hyun Kwon. “Quantification of pleural plaques by computed tomography and correlations with pulmonary function: preliminary study”. en. In: *J. Thorac. Dis.* 10.4 (Apr. 2018), pp. 2118–2124.

[91] Philip H Quanjer et al. “Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations”. en. In: *Eur. Respir. J.* 40.6 (Dec. 2012), pp. 1324–1343.

[92] Brian L Graham et al. “2017 ERS/ATS standards for single-breath carbon monoxide uptake in the lung”. en. In: *Eur. Respir. J.* 49.1 (Jan. 2017), p. 1600016.

[93] Marleen Groenier et al. “Evaluation of the impact of technical physicians on improving individual patient care with technology”. en. In: *BMC Med. Educ.* 23.1 (Mar. 2023), p. 181.

[94] Fabian Isensee et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. en. In: *Nat. Methods* 18.2 (Feb. 2021), pp. 203–211.

[95] Laurent Plantier et al. “Physiology of the lung in idiopathic pulmonary fibrosis”. en. In: *Eur. Respir. Rev.* 27.147 (Mar. 2018).

[96] Ibrahim Güven Çoşğun, Fatma Evyapan, and Nevzat Karabulut. “Environmental asbestos disease: pleural plaque volume measurement with Chest Tomography is there a correlation between pulmonary function?” en. In: *Sarcoidosis Vasc. Diffuse Lung Dis.* 34.4 (Apr. 2017), pp. 336–342.

[97] Marlies van Dijk et al. “The effects of lung volume reduction treatment on diffusing capacity and gas exchange”. en. In: *Eur. Respir. Rev.* 29.158 (Dec. 2020).

[98] Law Lords Department. *House of lords - Johnston (original appellant and cross-respondent) v. NEI international combustion limited (original respondents and cross-appellants) Rothwell (original appellant and cross-respondent) v. Chemical and insulating company limited and others (original respondents and cross-appellants) etc.* `https://publications.parliament.uk/pa/ld200607/ldjudgmt/jd071017/johns-1.htm`. Accessed: 2021-8-6. Oct. 2007.

[99] Jean-Claude Pairon et al. "Pleural plaques and the risk of pleural mesothelioma". en. In: *J. Natl. Cancer Inst.* 105.4 (Feb. 2013), pp. 293–301.

[100] Sam M Janes, Doraid Alrifai, and Dean A Fennell. "Perspectives on the Treatment of Malignant Pleural Mesothelioma". en. In: *N. Engl. J. Med.* 385.13 (Sept. 2021), pp. 1207–1218.

[101] Feng Shi et al. "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19". en. In: *IEEE Rev. Biomed. Eng.* 14 (Jan. 2021), pp. 4–15.

[102] Marco Francone et al. "Chest CT score in COVID-19 patients: correlation with disease severity and short-term prognosis". en. In: *Eur. Radiol.* 30.12 (Dec. 2020), pp. 6808–6817.

[103] Ran Yang et al. "Chest CT Severity Score: An Imaging Tool for Assessing Severe COVID-19". en. In: *Radiol Cardiothorac Imaging* 2.2 (Apr. 2020), e200047.

[104] Xiaofeng Wang et al. "Multicenter Study of Temporal Changes and Prognostic Value of a CT Visual Severity Score in Hospitalized Patients With Coronavirus Disease (COVID-19)". en. In: *AJR Am. J. Roentgenol.* 217.1 (July 2021), pp. 83–92.

[105] Ezio Lanza et al. "Quantitative chest CT analysis in COVID-19 to predict the need for oxygenation support and intubation". en. In: *Eur. Radiol.* 30.12 (Dec. 2020), pp. 6770–6778.

[106] Kajetan Grodecki et al. “Quantitative Burden of COVID-19 Pneumonia at Chest CT Predicts Adverse Outcomes: A Post Hoc Analysis of a Prospective International Registry”. In: *Radiology: Cardiothoracic Imaging* 2.5 (Oct. 2020), e200389.

[107] Mathias Prokop et al. “CO-RADS: A Categorical CT Assessment Scheme for Patients Suspected of Having COVID-19—Definition and Evaluation”. In: *Radiology* 296.2 (Aug. 2020), E97–E104.

[108] Arthur W E Lieveld et al. “Chest CT in COVID-19 at the ED: Validation of the COVID-19 Reporting and Data System (CO-RADS) and CT Severity Score: A Prospective, Multicenter, Observational Study”. en. In: *Chest* 159.3 (Mar. 2021), pp. 1126–1135.

[109] Mohamed Abdel-Tawab et al. “Comparison of the CO-RADS and the RSNA chest CT classification system concerning sensitivity and reliability for the diagnosis of COVID-19 pneumonia”. en. In: *Insights Imaging* 12.1 (Apr. 2021), p. 55.

