Journal of Digital Imaging (2021) 34:541–553
https://doi.org/10.1007/s10278-021-00460-3

ORIGINAL PAPER

Novel Autosegmentation Spatial Similarity Metrics Capture the Time Required to Correct Segmentations Better Than Traditional Metrics in a Thoracic Cavity Segmentation Workflow

Kendall J. Kiser[1] · Arko Barman[2] · Sonja Stieb[3] · Clifton D. Fuller[3] · Luca Giancardo[2]

Received: 19 May 2020 / Revised: 28 March 2021 / Accepted: 4 May 2021 / Published online: 23 May 2021

Abstract
Automated segmentation templates can save clinicians time compared to de novo segmentation but may still take substantial time to review and correct. It has not been thoroughly investigated which automated segmentation-corrected segmentation similarity metrics best predict clinician correction time. Bilateral thoracic cavity volumes in 329 CT scans were segmented by a UNet-inspired deep learning segmentation tool and subsequently corrected by a fourth-year medical student. Eight spatial similarity metrics were calculated between the automated and corrected segmentations and associated with correction times using Spearman's rank correlation coefficients. Nine clinical variables were also associated with metrics and correction times using Spearman's rank correlation coefficients or Mann–Whitney U tests. The added path length, false negative path length, and surface Dice similarity coefficient correlated better with correction time than traditional metrics, including the popular volumetric Dice similarity coefficient (respectively $\rho = 0.69$, $\rho = 0.65$, $\rho = -0.48$ versus $\rho = -0.25$; correlation p values < 0.001). Clinical variables poorly represented in the autosegmentation tool's training data were often associated with decreased accuracy but not necessarily with prolonged correction time. Metrics used to develop and evaluate autosegmentation tools should correlate with clinical time saved. To our knowledge, this is only the second investigation of which metrics correlate with time saved. Validation of our findings is indicated in other anatomic sites and clinical workflows. Novel spatial similarity metrics may be preferable to traditional metrics for developing and evaluating autosegmentation tools that are intended to save clinicians time.

Keywords Image segmentation · "Computer-assisted image analysis" [MeSH] · "AI artificial intelligence" [MeSH] · "Medical imaging" [MeSH] · "Clinical informatics" [MeSH]

Introduction

The advent of deep learning-based segmentation algorithms is expanding the range of automated segmentation (autosegmentation) use to clinical tasks and research questions that demand previously unattainable accuracy or reliability. Autosegmentation algorithms may soon assist neurologists to localize ischemic cores during a code stroke [1, 2] or anticipate Parkinson's disease onset in an outpatient setting [3]. They may inform 3D-printed implant designs for orthopedists [4, 5] or highlight posterior segment lesions [6–8] for ophthalmologists. They may help neurosurgeons spare microvessels [9], outline catheters for radiation oncologists during MRI-guided brachytherapy [10], or characterize vocal fold mobility for otorhinolaryngologists [11]. Dedicated imaging specialists—radiologists and pathologists—are likely to identify even more autosegmentation uses than clinicians whose primary clinical domain is not imaging. For example, segmenting regions-of-interest is a necessary step prior to extraction of quantitative imaging biomarkers ("radiomics" features) known to harbor information respecting disease prognoses and treatment response probabilities [12]. Radiomics feature computation methods were recently

✉ Kendall J. Kiser
K.j.kiser@wustl.edu

Clifton D. Fuller
cdfuller@mdanderson.org

Luca Giancardo
Luca.Giancardo@uth.tmc.edu

[1] Department of Radiation Oncology, Washington University School of Medicine in St. Louis, St. Louis, MO, USA

[2] Center for Precision Health, UT Health School of Biomedical Informatics, Houston, TX, USA

[3] Department of Radiation Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

standardized [13], overcoming a significant obstacle to clinical implementation. In the future, reviewing and vetting autosegmented regions-of-interest prior to radiomics analyses could become part of routine radiology [14].

Autosegmentations are useful if they obviate the need for a clinician to delineate segmentations de novo, which can be time-consuming [4, 15–18] and inconsistent [19–23] between observers and within the same observer at different time points. Several studies confirm that clinicians can save time from leveraging autosegmentation templates compared to de novo segmentation [15, 24–29], but in many circumstances, the time required for clinicians to review and correct autosegmentations is still substantial. For example, during online adaptive/stereotactic MRI-guided radiotherapy [30], radiation oncologists must carefully correct cancer and normal anatomy autosegmentations while a patient waits immobilized in the treatment device. Cardiologists may spend just as long to review and correct cardiac ventricle autosegmentations as segmenting them de novo [18]. Plastic surgeons can implant facial trauma repair plates faster with autosegmentation-based 3D-printed mandibular templates than without them [31], but autosegmentation review still consumes time in an urgent setting. Whenever autosegmentation algorithms are deployed to save clinical time, the metrics used to assess them should capture an expected time-savings benefit. Algorithm development should be optimized and evaluated by whatever metric or metrics best predict time savings.

