OPEN ACCESS

Original research

Fully automated volumetric measurement of malignant pleural mesothelioma by deep learning AI: validation and comparison with modified RECIST response criteria

Andrew C Kidd,[1] Owen Anderson,[2,3] Gordon W Cowell,[4] Alexander J Weir,[3] Jeremy P Voisey,[3] Matthew Evison,[5] Selina Tsim,[1,6] Keith A Goatman,[3] Kevin G Blyth[1,6,7]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/thoraxjnl-2021-217808).

[1]Glasgow Pleural Disease Unit, Queen Elizabeth University Hospital, Glasgow, UK
[2]School of Computing Science, University of Glasgow, Glasgow, UK
[3]Canon Medical Research Europe Ltd, Edinburgh, UK
[4]Department of Imaging, Queen Elizabeth University Hospital, Glasgow, UK
[5]Department of Respiratory Medicine, University Hospital of South Manchester, Manchester, UK
[6]Institute of Cancer Sciences, University of Glasgow, Glasgow, UK
[7]Beatson Institute for Cancer Research, Glasgow, UK

Correspondence to
Professor Kevin G Blyth, Institute of Cancer Sciences, University of Glasgow, Glasgow, Glasgow, UK; Kevin.Blyth@glasgow.ac.uk

Received 15 June 2021
Accepted 20 December 2021
Published Online First 2 February 2022

To cite: Kidd AC, Anderson O, Cowell GW, *et al*. *Thorax* 2022;**77**:1251–1259.

ABSTRACT

Background In malignant pleural mesothelioma (MPM), complex tumour morphology results in inconsistent radiological response assessment. Promising volumetric methods require automation to be practical. We developed a fully automated Convolutional Neural Network (CNN) for this purpose, performed blinded validation and compared CNN and human response classification and survival prediction in patients treated with chemotherapy.

Methods In a multicentre retrospective cohort study; 183 CT datasets were split into training and internal validation (123 datasets (80 fully annotated); 108 patients; 1 centre) and external validation (60 datasets (all fully annotated); 30 patients; 3 centres). Detailed manual annotations were used to train the CNN, which used two-dimensional U-Net architecture. CNN performance was evaluated using correlation, Bland-Altman and Dice agreement. Volumetric response/progression were defined as ≤30%/≥20% change and compared with modified Response Evaluation Criteria In Solid Tumours (mRECIST) by Cohen's kappa. Survival was assessed using Kaplan-Meier methodology.

Results Human and artificial intelligence (AI) volumes were strongly correlated (validation set r=0.851, p<0.0001). Agreement was strong (validation set mean bias +31 cm^3 (p=0.182), 95% limits 345 to +407 cm^3). Infrequent AI segmentation errors (4/60 validation cases) were associated with fissural tumour, contralateral pleural thickening and adjacent atelectasis. Human and AI volumetric responses agreed in 20/30 (67%) validation cases κ=0.439 (0.178 to 0.700). AI and mRECIST agreed in 16/30 (55%) validation cases κ=0.284 (0.026 to 0.543). Higher baseline tumour volume was associated with shorter survival.

Conclusion We have developed and validated the first fully automated CNN for volumetric MPM segmentation. CNN performance may be further improved by enriching future training sets with morphologically challenging features. Volumetric response thresholds require further calibration in future studies.

Key messages

What is the key question?
⇒ Can an artificial intelligence (AI) system be trained to accurately measure mesothelioma tumour volume on CT images, without any human input?

What is the bottom line?
⇒ Fully automated tumour segmentation was possible and the deep learning AI algorithm performed well in a diverse and previously unseen external validation set drawn from three UK centres.

Why read on?
⇒ This is the largest volumetry study performed in mesothelioma and the first description of a fully automated and externally validated AI segmentation tool.

INTRODUCTION

Malignant pleural mesothelioma (MPM) is an incurable cancer associated with previous asbestos exposure. For nearly two decades, platinum-pemetrexed chemotherapy has been the established standard of first-line care,[1] although recent data report superior survival with combination immune checkpoint inhibition.[2 3] Radiological assessment of treatment response is a critical part of routine care for most patients, and a key metric for clinical trials. Response assessment is, however, notoriously difficult in MPM because the primary tumour has a complex morphology, violating assumptions regarding spherical growth that underpin the Response Evaluation Criteria In Solid Tumours (RECIST) criteria used in other cancers.[4] Modified RECIST criteria (mRECIST) mitigate errors related to this by allowing the reporter to make six unidimensional tumour measurements at arbitrary positions, generating a summed value, which when compared with summed values from adjacent timepoints, can be codified into complete response (CR), partial response (PR), stable disease (SD) or progressive disease (PD).[5] However, mRECIST grossly oversimplifies true tumour burden and is associated with poor reproducibility, including up to 30% variation between readers,[6] which is large enough to cross response groups based on the same

data. This can be mitigated by multiple readers in clinical trials; however, this increases cost and slows drug development. The inherent technical difficulty also results in a threshold of 'minimally measurable disease', below which early stage patients cannot be reliably assessed and may not therefore be treated or offered entry to trials.[7]

Volumetric tumour measurement is established in lung[8] and other cancers[9–11] and eliminates decision-making about where to make, and how to replicate, unidimensional measures on serial scans. Volumetric measurements based on MRI have recently been shown to outperform traditional T-staging in predicting survival,[12] but MRI is not available routinely and CT remains the primary imaging modality used in MPM. Efforts therefore need to be focused on delivery of CT volumetry, at least in the short term. This requires development of accurate new techniques that minimise reader variation, ideally by reducing reliance on human annotation, which can be exceptionally time-consuming given the volume and complexity of the pleural space. Previous CT volumetry studies that have used human readers report interobserver variation broadly similar to mRECIST,[13 14] so a fully automated tool is needed. Deep learning Convolutional Neural Networks (CNNs) are uniquely well suited to image recognition and classification tasks.[15] Outside biomedicine, CNNs have demonstrated exceptional performance in such tasks, given sufficiently large training datasets. For example, in the ImageNet challenge,[15] which comprised over 14 million labelled images, CNNs now match or exceed human reader performance. In medical settings, millions of images are rarely available, and the labelling process is generally more arduous. In this study, we show that a CNN can be trained to segment MPM without human input, using a relatively small but extremely detailed set of ground truth images. The system's performance is also evaluated in an independent validation set to verify that the CNN did not overfit to the training data.

METHODS

Study design

A multicentre retrospective cohort study was performed. The training and internal validation set comprised 123 CT datasets from 108 patients with MPM from Glasgow. The external validation set comprised 60 CT datasets from 30 patients with MPM from Leicester (n=10), Manchester (n=10) and Glasgow (n=10). The study is reported according to Standards for Quality Improvement Reporting Excellence 2.0 guidelines.[16]

Study objectives

The training and internal validation set was used (1) to train the CNN and report initial performance, (2) to compare reproducibility between human and AI readers and (3) to compare the fidelity of AI segmentations to the reference human ground truth by region overlap Dice coefficient, and to compare that performance with a second human annotation and repeat human annotation by the first reader.

The external validation set was used for (1) a blinded comparison between human and AI volumes by correlation and agreement (Bland-Altman and Dice region overlap), (2) an analysis of anatomical features associated with AI segmentation errors, (3) a comparison between volumetric and mRECIST classification of chemotherapy response and (4) survival analyses based on PD versus non-PD, as defined by human volumes, AI volumes and mRECIST.