[110] Shohei Inui et al. “Comparison of Chest CT Grading Systems in COVID-19 Pneumonia”. In: *Radiology: Cardiothoracic Imaging* 2.6 (Dec. 2020), e200492.

[111] Chintan Shah et al. “A translational clinical assessment workflow for the validation of external artificial intelligence models”. en. In: *Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications*. Vol. 11601. SPIE, Feb. 2021, pp. 92–102.

[112] Michael Roberts et al. “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans”. en. In: *Nature Machine Intelligence* 3.3 (Mar. 2021), pp. 199–217.

[113] Foivos I Diakogiannis et al. “ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data”. In: (Apr. 2019).

[114] Chava L Ramspek et al. “External validation of prognostic models: what, why, how, when and where?” en. In: *Clin. Kidney J.* 14.1 (Jan. 2021), pp. 49–58.

[115] Vitaly Feldman, Roy Frostig, and Moritz Hardt. “The advantages of multiple classes for reducing overfitting from test set reuse”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1892–1900.

[116] Nikolas Lessmann et al. “Automated Assessment of COVID-19 Reporting and Data System and Chest CT Severity Scores in Patients Suspected of Having COVID-19 Using Artificial Intelligence”. en. In: *Radiology* 298.1 (Jan. 2021), E18–E28.

[117] Shuo Wang et al. “A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis”. en. In: *Eur. Respir. J.* 56.2 (Aug. 2020).

[118] Harrison X Bai et al. “Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT”. en. In: *Radiology* 296.3 (Sept. 2020), E156–E165.

[119] Lin Li et al. “Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy”. en. In: *Radiology* 296.2 (Aug. 2020), E65–E71.

[120] Kang Zhang et al. “Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography”. en. In: *Cell* 181.6 (June 2020), 1423–1433.e11.

[121] Cheng Jin et al. “Development and evaluation of an artificial intelligence system for COVID-19 diagnosis”. en. In: *Nat. Commun.* 11.1 (Oct. 2020), p. 5088.

[122] Minghuan Wang et al. “Deep learning-based triage and analysis of lesion burden for COVID-19: a retrospective study with external validation”. en. In: *Lancet Digit Health* 2.10 (Oct. 2020), e506–e515.

[123] Florian Jungmann et al. “Commercial AI solutions in detecting COVID-19 pneumonia in chest CT: not yet ready for clinical implementation?” en. In: *Eur. Radiol.* 32.5 (May 2022), pp. 3152–3160.

[124] Zhang Li et al. “From community-acquired pneumonia to COVID-19: a deep learning-based method for quantitative analysis of COVID-19 on thick-section CT scans”. en. In: *Eur. Radiol.* 30.12 (Dec. 2020), pp. 6828–6837.

[125] Jiantao Pu et al. “Automated quantification of COVID-19 severity and progression using chest CT images”. en. In: *Eur. Radiol.* 31.1 (Jan. 2021), pp. 436–446.

[126] Nastaran Enshaei et al. “COVID-rate: an automated framework for segmentation of COVID-19 lesions from chest CT images”. en. In: *Sci. Rep.* 12.1 (Feb. 2022), p. 3212.

[127] Bo Wang et al. “AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system”. en. In: *Appl. Soft Comput.* 98 (Jan. 2021), p. 106897.

[128] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. “Instance Normalization: The Missing Ingredient for Fast Stylization”. In: (July 2016).

[129] Samuel G Armato 3rd and Anna K Nowak. “Revised Modified Response Evaluation Criteria in Solid Tumors for Assessment of Response in Malignant Pleural Mesothelioma (Version 1.1)”. en. In: *J. Thorac. Oncol.* 13.7 (July 2018), pp. 1012–1021.

[130] J M S Wason and S R Seaman. “Using continuous data on tumour measurements to improve inference in phase II cancer studies”. In: *Stat. Med.* (2013).

[131] Aurélie Lombard et al. “Impact of tumour size measurement inter-operator variability on model-based drug effect evaluation”. In: *Cancer Chemother. Pharmacol.* 85.4 (Apr. 2020), pp. 817–825.

[132] James M S Wason, Adrian P Mander, and Tim G Eisen. “Reducing sample sizes in two-stage phase II cancer trials by using continuous tumour shrinkage end-points”. In: *Eur. J. Cancer* 47.7 (May 2011), pp. 983–989.

[133] Giorgio V Scagliotti et al. “Nintedanib in combination with pemetrexed and cisplatin for chemotherapy-naive patients with advanced malignant pleural mesothelioma (LUME-Meso): a double-blind, randomised, placebo-controlled phase 3 trial”. en. In: *Lancet Respir Med* 7.7 (July 2019), pp. 569–580.

[134] S Popat et al. “A multicentre randomised phase III trial comparing pembrolizumab versus single-agent chemotherapy for advanced pre-treated malignant pleural mesothelioma: the European Thoracic Oncology Platform (ETOP 9-15) PROMISE-meso trial”. en. In: *Ann. Oncol.* 31.12 (Dec. 2020), pp. 1734–1745.