Autosegmentations are usually compared with a reference segmentation by spatial similarity metrics that compare (1) volumetric overlap between an autosegmented structure and the same manually segmented structure [4, 6, 10, 11, 16, 22–26, 28, 29, 32–47], such as the volumetric Dice similarity coefficient/index [48] (DSC); or (2) geometric distance between two structures' surfaces [4, 10, 22, 24–26, 28, 32–35, 37, 40, 42, 43, 46, 47], such as the Hausdorff distance [49] (HD); or (3) structure centricity [32, 39], such as differences between centers-of-mass. Critically, these metrics do not necessarily correlate with time savings in clinical practice [40, 50]. To our knowledge, the only investigation of *which* metrics best predict the time clinicians spend correcting autosegmentations was published in 2020 by Vaassen et al. [28].

Vaassen et al. compared automatically generated and manually corrected thoracic structure segmentations in 20 CT cases acquired from patients with non-small-cell lung cancer (NSCLC). They found that the "added path length" (APL; a novel metric they introduced) and the surface Dice similarity coefficient (a novel metric introduced by Nikolov et al. [51]) correlated better with the time it took a clinician to review and correct autosegmentations than other metrics that are popular for autosegmentation evaluation. Here, we corroborate and extend their findings. We also experiment with APL by calculating variations of it, which we term the false negative path length (FNPL) and false negative volume (FNV). We correlate the APL, FNPL, FNV, surface DSC, volumetric DSC, Jaccard index (JI), average surface distance (ASD), and HD metrics calculated between automatically generated and manually corrected thoracic cavity segmentations with time required for correction. We contribute evidence that the surface DSC may be superior to popular volumetric DSC for optimizing autosegmentation algorithms. We also investigate how anatomic and pathologic variables impact autosegmentation correction time. In the process, we have generated a library of 402 expert-vetted left and right thoracic cavity segmentations, as well as 78 pleural effusion segmentations, which we made publicly available [52] through The Cancer Imaging Archive (TCIA). The CT scans on which the segmentations were delineated are likewise publicly available [53] from TCIA.

Materials and Methods

CT Datasets

A collection of four hundred twenty-two CT datasets acquired in Digital Imaging and Communications in Medicine (DICOM) format from patients with NSCLC was downloaded from NSCLC-Radiomics [53], a TCIA data collection, in January 2019. Accompanying clinical data in tabular format and gross tumor volume, segmentations available for a subset of cases were also downloaded. CT scans were converted from DICOM to Neuroimaging Informatics Technology Initiative (NIfTI) format using a free program called "dcm2niix." [54, 55] Four-hundred-two CT datasets were successfully converted and subsequently underwent autosegmentation and manual correction.

Segmentations

We leveraged a publicly available, UNet-inspired deep learning autosegmentation algorithm [56] to segment lungs in the 402 CT datasets described above. This algorithm was trained to segment bilateral lungs (under a single label) with approximately 200 CTs acquired in patients who—importantly—did not have lung cancer. A fourth-year medical student reviewed and corrected the autosegmentations using an image segmentation software called ITK-SNAP v 3.6. [57] The corrections included the bilateral thoracic cavity spaces that healthy lung parenchyma normally occupies but in our dataset were occasionally occupied by atelectatic parenchyma, tumor, pleural effusion, or other changes. Because the idea to capture correction time and correlate it with autosegmentation similarity metrics developed after this project had commenced, the medical student recorded the time it took to correct autosegmentations

for only 329 of 402 corrected cases. Specifically, correction times comprised the times required to load autosegmentations, correct them slice-wise with size-adjustable brush and interpolation tools, and save the corrections. Because the autosegmentation algorithm was trained on scans without cancer but deployed on scans with NSCLC, its accuracy varied with the severity of disease-induced anatomic change in each case. For example, cases with massive tumors or pleural effusions were sometimes poorly autosegmented, whereas cases with minimal anatomic changes were autosegmented well. This effectively simulated a range of major and minor manual corrections. Subsequently, the medical student's manually corrected segmentations were vetted and corrected as necessary by a radiation oncologist or a radiologist. The 402 physician-corrected thoracic cavity segmentations—so named to reflect inclusion of primary tumor and pleural pathologies in the thoracic cavity rather than lung parenchyma alone—have been made publicly available [52].