Case selection

Cases were selected from two multicentre MPM biomarker studies, led by the senior author, that have recently completed recruitment (DIAPHRAGM, Diagnostic and Prognostic Biomarkers in the Rational Assessment of Mesothelioma[17] and PRISM, Prediction of Resistance to chemotherapy using Somatic Copy Number Variation in Mesothelioma[18]). DIAPHRAGM prospectively recruited 649 patients at presentation for evaluation of diagnostic blood biomarkers, 152 of whom were diagnosed with MPM. In PRISM, 266 MPM tumour blocks and CT scans were retrospectively retrieved for discovery and validation of a genomic predictor classifier of chemotherapy resistance. Cases were selected for the training set based on the following inclusion criteria: (1) recruited to DIAPHRAGM or PRISM in Glasgow; (2) histological diagnosis of MPM; (3) venous phase contrast-enhanced CT available. DIAPHRAGM cases were specifically included in the training set as these had contemporaneous contrast-enhanced MRI. These scans were used to disambiguate tumour from adjacent structures on CT, using the superior soft tissue contrast offered by MRI,[19 20] enhancing ground truth quality. Validation set cases were selected using the following criteria: (1) recruited to PRISM in Glasgow, Leicester or Wythenshawe; (2) histological diagnosis of MPM; (3) prechemotherapy and postchemotherapy CT available, where the postchemotherapy scan was >4 weeks after chemotherapy initiation. These cases were selected to provide a diverse CT dataset collected in different UK centres using different CT scanners and imaging protocols.

Clinical data

Data were extracted from study databases and supplemented by electronic records, including demographics, histological subtype, date of diagnosis, details of chemotherapy and CT imaging. Missing data were recorded as not available. Overall survival (OS, days) was recorded from the date of prechemotherapy CT to death from any cause.

CT image acquisition and mRECIST reporting

CT examinations were performed within routine care, using a variety of scanners (GE Medical Systems BrightSpeed, LightSpeed or Optima 660 or Canon Medical Aquilion). Although local imaging protocols will have varied, all imaging was acquired in the portal venous phase (~65 s following injection of 75–95 mL of iodinated contrast). Multislice helical axial images were reconstructed with a maximum contiguous slice thickness of 2 mm. The mean number of slices was 225. For the purpose of mRECIST classification,[5] validation set scans were re-reported centrally by an expert MPM radiologist (GWC), who was blinded to all other data.

Manual tumour annotation for ground truth

Respiratory physicians with PhD training in MPM imaging (ACK and ST) performed the human tumour annotations, using a trackball mouse and cursor and Myrian Intrasense software (Paris, France). ACK generated the reference ground-truth annotations; ST generated second reader annotations for interobserver data. Tumour boundaries were outlined on every slice in the CT stack, generating a tumour volume integrating the summed areas and slice thickness (see figure 1A). In the DIAPHRAGM cases in the training set, ACK and ST used contemporaneous MRI scans to refine contour drawings, for example, disambiguating loculated fluid from tumour. In 43/123 training datasets, a sparser annotation was performed, with only every fifth slice annotated. This

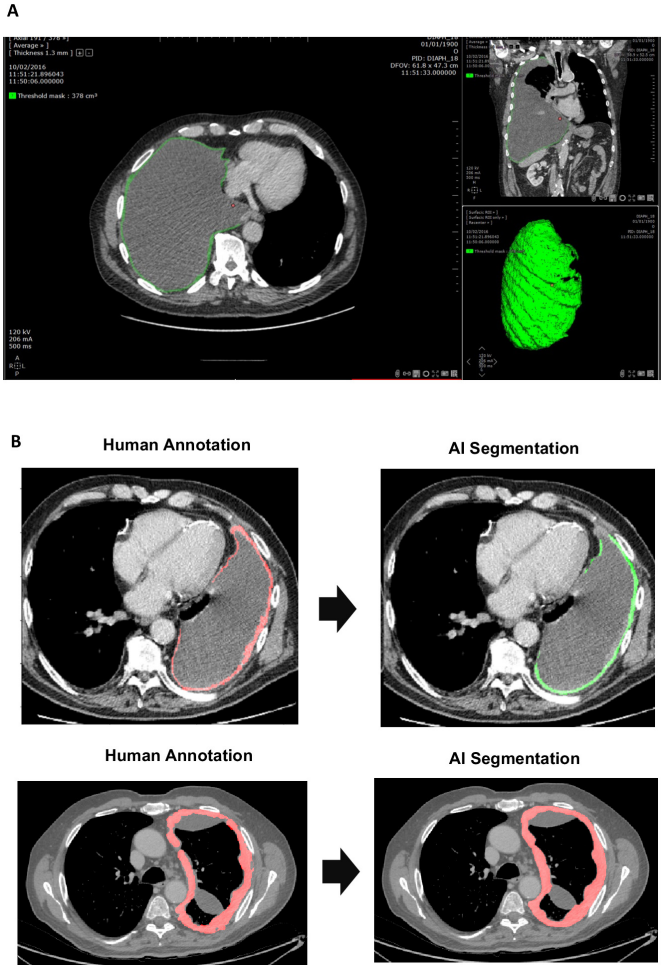


Figure 1 Panel A shows an example of expert human ground truth, with mesothelioma primary tumour volume outlined using trackball mouse and cursor on Myrian Intrasense software. The segmented volume is shown in the bottom right. Panel B shows examples of a human annotated tumour volume (in the two left hand images) and the artificial intelligence (AI)-derived volume from the same case at the same slice position (in the two right hand images). The AI volumes were generated by automatic segmentation without any user prompts.

followed an interim analysis, which demonstrated that adjacent slices were highly correlated, and enabled more cases to be included since each full annotation required ~2.5 hours. The sparse annotations were not included in the internal validation accuracy metrics, which report only the 80 fully annotated cases.

Convolutional Neural Network architecture

A CNN with a two-dimensional U-Net architecture was used[21] on each axial CT slice. The network has three input CT slices: the slice to be segmented plus its two adjacent slices. CT intensities are clipped to −1025 to +1100 Hounsfield units and then normalised to the range −1 to +1. The output from the CNN is a two-dimensional map predicting the likelihood that each pixel contains tumour. An optimal threshold (as determined on a subset of the training data) is applied to this map resulting in a binary mask of the tumour pixels. This is repeated for all slices in the stack, generating a full volumetric segmentation. To increase robustness and to provide an estimation of confidence, seven CNN models were trained using a sevenfold division of the training data and the volume measurements ensembled. A more detailed technical description of the method has previously been reported by us in conference proceedings.[22]

Training and internal validation

Internal validation was by sevenfold cross validation. The sparse annotations were included in the training sets for all folds but were not used for volume accuracy metrics. The training data were divided as follows: the 80 fully annotated datasets were randomly assigned to the seven folds, resulting in 11 or 12 datasets per fold. During evaluation of each fold, the remaining six folds were split 30:70 between (a) a set used to select the best performing model and determine the optimal threshold and (b) the training set, which also included the 43 sparsely annotated volumes. To avoid biasing the algorithm towards images with more tumour-containing slices, all training annotations were sparsely sampled during training.

Volumetric response classification

In the external validation set only, human and AI volume change following chemotherapy was computed for each case as ((post-chemotherapy volume minus prechemotherapy volume)/prechemotherapy volume)×100 (%). Volumetric PR required ≥30% reduction, PD required ≥20% increase. SD was recorded if volume change did not meet PR/PD thresholds. The selection of these criteria was based on previous mathematical modelling reported by Oxnard *et al*,[23] suggesting these volumetric thresholds approximate accurately to unidimension-based mRECIST, assuming the volume imaged approximates a crescent-shaped prism (see figure 1A for images supporting this assumption).