[135] Tomer Meirson et al. “Comparison of 3 Randomized Clinical Trials of Frontline Therapies for Malignant Pleural Mesothelioma”. en. In: *JAMA Netw Open* 5.3 (Mar. 2022), e221490.

[136] R L Prentice. “Surrogate endpoints in clinical trials: definition and operational criteria”. en. In: *Stat. Med.* 8.4 (Apr. 1989), pp. 431–440.

[137] L S Freedman, B I Graubard, and A Schatzkin. “Statistical validation of intermediate endpoints for chronic diseases”. en. In: *Stat. Med.* 11.2 (Jan. 1992), pp. 167–178.

[138] M Buyse and G Molenberghs. “Criteria for the validation of surrogate endpoints in randomized experiments”. en. In: *Biometrics* 54.3 (Sept. 1998), pp. 1014–1029.

[139] T Frauenfelder et al. “Volumetry: an alternative to assess therapy response for malignant pleural mesothelioma?” en. In: *Eur. Respir. J.* 38.1 (July 2011), pp. 162–168.

[140] Christian Plathow et al. “Therapy response in malignant pleural mesothelioma-role of MRI using RECIST, modified RECIST and volumetric approaches in comparison with CT”. en. In: *Eur. Radiol.* 18.8 (Aug. 2008), pp. 1635–1643.

[141] Andrew C Kidd et al. “Fully automated volumetric measurement of malignant pleural mesothelioma by deep learning AI: validation and comparison with modified RECIST response criteria”. en. In: *Thorax* (Feb. 2022).

[142] C G Moertel and J A Hanley. “The effect of measuring error on the results of therapeutic trials in advanced cancer”. en. In: *Cancer* 38.1 (July 1976), pp. 388–394.

[143] Geoffrey R Oxnard, Samuel G Armato 3rd, and Hedy L Kindler. “Modeling of mesothelioma growth demonstrates weaknesses of current response criteria”. en. In: *Lung Cancer* 52.2 (May 2006), pp. 141–148.

[144] Cornedine J de Gooijer et al. “Switch Maintenance Gemcitabine after First Line Chemotherapy in Patients with Malignant Mesothelioma: A Randomized Open Label Phase II Trial (NVALT19)”. Jan. 2020.

[145] Maria J Disselhorst et al. “Ipilimumab and nivolumab in the treatment of recurrent malignant pleural mesothelioma (INITIATE): results of a prospective, single-arm, phase 2 trial”. en. In: *Lancet Respir Med* 7.3 (Mar. 2019), pp. 260–270.

[146] Josine Quispel-Janssen et al. “OA13.01 A Phase II Study of Nivolumab in Malignant Pleural Mesothelioma (NivoMes): with Translational Research (TR) Biopies”. In: *J. Thorac. Oncol.* 12.1 (Jan. 2017), S292–S293.

[147] Kevin Bernardus Wilhelmus Lipman et al. “Pleural plaque volume correlation to lung function and artificial intelligence-driven pleural plaque quantification”. en. In: *Eur. Respir. J.* 58.suppl 65 (Sept. 2021).

[148] Sergios Gatidis et al. “A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions”. en. In: *Sci Data* 9.1 (Oct. 2022), p. 601.

[149] Stanislav Nikolov et al. “Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy”. In: (Sept. 2018).

[150] James D Brierley, Mary K Gospodarowicz, and Christian Wittekind. *TNM Classification of Malignant Tumours*. en. John Wiley & Sons, Jan. 2017.

[151] Lawek Berzenji, Paul E Van Schil, and Laurens Carp. “The eighth TNM classification for malignant pleural mesothelioma”. en. In: *Transl Lung Cancer Res* 7.5 (Oct. 2018), pp. 543–549.

[152] Marc Buyse et al. “Surrogacy Beyond Prognosis: The Importance of “Trial-Level” Surrogacy”. en. In: *Oncologist* 27.4 (Apr. 2022), pp. 266–271.

[153] Eyjolfur Gudmundsson et al. “Deep learning-based segmentation of malignant pleural mesothelioma tumor on computed tomography scans: application to scans demonstrating pleural effusion”. en. In: *J Med Imaging (Bellingham)* 7.1 (Jan. 2020), p. 012705.

[154] Wael Brahim et al. “Malignant pleural mesothelioma segmentation from thoracic CT scans”. In: *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. May 2017, pp. 1–5.

[155] Alessandra Curioni Fontecedro. *SAKK 17/18 (ORIGIN) MPM & NSCLC >1st Line Gemci & Atezo Ph II*. `https://clinicaltrials.gov/ct2/show/NCT04480372`. Accessed: 2023-8-16. July 2020.

[156] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241.