Metrics

Automated and corrected segmentations were compared by the volumetric DSC; the JI; the surface DSC at 0-mm, 4-mm, 8-mm, and 10-mm tolerances; the APL; the FNPL; the FNV; the 100th, 99th, 98th, and 95th percentile HDs; and the ASD. Each metric is illustrated in Fig. 1. The volumetric DSC is twice the overlap between volumes A and B, divided by their sum. A DSC of 1 indicates perfect overlap while 0 indicates no overlap. The JI is a related volumetric measure and is the overlap between volumes A and B divided by their union. The DSC and JI converge at 1 [58]. The surface DSC is calculated by the same formula as the volumetric DSC, but its inputs A and B are the segmentations' surface areas rather than their volumes. To permit small differences between surfaces to go unpunished, Nikolov et al. programmed a tolerance parameter: if points in two surfaces are separated by a distance that is within the tolerance parameter, they are considered part of the intersection of A and B. The APL is the number of pixels in the corrected segmentation surface (edge) that are not in the autosegmentation surface [28]. We experiment with metrics related to the APL that we term the FNPL and the FNV. The FNPL is the APL less the pixels from any edits that shrink the autosegmentation. That is, edits that erase pixels from the autosegmentation volume are excluded. The FNV is the number of pixels in the corrected segmentation volume that are not in the autosegmentation volume. The Python code we developed to calculate the APL, FNPL, and FNV has been made available at GitHub at https://github.com/kkiser1/Autosegmentation-Spatial-Similarity-Metrics. The Hausdorff distance calculates the minimum distance from every point in surface A to every point in surface B, and vice versa; arranges all distances in ascending order; and returns the maximum distance (100th percentile) or another percentile if so specified (e.g., 95th percentile). The ASD calculates the average of the minimum distances from every point in surface A to every point in surface B, and vice versa, and returns the average of the two average distances. All metric calculations were made using custom Python scripts that leveraged common scientific libraries [51, 59–61].

Clinical Variables

To describe clinical variation in the NSCLC-Radiomics CT datasets and study the effects of variation in tumor volume, tumor laterality and location, pleural effusion presence, pleural effusion volume, and thoracic cavity volume on autosegmentation spatial similarity metrics and on manual correction time, we collected these variables for each case. Furthermore, we studied how primary tumor stage, tumor overall stage, and tumor histology associated with accuracy and correction time, but these variables were already collected in the NSCLC-Radiomics collection [53] in a spreadsheet named "NSCLC Radiomics Lung1.clinical-version3-Oct 2019.csv." Left and right thoracic cavity volumes were collected from physician-vetted thoracic cavity segmentations using ITK-SNAP. Tumor volume and laterality were collected by referencing primary gross tumor volume segmentations ("GTV-1") and other tumor volume segmentations available from the NSCLC-Radiomics data collection [53]. Tumor location was classified as central, peripheral, or pan. There is no consensus in radiotherapy literature regarding the definition of centrality [62]; we used a definition based off that provided by the International Association for the Study of Lung Cancer [63]: tumors located within 2 cm of the proximal bronchial tree, spinal cord, heart, great vessels, esophagus, or phrenic nerves and recurrent laryngeal nerves and spanning up to 4 cm from these structures were classified as central. Tumors that were not within 2 cm of any central structure were classified as peripheral. Tumors within the central territory that extended further than 4 cm from central structures were classified as pan. The presence or absence of pleural effusion in each subject was noted by a medical student, and effusions were contoured by the student. Pleural effusion segmentations were reviewed and corrected by a radiologist. Pleural effusion volumes were collected from physician-vetted segmentations using ITK-SNAP.