Statistical analysis

Given the exploratory nature of the study, a sample size calculation was not performed. Individual data are summarised by median (IQR) or mean (SD) depending on their distribution, but since most variables were non-normally distributed non-parametric tests were used for all comparisons. The Wilcoxon matched-pairs signed rank test was used to compare paired volume data (human vs AI volumes, prechemotherapy vs postchemotherapy). Spearman's rho test was used for correlation and agreement was evaluated using Bland-Altman plots. The Dice coefficient (equivalent to the F1 score) was used to quantify region overlap between different volumetry methods or readers. Cohen's kappa statistic was used to quantify agreement between chemotherapy response classification by human volumetry, AI volumetry and mRECIST. Differences in volume between AI-defined and mRECIST-defined PR, SD and PD groups were compared by Kruskal-Wallis test, with Dunn's test for multiple comparisons. Human interobserver and intraobserver variability were assessed using intraclass correlation coefficient (ICC) for volume outputs, and Dice coefficient for voxel-level region overlap. The latter comparisons involved a mean total of 2250 CT slices (mean 225 slices, 10 patients). For interobserver data, 10 randomly selected DIAPHRAGM study scans were annotated by ST. For intraobserver data, ACK re-annotated 10 randomly selected training scans, no sooner than 3 weeks after the first annotation. Differences in OS were compared using Kaplan-Meier methodology. Statistical tests were performed in SPSS (V.24.0, Chicago, USA), GraphPad (V.9.1.0, San Diego, USA) and MATLAB (V.9.10, MathWorks, Natick, USA).

RESULTS

Study population

The training set comprised 123 CT datasets from 108 patients; 23 were drawn from DIAPHRAGM, 85 from PRISM; 80/123

Table 1 Demographics and clinical findings in patients with malignant pleural mesothelioma, split into training (n=80 subject to full annotation) and external validation (n=30) sets

	Training set n=80	External validation set n=30
Age at diagnosis	70 (8%)	69 (7%)
Male gender	71 (89%)	22 (73%)
Histological subtype		
Epithelioid	62 (78%)	24 (80%)
Non-epithelioid	11 (14%)	6 (20%)
Not available	7 (9%)	
Disease stage		
I	28 (35%)	12 (40%)
II	2 (3%)	1 (3%)
III	11 (14%)	2 (7%)
IV	6 (8%)	4 (13%)
Not available	33 (41%)	11 (37%)
ECOG performance status		
0	20 (25%)	6 (20%)
1	47 (59%)	15 (50%)
2	11 (14%)	3 (10%)
Not available	2 (3%)	6 (20%)

Values are n (%; NB: % may exceed 100 due to rounding).
ECOG, Eastern Cooperative Oncology Group.

CT datasets from 80 individual patients were fully annotated. The external validation set included 30 individual patients, each with a prechemotherapy and postchemotherapy scan (60 CT datasets). As summarised in table 1, the clinical characteristics of the fully annotated training and external validation sets were well balanced.

This included stage distribution overall, however the fully annotated training set contained some stage heterogeneity, with 16/23 (70%) DIAPHRAGM cases being stage I, compared with 10/57 (18%) PRISM cases. All cases in the external validation set were drawn from PRISM. All patients received cisplatin/carboplatin-pemetrexed chemotherapy (median number of cycles 4 (3.75–4)). The median interval between the last dose and the postchemotherapy CT scan was 22 (10–62) days. In 4/30 cases this interval exceeded 100 days, these cases were excluded from the response classification survival analyses. This was an a priori but arbitrary threshold. It was selected to allow only inclusion of cases in which imaging was acquired with reasonable proximity to chemotherapy conclusion. We defined this as 100 days, approximating to around 3 months, judging this to provide sufficient tolerance for a study reliant on routinely acquired, non-protocolised imaging.

Training and internal validation

Initial CNN performance

Human and AI volumes (see figure 1B for examples) were strongly correlated (training set r=0.847, p<0.0001, see figure 2A). AI volumes were significantly larger with a mean bias (AI-human volume) of +142 cm^3 (p<0.0001, 95% limits −226 to 511 cm^3) (see figure 2B).

Reproducibility

Human interobserver agreement was moderate (ICC 0.732, p=0.001), with a mean difference of 65.8 (70.9) cm^3. Human intraobserver agreement was excellent (ICC 0.997, p=<0.0001), with a mean difference of 29.6 (19.1) cm^3. There is, by definition, no AI intraobserver variation. AI interobserver variation would involve comparison with a different algorithm.

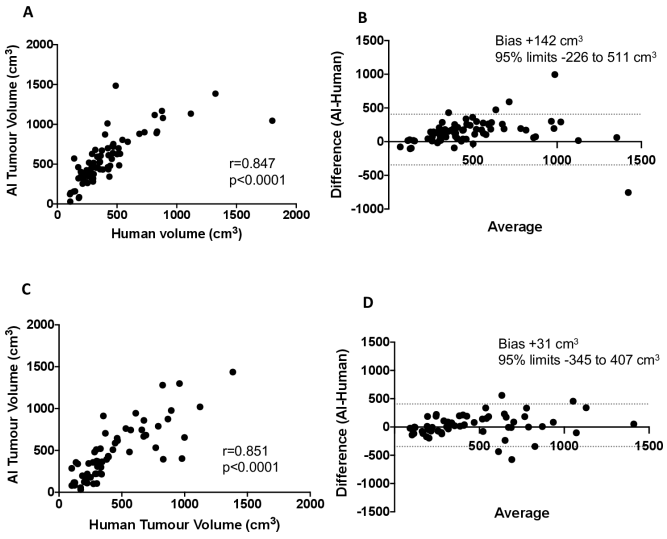


Figure 2 Panel A summarises Spearman's correlation between human and artificial intelligence (AI)-derived tumour volume measurements in 80 subjects (80 scans) in the training set. All scans were acquired prior to chemotherapy. Panel B shows a Bland-Altman plot comparing human versus AI-derived tumour volume measurements based on the same scans. Panel C summarises Spearman's correlation between human and AI-derived tumour volume measurements in 30 subjects (60 scans) in the validation set. Each subject had CT scans acquired prechemotherapy and following at least two cycles of chemotherapy. Panel D shows a Bland-Altman plot comparing human versus AI-derived tumour volume measurements based on the same scans.

Fidelity to reference human annotations by region overlap

The mean Dice overlap between reference human ground truth annotation and AI segmentation was 0.54 (0.08) and 0.54 (0.16), respectively, for the two sets of 10 scans used for interobserver and intraobserver analyses. In direct comparisons of these CT datasets, this was superior to agreement with the second human reader (ST, mean DICE 0.36 (0.1), p=0.002) but inferior to agreement with re-annotation by the reference human (AK, mean DICE 0.61 (0.09), p=0.014).

External validation

Human versus AI volumes

Human and AI volumes were strongly correlated (validation set r=0.851, p<0.0001) (see figure 2C). Bland-Altman plots revealed a mean bias of +31 cm^3, which was not significantly different to zero (p=0.182) and 95% limits of −345 to +407 cm^3 (see figure 2D). Similar results were found when prechemotherapy and postchemotherapy scans were analysed separately (see online supplemental figures 1 and 2). Analysis of the four datapoints outside the 95% limits revealed that the two undersegmented scans both reflected failures to include fissural tumour (see figure 3A) while the two oversegmentation scans reflected erroneous inclusion of atelectatic lung overlying the hemidiaphragm (see figure 3B) and contralateral segmentation in a patient with contralateral benign pleural thickening (see figure 3C).