[157] Camila Gonzalez et al. “Detecting when pre-trained nnU-Net models fail silently for Covid-19 lung lesion segmentation”. In: (July 2021).

[158] Camila González et al. “Lifelong nnU-Net: a framework for standardized medical continual learning”. en. In: *Sci. Rep.* 13.1 (June 2023), p. 9381.

[159] Neal Corson et al. “Characterization of mesothelioma and tissues present in contrast-enhanced thoracic CT scans”. en. In: *Med. Phys.* 38.2 (Feb. 2011), pp. 942–947.

[160] Anna Zapaishchykova et al. “SegmentationReview: A Slicer3D extension for fast review of AI-generated segmentations”. In: *Software Impacts* 17 (Sept. 2023), p. 100536.

[161] Andrew T Jebb, Vincent Ng, and Louis Tay. “A Review of Key Likert Scale Development Advances: 1995-2019”. en. In: *Front. Psychol.* 12 (May 2021), p. 637547.

[162] Hamparsum Bozdogan. “Model selection and Akaike’s Information Criterion (AIC): The general theory and its analytical extensions”. In: *Psychometrika* 52.3 (Sept. 1987), pp. 345–370.

[163] R D C Team. “A language and environment for statistical computing”. In: *http://www.R-project.org* (2009).

[164] Cameron Davidson-Pilon. “lifelines: survival analysis in Python”. In: *J. Open Source Softw.* 4.40 (Aug. 2019), p. 1317.

[165] Eric J Topol. “High-performance medicine: the convergence of human and artificial intelligence”. en. In: *Nat. Med.* 25.1 (Jan. 2019), pp. 44–56.

[166] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. “Resistance to Medical Artificial Intelligence”. In: *J. Consum. Res.* 46.4 (Dec. 2019), pp. 629–650.

[167] Mohamed A Abdou. “Literature review: efficient deep neural networks techniques for medical image analysis”. In: *Neural Comput. Appl.* 34.8 (Apr. 2022), pp. 5791–5812.

[168] Susan Cheng Shelmerdine et al. “Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare”. en. In: *BMJ Health Care Inform* 28.1 (Aug. 2021).

[169] Hiroaki Arakawa et al. “Asbestosis and other pulmonary fibrosis in asbestos-exposed workers: high-resolution CT features with pathological correlations”. en. In: *Eur. Radiol.* 26.5 (May 2016), pp. 1485–1492.

[170] Alice C Yu, Bahram Mohajer, and John Eng. “External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review”. en. In: *Radiol Artif Intell* 4.3 (May 2022), e210064.

[171] Nathaniel Yang. “On the Importance of Proper Window and Level Settings in Temporal Bone CT Imaging”. en. In: *Philipp J Otolaryngol Head Neck Surg* 35.2 (Dec. 2020), pp. 51–51.

[172] Dandu Ravi Varma. “Managing DICOM images: Tips and tricks for the radiologist”. en. In: *Indian J. Radiol. Imaging* 22.1 (Jan. 2012), pp. 4–13.

[173] François Pontana et al. “Chest computed tomography using iterative reconstruction vs filtered back projection (Part 2): image quality of low-dose CT examinations in 80 patients”. en. In: *Eur. Radiol.* 21.3 (Mar. 2011), pp. 636–643.

[174] K Eldevik, W Nordhøy, and A Skretting. “Relationship between sharpness and noise in CT images reconstructed with different kernels”. en. In: *Radiat. Prot. Dosimetry* 139.1-3 (Feb. 2010), pp. 430–433.

[175] Priyanka Prakash et al. “Diffuse lung disease: CT of the chest with adaptive statistical iterative reconstruction technique”. en. In: *Radiology* 256.1 (July 2010), pp. 261–269.

[176] Trieu-Nghi Hoang-Thi et al. “Deep learning for lung disease segmentation on CT: Which reconstruction kernel should be used?” en. In: *Diagn. Interv. Imaging* 102.11 (Nov. 2021), pp. 691–695.

[177] Stephan P Blazis et al. “Effect of CT reconstruction settings on the performance of a deep learning based lung nodule CAD system”. en. In: *Eur. J. Radiol.* 136 (Mar. 2021), p. 109526.

[178] Rahul Pandey et al. “Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning”. In: *Int. J. Hum. Comput. Stud.* 160 (Apr. 2022), p. 102772.

[179] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. “An analysis of human factors and label accuracy in crowdsourcing relevance judgments”. In: *Inf. Retr. Boston.* 16.2 (Apr. 2013), pp. 138–178.

[180] Anne Kathrine Petersen Bach et al. ““If I Had All the Time in the World”: Ophthalmologists’ Perceptions of Anchoring Bias Mitigation in Clinical AI Support”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23 Article 16. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–14.

[181] C Schulz. “Design and Evaluation of a Prototype for a Platform for AI algorithms in Medical Imaging : A Human-Centered Approach”. PhD thesis. University of Twente, May 2023.