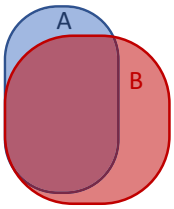
Statistics

We correlated eight autosegmentation spatial similarity metrics with the time expended to correct the autosegmentations. Segmentation correction time; volumetric DSC; surface

Volumetric Dice Similarity Coefficient

$$\frac{2(A \cap B)}{A + B}$$

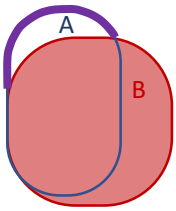
where A and B are volumes



False Negative Path Length (pixels)

$$A - (A \cap B)$$

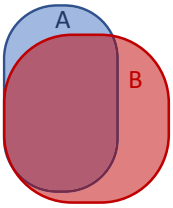
where A is the manually corrected segmentation surface and B is the autosegmentation volume



Jaccard Index

$$\frac{A \cap B}{A \cup B}$$

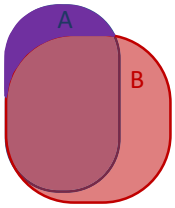
where A and B are volumes



False Negative Volume (pixels)

$$A - (A \cap B)$$

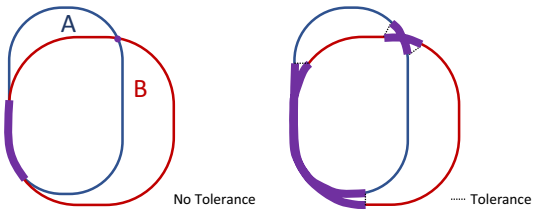
where A is the manually corrected segmentation volume and B is the autosegmentation volume



Surface Dice Similarity Coefficient

$$\frac{2(A \cap B)}{A + B}$$

where A and B are surfaces

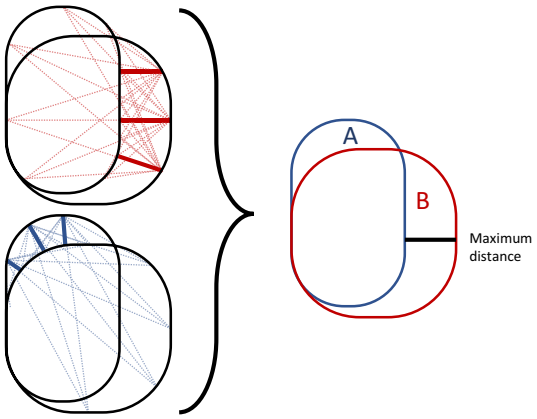


Maximum Hausdorff Distance (mm)

$$\max\big(h(A,B), h(B,A)\big)$$

where

$$h(A,B) = \max_{a \in A} \min_{b \in B} \|a - b\|$$

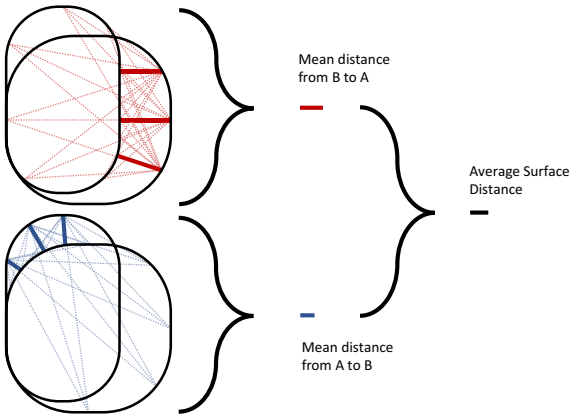


Average Surface Distance (mm)

$$\frac{1}{2}\big(h(A,B) + h(B,A)\big)$$

where

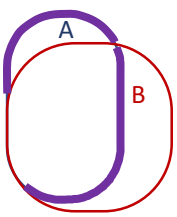
$$h(A,B) = \frac{1}{A}\sum_{\alpha \in A} \min_{b \in B} \|a - b\|$$



Added Path Length (pixels)

$$A - (A \cap B)$$

where A is the manually corrected segmentation surface and B is the autosegmentation surface



◂**Fig. 1** Eight metrics for evaluating spatial similarity between segmentations. Traditional (volumetric DSC, Jaccard index, Hausdorff distance, and average surface distance) or novel (surface DSC, added path length, false negative path length, false negative volume) was used to compare autosegmentations with manually corrected segmentations. The surface DSC calculation permits a tolerance parameter whereby non-intersecting segments of surfaces **A** and **B** that are separated by no more than the parameter distance are considered part of the intersection between **A** and **B**. The Hausdorff distance illustration and equation represent the 100th percentile (maximum) distance but can be adapted to any other percentile distance

DSC, JI, APL, FNPL, FNV, HD; and ASD distributions were assessed for normality using the Shapiro–Wilk's test [64]. The null hypotheses (that distributions were normal) were rejected in each case (p values < 0.001). Additionally, we sought to describe the influence of common disease-induced anatomic changes on the accuracy of the UNet autosegmentation algorithm. The null hypotheses that tumor, thoracic cavity, and pleural effusion volume distributions were normal were likewise each rejected (p values < 0.001). Therefore, non-parametric statistical tests were employed. Pairwise Spearman's rank correlation coefficients [65] described linear associations between spatial similarity metrics and correction times. The pairwise Mann–Whitney U test [66] assessed significant differences between numeric variable distributions stratified by a categorical variable with two categories and followed a Kruskal-Wallis [67] test if the categorical variable had three or more categories and the Kruskal–Wallis result was significant. A significance threshold of $\alpha = 0.05$ was used, and Bonferroni corrections [68] were assessed to account for multiple comparisons as needed. Statistics were computed in Python using the SciPy library [59].