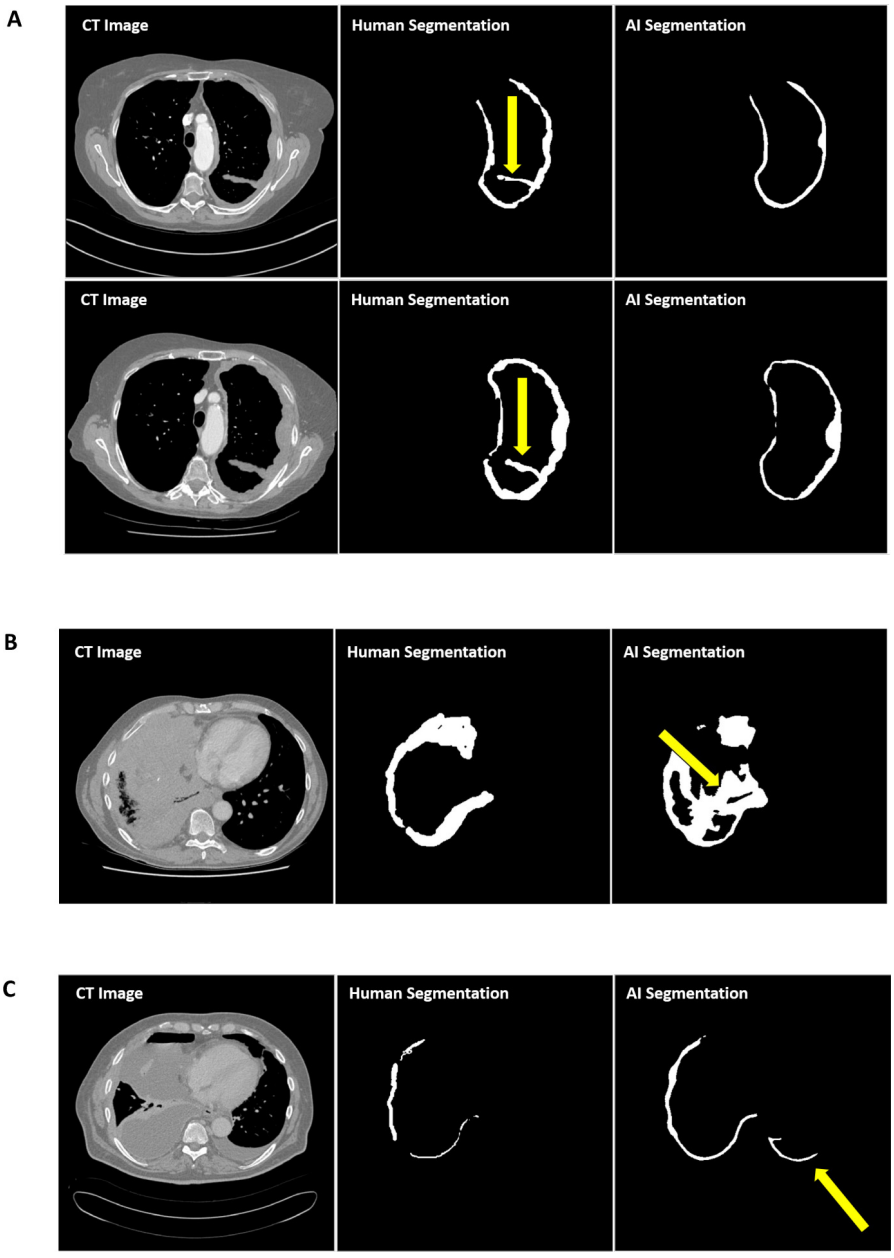


Figure 3 In 4/60 validation set datasets, artificial intelligence (AI)-human differences exceeded 95% Bland-Altman limits (see figure 2D), with AI undersegmentation in 2/60 and oversegmentation in 2/60. CT images, human annotations and AI segmentations were examined in these cases; representative images are presented (CT images on the left, human annotation in the middle, AI segmentations on the right of all panels). Panel A shows prechemotherapy and postchemotherapy images (upper and lower rows, respectively) from the same patient in whom the AI undersegmented compared with the human at both timepoints, reflecting failure to include fissural tumour (arrow). Panel B is from the first oversegmented case, in which the AI erroneously included an area of atelectatic lung overlying the right hemidiaphragm (arrow), which was not included by the human reader. Panel C is from the second oversegmented case and shows erroneous inclusion of contralateral benign pleural thickening by the AI (arrow), but not by the human reader.

Volumetric change following chemotherapy

There were trends towards lower human and AI tumour volumes following chemotherapy, but neither change reached statistical significance (human: 366 cm^3 (244 to 656) vs 328 cm^3 (225 to 663), p=0.196; AI: 427 cm^3 (220 to 682) vs 371 cm^3 (122 to 689), p=0.081). Human and AI volume changes were closely correlated (r=0.611, p=0.0003) (see figure 4A), with a mean bias (AI minus human) of +2.1% that was not significantly different to zero (p=0.425), 95% limits of agreement −59.6% to 55.5% (see figure 4B). When human and AI volume changes were codified into PR, SD and PD, there was agreement in 20/30 (67%) cases, kappa=0.439 (0.178 to 0.700) (see figure 4C). When response was simplified to non-PD versus PD, the number of agreements increased to 26/30 (87%), kappa=0.586 (0.227 to 0.945) (see figure 4D).

mRECIST versus AI volumetric response

The number of PR/SD/PD and non-PD/PD cases by mRECIST, AI and human volumetrics are summarised in table 2 (overleaf). In 16/30 (55%) cases, there was