[182] Liza C Villaruz and Mark A Socinski. “The clinical viewpoint: definitions, limitations of RECIST, practical considerations of measurement”. en. In: *Clin. Cancer Res.* 19.10 (May 2013), pp. 2629–2636.

[183] Serena Grimaldi, Marie Terroir, and Caroline Caramella. “Advances in oncological treatment: limitations of RECIST 1.1 criteria”. en. In: *Q. J. Nucl. Med. Mol. Imaging* 62.2 (June 2018), pp. 129–139.

Impact

The contributions of this thesis promise transformative effects in healthcare. While **Part I** provides essential foundational insights, it's primarily **Part II**'s findings that present a potential paradigm shift in treating pleural mesothelioma, which holds societal ramifications.

This research illustrates AI models' potential to disseminate expert knowledge from specialized centers to regional hospitals, extending even to countries with high mesothelioma prevalence but limited expertise. This distribution of medical expertise not only narrows the disparity in diagnostic and therapeutic capabilities across healthcare facilities but also enables a more standardized evaluation in clinical trials on a global scale. Such standardization ensures high-quality care for patients worldwide, regardless of their geographical locations.

Furthermore, by adopting AI-driven volumetric assessments, uncertainties surrounding tumor growth are reduced, which is one of the stress factors for patients. Accurate tumor evaluations allow treating physicians to provide clearer feedback, enabling patients to make well-informed decisions about their treatment trajectories. This clarity can potentially lead to better quality of life by helping patients transition away from ineffective treatments causing detrimental side effects.

In essence, the arrival of AI-powered standardized methodologies for patient disease assessment holds the potential to improve clinical trial evaluations and patient care. By improving the precision, consistency, and reliability of disease evaluations, this research enhances informed clinical decisions and offers a brighter, more equitable future for patients and society alike.

Summary

This PhD thesis investigates the application of artificial intelligence (AI) in quantifying disease status and response to therapy for patients with asbestosis, pleural plaques, COVID-19, and pleural mesothelioma. The thesis is divided into two parts: Enhancing Disease Quantification at Baseline and Evaluating Therapeutic Response in Pleural Mesothelioma.

In Part I, our first study in ***Chapter 2*** explores using AI to assist in deciding which asbestosis patients (a lung disease caused by asbestos exposure) should receive government support. We analyzed 523 cases in the Netherlands, using AI to review chest CT scans and lung function tests. The AI's decisions were compared with those of a panel of lung doctors. Results showed the AI system was quite accurate, even more so when combined with lung function test data. This research suggests AI could be a valuable tool in streamlining and improving the fairness of the support application process for asbestosis patients.

Chapter 3 tested the AI model developed in *Chapter 2* in a real-life setting without having an impact on the decision. We included all applicants seeking asbestosis compensation in a Dutch nationwide cohort from September 2020 to July 2022. The AI's assessments were compared with the evaluations of the three pulmonologists. If the AI was unsure, two more reviewers joined the assessment. The results showed that the AI was quite accurate, but it didn't hit our target for sensitivity – the ability to correctly identify those with asbestosis. The AI did well in terms of specificity – correctly identifying those without the disease – and overall accuracy. However, because it didn't meet our sensitivity goal, we believe more work is needed.

In this study of ***Chapter 4***, we focused on pleural plaques (PP), which

are signs of long-term asbestos exposure, and their impact on lung function. We also aimed to speed up the process of measuring PP by using AI. We trained an AI model to identify PP in CT scans of patients who had been exposed to asbestos. This model was compared with the work of radiologists. We also looked at how the volume of PP related to different lung function tests. The AI was trained on 422 CT scans and tested its accuracy in predicting PP volume. The results showed a strong correlation between the AI's predictions and the actual measurements. We also found that larger PP volumes are associated with a decrease in some lung function measures, giving us new insights into the effects of asbestos exposure on lung health.

Chapter 5 evaluates an AI model for analyzing COVID-19 in CT scans. The model was trained to identify lung infections and estimate COVID-19 likelihood. To test the performance in real-world situations, we used a different set of 400 CT scans from various centers. The results showed that while the model was excellent at identifying lung contours both in our tests and the external ones, it struggled with detecting lung infections when used outside the initial testing environment. Additionally, its effectiveness in determining the likelihood of COVID-19 also dropped in these external tests. The takeaway from this study is that while the AI model showed promise, its performance varied significantly in different settings. This highlights the importance of testing AI models in various real-world conditions, especially for clinical tools, to ensure they are reliable and effective in all potential environments.

In Part II of this thesis, we developed in ***Chapter 6*** a new way to measure how well treatments work for pleural mesothelioma (PM), a type of cancer linked to asbestos exposure. We created an AI algorithm that can automatically measure the volume of PM tumors in CT scans. This AI tool was designed to make it easier and more accurate to track changes in tumor size over time, which is important for understanding how well treatments are working. The AI's performance was impressive in our tests. It matched the expert-segmented tumor volumes with 89% accuracy in an internal test set. When we used it in a large European dataset, the AI showed 98% overlap with expert corrections,

demonstrating its high reliability. In a side-by-side comparison of CT scans of a smaller phase II trial, radiologists often preferred the AI's segmentation over the manual volume extraction of a radiologist.

We also used the AI to see if the size of the tumor at the beginning of treatment could predict how patients would fare. We divided the tumors into four size groups and found that the initial tumor size was significantly linked to the patient's overall survival.

Furthermore, we introduced new criteria called ARTIMES to evaluate how PM tumors respond to treatment. These criteria use both a fixed size change and a percentage increase to determine if the cancer is getting worse. In our studies, ARTIMES spotted tumor growth about 7 weeks earlier than traditional methods. It also proved to be more effective in predicting patients' overall survival.

Samenvatting

Dit proefschrift onderzoekt de toepassing van kunstmatige intelligentie (AI) bij het kwantificeren van ziektestatus voor patiënten met asbestose, pleurale plaques, en COVID-19. Verder stellen we een nieuwe methode voor om respons op therapie bij pleuraal mesothelioom te bepalen. Het proefschrift is verdeeld in twee delen: verbetering van ziektekwantificering op de eerste scan en evaluatie van de respons op de behandeling bij pleuraal mesothelioom.

In deel I onderzoeken we in ***Hoofdstuk 2*** het gebruik van AI om te helpen bepalen welke patiënten asbestose hebben (een longziekte veroorzaakt door asbestblootstelling), en daardoor in aanmerking komen voor overheidssteun. We analyseerden 523 casussen in Nederland, waarbij AI werd gebruikt om CT-scans en longfunctietests te beoordelen. De beslissingen van de AI werden vergeleken met die van een panel van longartsen. De resultaten toonden aan dat het AI-systeem vrij nauwkeurig was, en zelfs verbeterde toen het gecombineerd werd met longfunctietestgegevens. Dit onderzoek suggereert dat AI een waardevol hulpmiddel kan zijn bij het stroomlijnen en verbeteren van de consistentie van het aanvraagproces voor overheidssteun aan asbestosepatiënten.

Hoofdstuk 3 testte het AI-model ontwikkeld in *Hoofdstuk 2* in de echte beoordelingopzet zonder invloed op de uiteindelijke beslissing. We hebben alle aanvragers voor asbestosecompensatie in een landelijke Nederlandse cohort van september 2020 tot juli 2022 opgenomen. De beoordelingen van de AI werden vergeleken met de evaluaties van de drie longartsen. Als de AI onzeker was, sloten twee extra beoordelaars zich aan bij de beoordeling. De resultaten toonden aan dat de AI vrij nauwkeurig was, maar ons doel voor gevoeligheid – het correct identificeren van mensen met asbestose – niet haalde. De AI deed het goed

qua specificiteit – het correct identificeren van mensen zonder de ziekte – en algehele nauwkeurigheid. Omdat het echter ons gevoeligheidsdoel niet bereikte, is er meer werk nodig.

In de studie van ***Hoofdstuk 4*** hebben we ons gericht op pleurale plaques (PP), die kunnen voorkomen na langdurige blootstelling aan asbest. Mensen met PP komen niet in aanmerking voor overheidssteun, omdat er niet duidelijk is aangetoond dat PP een negatieve impact op de kwaliteit van leven heeft. Om dit te onderzoeken, beoogden we het proces van het meten van PP versnellen door het gebruik van AI. We trainden een AI-model om PP te identificeren in CT-scans van patiënten die blootgesteld waren aan asbest. Dit model werd vergeleken met het werk van radiologen. We keken ook naar hoe het volume van PP gerelateerd was aan verschillende longfunctietests. De AI werd getraind op 422 CT-scans en testte de nauwkeurigheid bij het voorspellen van PP-volume. De resultaten toonden een sterke correlatie tussen de voorspellingen van de AI en de werkelijke metingen van de radiologen. We vonden ook dat grotere PP-volumes geassocieerd zijn met een afname van sommige longfunctiemetingen, waardoor we nieuwe inzichten kregen in de effecten van pleurale plaques op de longgezondheid.

Hoofdstuk 5 evalueert een AI-model voor het analyseren van COVID-19 in CT-scans. Het model werd getraind om longinfecties te identificeren en de waarschijnlijkheid van COVID-19 te schatten. Om de prestaties in een realistische situatie te testen, gebruikten we een andere set van 400 CT-scans uit verschillende centra. De resultaten toonden aan dat het model uitstekend was in het identificeren van longcontouren, maar het moeite had met het detecteren van longinfecties wanneer deze in een nieuw ziekenhuis werd gebruikt. Bovendien daalde de effectiviteit in het bepalen van de waarschijnlijkheid van COVID-19 ook. De conclusie van deze studie is dat hoewel het AI-model belovend was, de prestaties aanzienlijk varieerden in nieuwe omgevingen. Dit benadrukt het belang van het testen van AI-modellen in verschillende omstandigheden om ervoor te zorgen dat ze betrouwbaar en effectief zijn in meerdere ziekenhuizen.