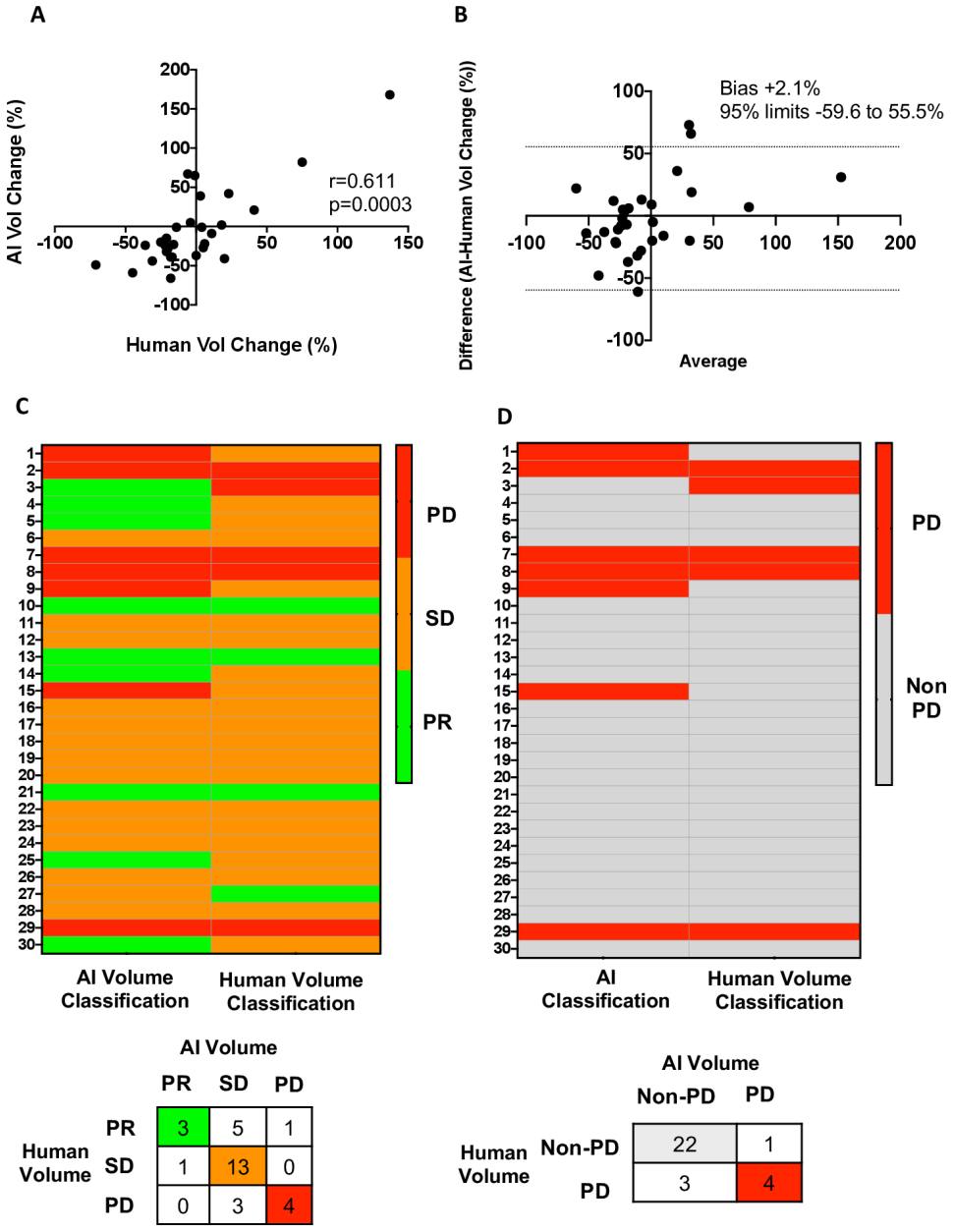


Figure 4 Panel A summarises Spearman's correlation between human and artificial intelligence (AI) volume change following chemotherapy in 30 subjects (60 scans) in the validation set. Panel B shows a Bland-Altman plot comparing human and AI volume change based on the same scans. In panel C, human and AI volumetric responses for each patient (rows 1–30) are compared (as partial response (PR), stable disease (SD) and progressive disease (PD)), accompanied by a confusion matrix summarising agreement, which was present in 20/30 (67%) cases, kappa=0.439 (0.178 to 0.700). Similar results are summarised in panel D, in which responses have been simplified to non-PD (PR and SD) and PD, generating 26/30 (87%) agreements (kappa=0.586 (0.227 to 0.945)).

agreement between mRECIST and AI response classification, kappa=0.284 (0.026 to 0.543), constituting fair agreement (see figure 5A). When response was simplified to non-PD versus PD, the number of agreements increased to 20/30 (67%), kappa=0.223 (−0.128 to 0.574) (see figure 5B).

Using the human ground truth volumes as a reference standard, there was no significant difference in volume change between mRECIST PR, SD and PR groups (median volume change (%) −52 cm^3, −21 cm^3, −18 cm^3, p=0.072, see figure 5C). However, volume change did differ between AI classified PR, SD and PR (median volume change (%) −18 cm^3, −15 cm^3, +23 cm^3, p=0.009, see figure 5D).

Survival analyses

Median OS in the validation cohort was 377 days (median follow-up 1729 days (4.7 years)). Higher prechemotherapy tumour volume was a strong predictor of OS, when the validation cohort was dichotomised around the median human or AI volume (see online supplemental figure 3a,b), respectively. There were non-significant trends towards shorter OS in cases with PD versus non-PD as defined by mRECIST (293 vs 399 days, HR 1.78 (0.71 to 4.46), p=0.149, n=26); human volumes (271 vs 375 days, HR 1.61 (0.51 to 5.07), p=0.326, n=26) and AI volumes (271 vs 375 days, HR 1.58 (0.37 to 6.75), p=0.326, n=26) (see online supplemental figure 3c-e).

Table 2 Primary tumour volume was measured by manual human segmentation and a fully automated AI algorithm on CT scans before and after palliative chemotherapy in 30 patients with MPM, allowing calculation of volumetric responsemRECIST criteria were also used by an expert MPM radiologist to score response on the same scans

	Human volume	AI volume	Human mRECIST
PR	4/30 (13%)	9/30 (30%)	6/30 (20%)
SD	21/30 (70%)	14/30 (47%)	13/30 (43%)
PD	5/30 (17%)	7/30 (23%)	11/30 (37%)
Non-PD	25/30 (83%)	23/30 (77%)	19/30 (63%)
PD	5/30 (17%)	7/30 (23%)	11/30 (37%)

Values are n (%) unless stated.
MPM, malignant pleural mesothelioma; mRECIST, modified Response Evaluation Criteria In Solid Tumours; PD, progressive disease; PR, partial response; SD, stable disease.

DISCUSSION

In this study, we trained an automated deep learning CNN capable of accurately segmenting primary tumour volume in MPM, without any human input. In an independent validation set (60 CT datasets), the mean difference between AI and human volumes was not significantly different to zero (mean bias +31 cm^3 (p=0.182), 95% limits −345 to +407 cm^3, see figure 2C and D). Segmentation errors exceeding 95% limits were observed in 4/60 cases, reflecting important morphological features of MPM (fissural tumour, contralateral pleural thickening and adjacent lung atelectasis), suggesting CNN performance can be further improved by enriching future training sets for these features. In the training set, the positive bias observed for AI volumes may have reflected the stage heterogeneity of the training set, that is, training on predominantly later stage PRISM patients followed by initial internal validation which included early stage DIAPHRAGM patients. Inclusion of lower volume disease patients in future training steps may therefore also enhance CNN performance.

In the current study, higher prechemotherapy tumour volume by human reader was strongly associated with survival, replicating findings from multiple previous volumetry studies.[13 24] However, to our knowledge, ours is the first report of an entirely independent AI-generated volume generating the same prognostic information. The chemotherapy response rate reported here (20% PR by mRECIST) is concordant with previous reports, including a large, expanded access programme (n=1704, 21.7%–26.3% PR)[25] and a meta-analysis of nine chemotherapy trials (n=526, 11% PR).[26] In these studies, SD was the most frequent radiological response (51.4%–54.1%[25] and 75%[26]), reflecting the low efficacy of this treatment, which has been supplanted by platinum-pemetrexed bevacizumab[27] and combination immune checkpoint blockade[2] in recent phase III trials.

We observed only fair agreement between mRECIST and AI volumetric response classifications (kappa=0.284, see figure 5A). Agreement was better between AI and human volume responses (kappa=0.439, increasing to 0.586 when simplified to PD versus non-PD) with no significant difference between these values on Bland-Altman analysis (see figure 2D).[13 24] The relatively poor agreement between mRECIST and AI classification is therefore most likely to reflect poor calibration of the volumetric response criteria chosen, which may also explain the dislocation between volume change and subsequent survival observed (see online supplemental figure 3). We defined PR and PD as −30% and +20% changes in tumour volume based on previous mathematical modelling reported by Oxnard *et al*,[23] as reported in the 'Methods' section. However, like previous volumetric response studies[23 28 29] using different response thresholds, we observed significantly more SD and less PR using volume-defined response than with mRECIST. Frauenfelder *et al* used alternative 'volume equivalent' criteria (−65.7%, +72.8%)[28] proposed by Oxnard *et al*, while acknowledging their inherent limitations when applied to a non-spherical tumour. Based on median scores from three readers, this study reported volumetry-defined SD in 20/30 (67%) cases and a PR in only 2/30 (7%), which is lower than expected based on previous chemotherapy studies,[25 26] and significantly different to the mRECIST-defined PR and SD rates in the same study (7/30 (23%) and 16/30 (53%), respectively). This may reflect the broad SD category used but is concordant with our own and other studies,[23 28 29] which used different criteria. This poor calibration of volumetry response criteria currently offsets the potentially exquisite sensitivity of volumetric measures to minimal change during therapy.[11] Future large studies are therefore essential to determine the optimal cut points for volumetric PR and PD in MPM. Validated cut points from future studies could also help define what constitutes a clinically important difference (or bias) between human and AI volumetry measurements.

Previous studies report increasingly capable computer-aided systems for volumetric segmentation in MPM, including methods based on the Cavalieri stereological principle,[29] semi-automated segmentation with linear interpolation,[28] a random walk-based method[30] and a previous deep learning CNN that required the user to define the laterality of the disease.[31] The CNN developed here is, to our knowledge, the first fully automated and validated system that requires no user input. However, this highly evolved technical system remains constrained by the clarity of the imaging acquired. This will be of critical importance if CNNs for MPM segmentation are to be made ready for clinical practice, wherein variable image quality may lead to inconsistent response calls. This is reflected in previous studies that report highly discordant volumetric MPM measurements when CT tissue contrast is poor,[13 14] emphasising the need for highly protocolised acquisition and further large-scale validation.