In deel II van dit proefschrift hebben we in ***Hoofdstuk 6*** een nieuwe manier ontwikkeld om te meten hoe goed behandelingen werken voor pleuraal mesothelioom (PM), een soort kanker veroorzaakt door asbestblootstelling. We creëerden een AI-algoritme dat automatisch het volume van PM-tumoren in CT-scans kan meten. Dit AI-hulpmiddel is ontworpen om het gemakkelijker en nauwkeuriger te maken om veranderingen in tumorgrootte over tijd bij te houden, wat belangrijk is om te begrijpen hoe goed behandelingen werken. De prestaties van de AI waren indrukwekkend in onze test. Het kwam overeen met de door experts bepaalde tumorvolumes met 89% nauwkeurigheid in een onafhankelijke interne testset. Toen we de AI inzette in een grote Europese dataset, toonde het 98% overlap na correcties van medische experts, wat een hoge betrouwbaarheid aantoont. In een vergelijking van CT-scans van een kleinere fase II-studie gaven radiologen vaak de voorkeur aan de volume-bepaling van de AI ten opzichte van de handmatige volumebepaling van een radioloog.

We gebruikten ook de AI om te zien of de grootte van de tumor aan het begin van de behandeling kon voorspellen hoe het met patiënten zou gaan. We verdeelden de tumoren in vier groepen van tumorvolumes en ontdekten dat de aanvangsgrootte van de tumor significant gekoppeld was aan de algehele overleving van de patiënt.

Verder introduceerden we nieuwe criteria genaamd ARTIMES om te evalueren hoe PM-tumoren reageren op behandeling. Deze criteria gebruiken zowel een absolute verandering in volume als een percentage toename om te bepalen of de kanker erger wordt. In onze studies ontdekte ARTIMES tumorgroei ongeveer 7 weken eerder dan traditionele methoden. Het bleek ook effectiever te zijn in het voorspellen van de algehele overleving van patiënten.

Acknowledgments

Kevin B.W. Groot Lipman
Maastricht
January 22nd, 2024

I would like to express my appreciation to my promotor, **Prof. Dr. Regina Beets-Tan**, and co-copromotors **Dr. Sjaak Burgers** and **Dr. Stefano Trebeschi** for their guidance throughout this journey. Your insights and consistent focus on clinical relevance have shaped my development as a researcher, and I am grateful for the autonomy you allowed me in choosing the projects. Furthermore, working alongside dedicated and skillful individuals in your departments has been a pleasure.

I wish to extend my gratitude to the assessment committee: **Prof. Dr. Frits Franssen**, **Prof. Dr. André Dekker**, **Prof. Dr. Christiane Kuhl**, and **Dr. Jonas Teuwen** for the evaluation of this thesis. Additionally, I am grateful to **Prof. Dr. Philippe Lambin** and **Dr. Wouter van Geffen** for taking part in the defense committee.

I am also indebted to **Zuhir Bodalal**, **Laurens Topff**, **Dianne de Gooijer**, and **Illaa Smesseim** for the outstanding collaboration during the projects. Your support, willingness to invest countless hours in discussions, and ability to help me overcome challenges have made an immense difference in my research experience, and I cannot thank you enough.

My heartfelt gratitude goes to all the radiologists who have contributed to the projects, and in specific **Thierry Boellaard**, **Rianne Wittenberg** and the **ESOR fellows**; the remarkable amount of work per-

formed and dedication are truly appreciated. Your feedback from the radiological perspective has shaped the projects for the better.

I also want to acknowledge my wonderful **colleagues in the O-building**, whose camaraderie and daily interactions have made this PhD an enjoyable one. Watching our offices fill up with people with such diverse backgrounds and personalities has been a true pleasure. I always enjoyed the sight of the hallways being alive with engaging conversations about research and life. Mostly due to you, I cannot remember a day I didn't want to go to work.

I want to give a special shoutout to all the **unsung heroes** who have been essential in supporting this research journey. First, the Radiology Datadesk team, **Joost van Griethuysen** and **Artem Khmelinskii**, for their continuous work on creating streamlined software packages to anonymize and export imaging scans for research purposes. Second, the RHPC administrators, particularly **Torben Wriedt** and **Ameer Alkhier**, whose dedication has been essential in keeping the GPU servers running to train our AI models. At last, thank you to all the other individuals working in supporting research; your efforts have been the backbone of our achievements.