Clinical implications

User independence and the high fidelity to human ground truth make the AI tool reported here a potentially important clinical development. Following further optimisation, including validation of optimal response thresholds, it could improve clinical decision making, by enabling practical deployment of volumetric tumour measurements for the first time, allowing earlier cessation of toxic treatment, and enhancing clinical trials by increasing statistical power and reducing costs.[11] AI-generated volumes could also obviate the current need for a minimal measurable disease threshold, since the tool reported here was able to accurately segment tumour volumes as low as ~100 cm^3.

Strengths and limitations

The current study is the largest report of MPM volumetry, comprising 183 datasets, of which 80 fully annotated datasets were used for internal validation, and 60 were used for external validation. However, this sample size is modest when compared with many deep learning projects. Nevertheless, extremely detailed ground truth was used for training, validation and comparisons between volumetry methods and readers, reducing the number of

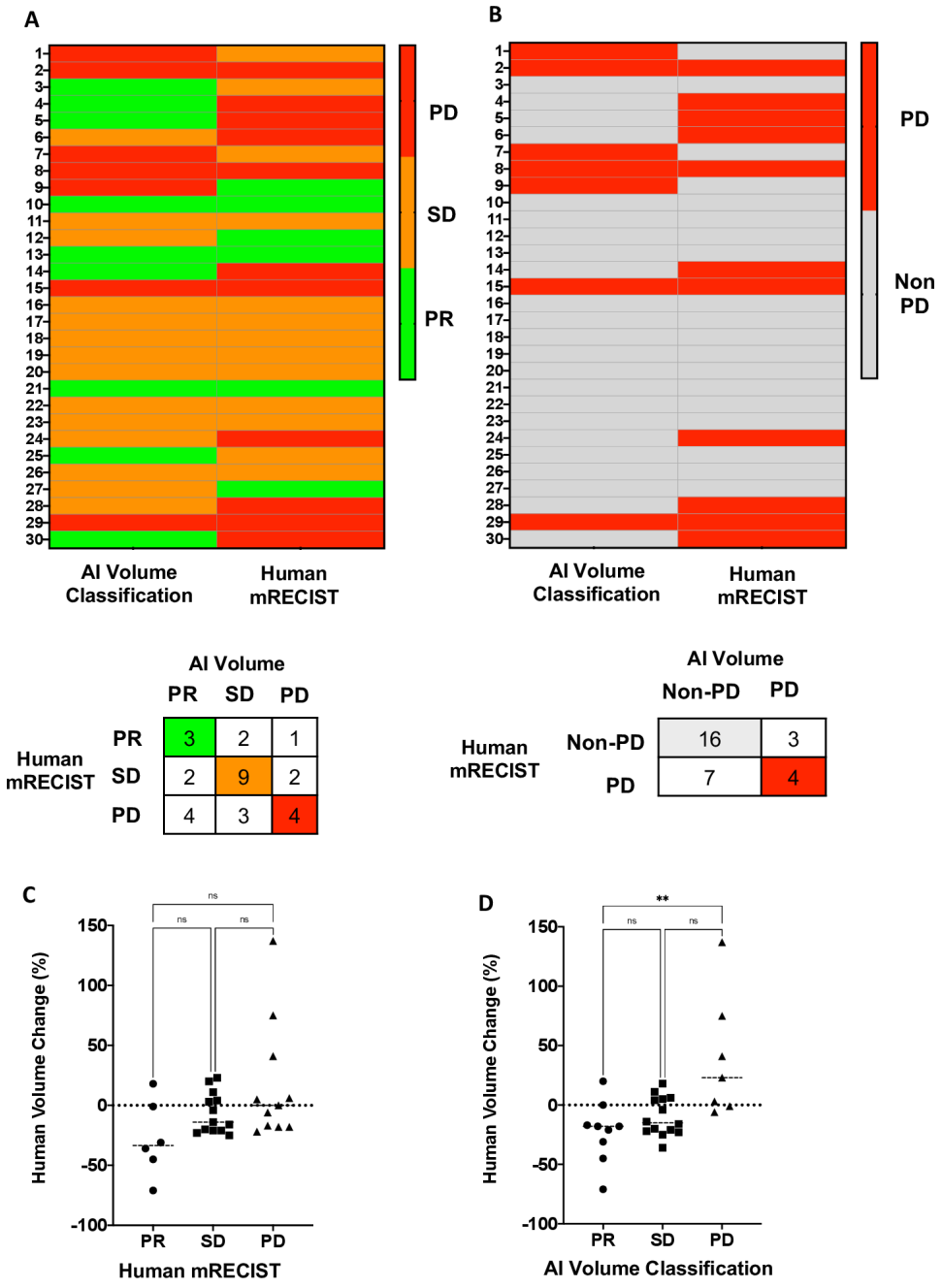


Figure 5 In panel A, modified Response Evaluation Criteria In Solid Tumours (mRECIST) and artificial intelligence (AI) volumetric responses for each patient (rows 1–30) are compared (as partial response (PR), stable disease (SD) and progressive disease (PD)), accompanied by a confusion matrix summarising agreement, which was present in 16/30 (55%) cases, kappa=0.284 (0.026 to 0.543). Similar results are summarised in panel B, in which responses have been simplified to non-PD (PR and SD) and PD, generating 20/30 (67%) agreements, kappa=0.223 (–0.128 to 0.574). Panels C and D use the manually annotated human volumes as a reference standard to compare volume changes in mRECIST (panel C) and AI volume (panel B)-defined PR, SD and PR groups. By mRECIST, the median volume changes (%) were not significantly different (–52 cm^3, –21 cm^3, –18 cm^3, p=0.072). By AI volume, there was significant volume difference between response classes (median volume changes (%) –18 cm^3, –15 cm^3, +23 cm^3, respectively, p=0.009), driven by a significant difference between PR and PD cases (**p=0.008). ns, not significant.

individual patients needed. The selection of 10 cases for inter-reader comparisons was arbitrary but followed precedents set in recent similar publications, including Brahim *et al*,[32] in which an identical number were evaluated and Sensakovic *et al*[33] and Gudmundsson *et al*[34] in which a larger number of patients, but a significantly smaller number of CT sections were interrogated. Sensakovic *et al* compared a total of 155 CT slices between readers (31 patients, 5 CT slices each), while Gudmundsson *et al* compared a total of 69 CT slices from 27 patients. In the current study, we computed a voxel-wise metric, the Dice co-efficient and compared this between readers over a mean total of 2250 CT slices (10 patients, mean of 225 CT slices each), far exceeding the comparisons made in previous studies. Unlike previous studies which used selected CT slices only, comparisons in the current study also benefit from fully volumetric datasets, encompassing a wide variety of tumour appearances and features. The diversity of the imaging data (centres, scanner vendor and model) is an additional strength that reduces the chance of overfitting.

CONCLUSIONS

We have developed and validated the first fully automated deep learning CNN for the volumetric assessment of MPM. Volumetric classification of response requires further calibration in large-scale studies. Given the complexity of MPM tumour morphology, these data represent a strong proof-of-principle for development of similar tools for other cancers.

Twitter Gordon W Cowell @GWCowell and Kevin G Blyth @kevingblyth

Acknowledgements Vismantas Dilys and Jan P Siebert are acknowledged for input to development of the CNN. DIAPHRAGM and PRiSM study patients and study teams are acknowledged for their support in collection of the original study data.