My sincerest appreciation goes to **my parents Theo and Monique Groot Lipman and brother Justin** for their unwavering support throughout my life. Your unconditional backing, even when I made choices you might not have agreed with, means the world to me.

Finally, I am immensely grateful to **Kaylee van Duren** for being a phenomenal partner, where the positivity has carried over from our home life into work. Our shared commutes helped maintain a balanced work/life schedule, and I have cherished both our quiet morning rides to work and the lively conversations on our way back home.

Papers outside of Thesis

Doremalen, Rob F. M. van, **Kevin B. W. Groot Lipman**, Esther van 't Riet, Hans Torrenga, Maria M. Smits, and Ferdinand van der Heijden. 2021. "Novel Breast Specimen Orientation Approach through 3D Visualizations for Relocating Inadequate Margins Based on the Surgical Clips: Feasibility Study." Research Square. https://doi.org/10.21203/rs.3.rs-819800/v1.

Bodalal, Zuhir, Stefano Trebeschi, Ivar Wamelink, **Kevin Groot Lipman**, Teresa Bucho, Nick van Dijk, Thierry Boellaard, Selam Waktola, and Regina G. H. Beets-Tan. 2022. "The Future of Artificial Intelligence Applied to Immunotherapy Trials." In Neoadjuvant Immunotherapy Treatment of Localized Genitourinary Cancers: Multidisciplinary Management, edited by Andrea Necchi and Philippe E. Spiess, 265–84. Cham: Springer International Publishing.

Bucho, Teresa T., Renaud Tissier, **Kevin Groot Lipman**, Zuhir Bodalal, Andrea Delli Pizzi, Thi Dan Linh Nguyen-Kim, Regina Beets-Tan, and Stefano Trebeschi. 2022. "How Does Target Lesion Selection Affect RECIST? A Computer Simulation Study." Invest Radiol. 2023 Nov 3. doi: 10.1097/RLI.0000000000001045. Epub ahead of print. PMID: 37921780.

Coorens, Nadine A., **Kevin Groot Lipman**, Sanjith P. Krishnam, Can Ozan Tan, Lejla Alic, and Rajiv Gupta. 2023. "Intracerebral Hemorrhage Segmentation on Noncontrast Computed Tomography Using a Masked Loss Function U-Net Approach." Journal of Computer Assisted Tomography 47 (1): 93–101.

Ponsiglione, Andrea, Arnaldo Stanzione, Gaia Spadarella, Agah Baran, Luca Alessandro Cappellini, **Kevin Groot Lipman**, Peter Van Ooijen & Renato Cuocolo. 2023. "Ovarian Imaging Radiomics Quality

Score Assessment: An EuSoMII Radiomics Auditing Group Initiative." European Radiology 33 (3): 2239–47.

Bodalal, Zuhir, Nino Bogveradze, Leon C. Ter Beek, Jose G. van den Berg, Joyce Sanders, Ingrid Hofland, Stefano Trebeschi, **Kevin B. W. Groot Lipman**, Koen Storck, Eun Kyoung Hong, Natalya Lebedyeva, Monique Maas, Regina G. H. Beets-Tan, Fernando M. Gomez & Ieva Kurilova. 2023. "Radiomic Signatures from T2W and DWI MRI Are Predictive of Tumour Hypoxia in Colorectal Liver Metastases." Insights into Imaging 14 (1): 133.

About the author

Kevin Bernardus Wilhelmus Groot Lipman was born on December 9, 1993, in Deventer, The Netherlands, and grew up in Lettele with his parents and brother. He began his education at Etty Hillesum Lyceum, Deventer, where he completed Atheneum in NG/NT in 2012. Subsequently, he attended the University of Twente, Enschede, and completed his Bachelor's (2017) and Master's (2020) in Technical Medicine - Medical Imaging & Intervention.

His professional journey began as a Technical Medicine intern at various hospitals, such as Antoni van Leeuwenhoek, UMC Utrecht, Deventer Ziekenhuis, and Massachusetts General Hospital / Harvard Medical School. He gained insights into medical imaging, deep learning, and clinics during these internships.

In 2020, he started his PhD candidacy at the Netherlands Cancer Institute on the standardization of respiratory disease evaluation by means of Artificial Intelligence, where he developed AI models of which one is in the process of clinical implementation. During his PhD, he also took up the role of a visiting researcher at the University Hospital of Zurich (USZ), Switzerland. His primary responsibility at USZ was the external validation of the developed AI model.

Throughout his PhD journey, he actively contributed to various conferences and seminars, delivering talks at events like European Congress of Radiology and European Respiratory Society International Congress. He reviewed for European Radiology, European Journal of Radiology, and Medical Physics. Additionally, he was the main supervisor for two Bachelor's students in Information Science and eight Master's students in Artificial Intelligence, Data Science, Information Science, Human Factors and Ergonomics, and Technical Medicine.