Contributors ACK: data curation, formal analysis, investigation, methodology, validation, writing—review and editing. OA: data curation, formal analysis, investigation, methodology, validation, visualisation, writing—review and editing. GWC: data curation, formal analysis, investigation, software, validation, visualisation, writing—review and editing. AJW: conceptualisation, funding acquisition, project administration, resources, writing—review and editing. JPV: methodology, project administration, resources, software, supervision, writing—review and editing. ME: data curation, formal analysis, investigation, methodology, validation, writing—review and editing. ST: data curation, formal analysis, investigation, methodology, validation, writing—review and editing. KAG: conceptualisation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualisation, writing—review and editing. KGB: conceptualisation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualisation, writing—original draft, writing—review and editing, guarantor of overall content.

Funding Scottish Funding Council via Cancer Innovation Challenge 2018; British Lung Foundation via Project Grant MPG16-7; Chief Scientist Office via Project Grant ETM/285.

Competing interests ACK, GWC, ME, ST and KGB have no conflicts of interest to declare. OA, AJW, JV and KAG are employees of Canon Medical Research Europe.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by West of Scotland Research Ethics Committee (REC) via NHSGCC Safe Haven (Ref GSH/18/ON/001). Given the retrospective nature of the study, most individuals involved were unfortunately deceased. The REC approval covered use of anonymised unconsented data.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data may be applied for through the PREDICT-Meso International Accelerator Network, via the corresponding author.

REFERENCES

1 Vogelzang NJ, Rusthoven JJ, Symanowski J, *et al*. Phase III study of pemetrexed in combination with cisplatin versus cisplatin alone in patients with malignant pleural mesothelioma. *Journal of Oncol* 2003;21:2636–44.
2 Baas P, Scherpereel A, Nowak AK, *et al*. First-Line nivolumab plus ipilimumab in unresectable malignant pleural mesothelioma (CheckMate 743): a multicentre, randomised, open-label, phase 3 trial. *Lancet* 2021;397:375–86.
3 Fennell DA, Kirkpatrick E, Cozens K, *et al*. Confirm: a double-blind, placebo-controlled phase III clinical trial investigating the effect of nivolumab in patients with relapsed mesothelioma: study protocol for a randomised controlled trial. *Trials* 2018;19:233.
4 van KRJ, JGJV A, de BH. Inadequacy of the RECIST criteria for response evaluation in patients with malignant pleural mesothelioma. Poster Abstracts of the 13th Annual British Thoracic Oncology Group Conference 2015; 28-30 January 2015, Dublin, Ireland, 2004:63–9.
5 Byrne MJ, Nowak AK. Modified RECIST criteria for assessment of response in malignant pleural mesothelioma. *Annals of Oncology* 2004;15:257–60.
6 Armato SG, Oxnard GR, MacMahon H, *et al*. Measurement of mesothelioma on thoracic CT scans: a comparison of manual and computer-assisted techniques. *Med Phys* 2004;31:1105–15.
7 Armato SG, Nowak AK, Francis RJ, *et al*. Observer variability in mesothelioma tumor thickness measurements: defining minimally measurable lesions. *J Thorac Oncol* 2014;9:1187–94.
8 Mozley PD, Bendtsen C, Zhao B, *et al*. Measurement of tumor volumes improves RECIST-Based response assessments in advanced lung cancer. *Transl Oncol* 2012;5:19–25.
9 Spira D, Sökler M, Vogel W, *et al*. Volume and attenuation computed tomography measurements for interim evaluation of Hodgkin and follicular lymphoma as an additional surrogate parameter for more confident response monitoring: a pilot study. *Cancer Imaging* 2011;11:155–62.
10 Mueller S, Wichmann G, Dornheim L, *et al*. Different approaches to volume assessment of lymph nodes in computer tomography scans of head and neck squamous cell carcinoma in comparison with a real gold standard. *ANZ J Surg* 2012;82:737–41.
11 Goldmacher GV, Conklin J. The use of tumour volumetrics to assess response to therapy in anticancer clinical trials. *Br J Clin Pharmacol* 2012;73:846–54.
12 Tsim S, Cowell GW, Kidd A, *et al*. A comparison between MRI and CT in the assessment of primary tumour volume in mesothelioma. *Lung Cancer* 2020;150:12–20.
13 Rusch VW, Gill R, Mitchell A, *et al*. A multicenter study of volumetric computed tomography for staging malignant pleural mesothelioma. *Ann Thorac Surg* 2016;102:1059–66.
14 Gill RR, Naidich DP, Mitchell A, *et al*. North American multicenter volumetric CT study for clinical staging of malignant pleural mesothelioma: feasibility and logistics of setting up a quantitative imaging study. *J Thorac Oncol* 2016;11:1335–44.
15 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017;60:84–90.
16 Ogrinc G, Davies L, Goodman D, *et al*. SQUIRE 2.0 (*Standards for QUality Improvement Reporting Excellence)* : revised publication guidelines from a detailed consensus process: Table 1. *BMJ Qual Saf* 2016;25:986–92.
17 Tsim S, Alexander L, Kelly C, *et al*. Serum proteomics and plasma fibulin-3 in differentiation of mesothelioma from asbestos-exposed controls and patients with other pleural diseases. *J Thorac Oncol* 2021;16:1705–17.
18 Blyth KG, Kidd AC, Winter A, *et al*. An update regarding the prediction of resistance to chemotherapy using somatic copy number variation in mesothelioma (PriSM) study. *Lung Cancer* 2018;115:S26–7.
19 Stewart D, Waller D, Edwards J, *et al*. Is there a role for pre-operative contrast-enhanced magnetic resonance imaging for radical surgery in malignant pleural mesothelioma? *Eur J Cardiothorac Surg* 2003;24:1019–24.
20 Plathow C, Staab A, Schmaehl A, *et al*. Computed tomography, positron emission tomography, positron emission tomography/computed tomography, and magnetic resonance imaging for staging of limited pleural mesothelioma: initial results. *Invest Radiol* 2008;43:737–44.
21 Ronneberger O, Fischer P. Brox T. U-Net: Convolutional networks for biomedical image segmentation. *Arxiv* 2015.
22 Anderson O, Kidd A, Goatman K. Fully automated volumetric measurement of malignant pleural mesothelioma from computed tomography images by deep learning: preliminary results of an internal validation 2020:64–73.
23 Oxnard GR, Armato SG, Kindler HL. Modeling of mesothelioma growth demonstrates weaknesses of current response criteria. Poster Abstracts of the 13th Annual British Thoracic Oncology Group Conference 2015; 28-30 January 2015, Dublin, Ireland, 2006:141–8.
24 Pass HI, Temeck BK, Kranda K, *et al*. Preoperative tumor volume is associated with outcome in malignant pleural mesothelioma. *J Thorac Cardiovasc Surg* 1998;115:310–8.
25 Santoro A, O'Brien ME, Stahel RA, *et al*. Pemetrexed plus cisplatin or pemetrexed plus carboplatin for chemonaïve patients with malignant pleural mesothelioma: results of the International expanded access program. *J Thorac Oncol* 2008;3:756–63.
26 Blayney JK, Ceresoli GL, Castagneto B, *et al*. Response to chemotherapy is predictive in relation to longer overall survival in an individual patient combined-analysis with pleural mesothelioma. *Eur J Cancer* 2012;48:2983–92.
27 Zalcman G, Mazieres J, Margery J, *et al*. Bevacizumab for newly diagnosed pleural mesothelioma in the mesothelioma Avastin cisplatin pemetrexed study (maps): a randomised, controlled, open-label, phase 3 trial. *The Lancet* 2016;387:1405–14.
28 Frauenfelder T, Tutic M, Weder W, *et al*. Volumetry: an alternative to assess therapy response for malignant pleural mesothelioma? *Eur Respir J* 2011;38:162–8.
29 Ak G, Metintas M, Metintas S, *et al*. Three-Dimensional evaluation of chemotherapy response in malignant pleural mesothelioma. *Eur J Radiol* 2010;74:130–5.
30 Chen M, Helm E, Joshi N, *et al*. Computer-Aided volumetric assessment of malignant pleural mesothelioma on CT using a random walk-based method. *Int J Comput Assist Radiol Surg* 2017;12:529–38.
31 Gudmundsson E, Straus CM, Armato SG. Deep convolutional neural networks for the automated segmentation of malignant pleural mesothelioma on computed tomography scans. *J Med Imaging* 2018;5:1–11.
32 Brahim W, Mestiri M, Betrouni N, *et al*. Malignant pleural mesothelioma segmentation for photodynamic therapy planning. *Comput Med Imaging Graph* 2018;65:79–92.
33 Sensakovic WF, Armato SG, Straus C, *et al*. Computerized segmentation and measurement of malignant pleural mesothelioma. *Med Phys* 2011;38:238–44.
34 Gudmundsson E, Straus CM, Li F, *et al*. Deep learning-based segmentation of malignant pleural mesothelioma tumor on computed tomography scans: application to scans demonstrating pleural effusion. *J Med Imaging* 2020;7:1.

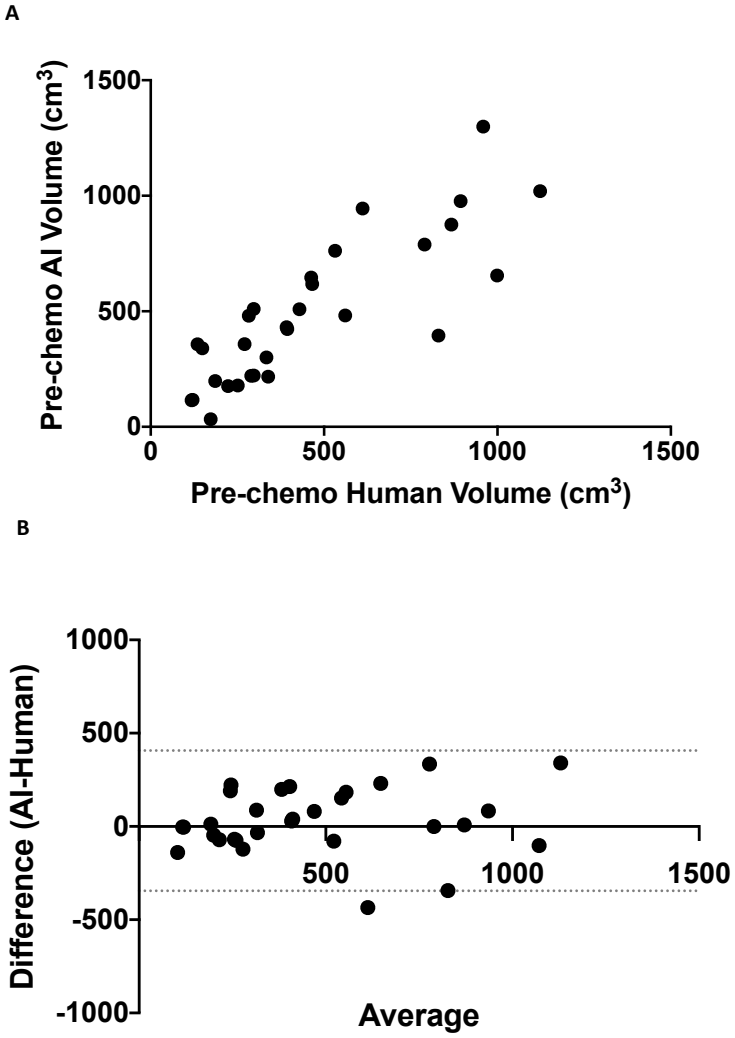
FULLY AUTOMATED VOLUMETRIC MEASUREMENT OF MALIGNANT PLEURAL MESOTHELIOMA BY DEEP LEARNING AI: VALIDATION AND COMPARISON WITH MODIFIED RECIST RESPONSE CRITERIA

ONLINE ONLY SUPPLEMENT

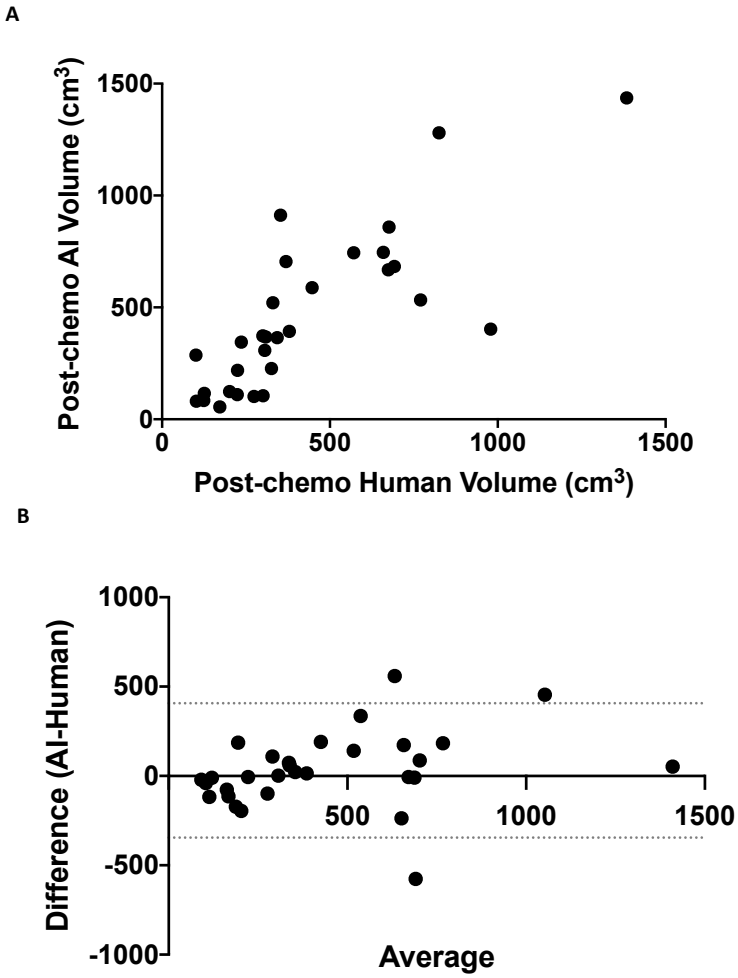
Contents

Supplementary Figure 1

Median pre-chemotherapy human and AI volumes (n=30) were not significantly different (366 cm^3 [244 -656] v 427 cm^3 [220-682], p=0.67). Panel A shows correlation between human and AI volumes (r=0.86, p<0.0001)). Panel B shows Bland-Altman agreement, with a mean bias of +29 cm^3 and 95% limits (-312.9 to 371.1 cm^3).



Supplementary Figure 2
Median post-chemotherapy human and AI- volumes (n=30) were not significantly different (328 cm^3 [225-63] v 371cm^3 [122-689], p=0.84. Panel A shows correlation between human and AI volumes (r=0.86, p<0.0001). Panel B shows Bland-Altman agreement, with a mean bias of +32 cm^3 and 95% limits of agreement (-381 to 445 cm^3).



Supplementary Figure 3

Overall survival (days) was calculated for cases in the validation set from the date of pre-chemotherapy CT scan to death from any cause. Survival analysis was by Kaplan-Meier methodology. Panels A and B report the statistically significant association between higher baseline (pre-chemotherapy) tumour volume and OS, dichotomised around the median volume measured by human (Panel A) and AI (Panel B) segmentation. Panels C-D report non-significant trends towards shorter OS in cases with PD v non-PD defined by human volume criteria (271 v 375 days, n=26), AI volume criteria (271 v 375 days, n=26) and mRECIST criteria (293 v 399 days, n=26), respectively.