Results

Four-hundred and two thoracic cavity segmentations were automatically generated and corrected manually (Fig. 2). Correction times were recorded in 329 cases. Among these cases, median right and left corrected thoracic cavity volumes were 2220 cm [3] and 1920 cm [3], respectively (Fig. 3a). Tumor overall stage was I in 10% of cases (33/329), II in 21% of cases (69/329), IIIA in 27% of cases (88/329), and IIIB in 42% of cases (139/329). Tumor stage was T1 in 24% of cases (78/329), T2 in 34% of cases (112/329), T3 in 13% of cases (42/329), and T4 in 29% of cases (97/329). Primary lung tumors were in the right hemithorax in 58% of cases (191/329) and in the left hemithorax in 42% of cases (138/329). Tumor locations were classified as central in 24% of cases (80/329), peripheral in 33% of cases (108/329), and pan in 43% of cases (141/329). Median tumor volumes were 29 cm [3], 17 cm [3], 2 cm [3], 4 cm [3], 3 cm [3], and 6 cm [3] for GTV1 through GTV6, respectively (Fig. 3b). Among 298 cases with recorded autosegmentation time and available tumor histology, the histology was squamous cell carcinoma in 40% of cases (120/298), large cell carcinoma in 30% of cases (90/298), adenocarcinoma in 14% of cases (43/298), and not otherwise specified in 15% of cases (45/298). Among 59 cases with recorded autosegmentation correction times and a pleural effusion in at least one hemithorax, median right and left pleural effusion volumes were 53 cm [3] and 51 cm [3], respectively (Fig. 3c).

Anatomic changes caused by disease significantly influenced the autosegmentation algorithm's similarity to manually corrected segmentations, but worse similarity did not always result in longer correction times. Tumor location (central, peripheral, or pan) was associated with similarity by several metrics (e.g., volumetric DSC for central tumors: 0.963, pan tumors: 0.945; $p < 0.001$), but there were no significant differences in correction time between central (median 18.61 min), peripheral (median 19.01 min), or pan (median 18.83 min) tumors ($p = 0.24$). Primary tumor volume correlated moderately with several similarity metrics but not with correction time ($p = 0.15$). Like tumor location, the presence of pleural effusion was associated with significantly worse similarity by all metrics (p values < 0.001), but this did not significantly prolong manual correction times (median with effusion: 19.13 min, without effusion: 18.71 min; $p = 0.18$). Furthermore, pleural effusion volume correlated weakly with volume overlap metrics (i.e. volumetric DSC, JI) but not with correction time ($p = 0.20$). Neither metric nor correction time distributions were significantly different between tumor histologies.

Few clinical variables were significantly associated with correction time. Autosegmentations delineated on CT scans with T4 tumors took marginally but significantly longer to correct (median 20.82 min) than those on CTs with T1 (median 19.0 min), T2 (median 18.13 min), or T3 (median 18.30 min) tumors (p values ≤ 0.01). Interestingly, the only metrics that captured significant differences between cases with T4 tumors and cases with any other T stage tumor were the maximum HD (median T4: 56 mm, median T3: 70 mm; $p = 0.02$), FNV (median T4: 104,991 pixels, median T1: 89,334 pixels; $p = 0.001$), FNPL (median T4: 61,928 pixels, median T2: 56,612 pixels, median T1: 57,489 pixels; p values ≤ 0.006), and APL (median T4: 69,707 pixels, median T2: 61,553 pixels, median T1: 61,788 pixels; p values ≤ 0.002). Autosegmentations on CT scans with overall stage II tumors took significantly less time to correct (median 13.53 min) than those from CTs with stage I (median 19.08 min), stage IIIA (median 18.94 min), or stage IIIB tumors (median 19.83 min) (p values ≤ 0.04). Only the surface DSC at 0 mm tolerance, FNPL, and APL captured a significant difference between these groups (surface DSC median for stage II: 0.77 vs. stage I: 0.70, stage IIIA: 0.70, stage

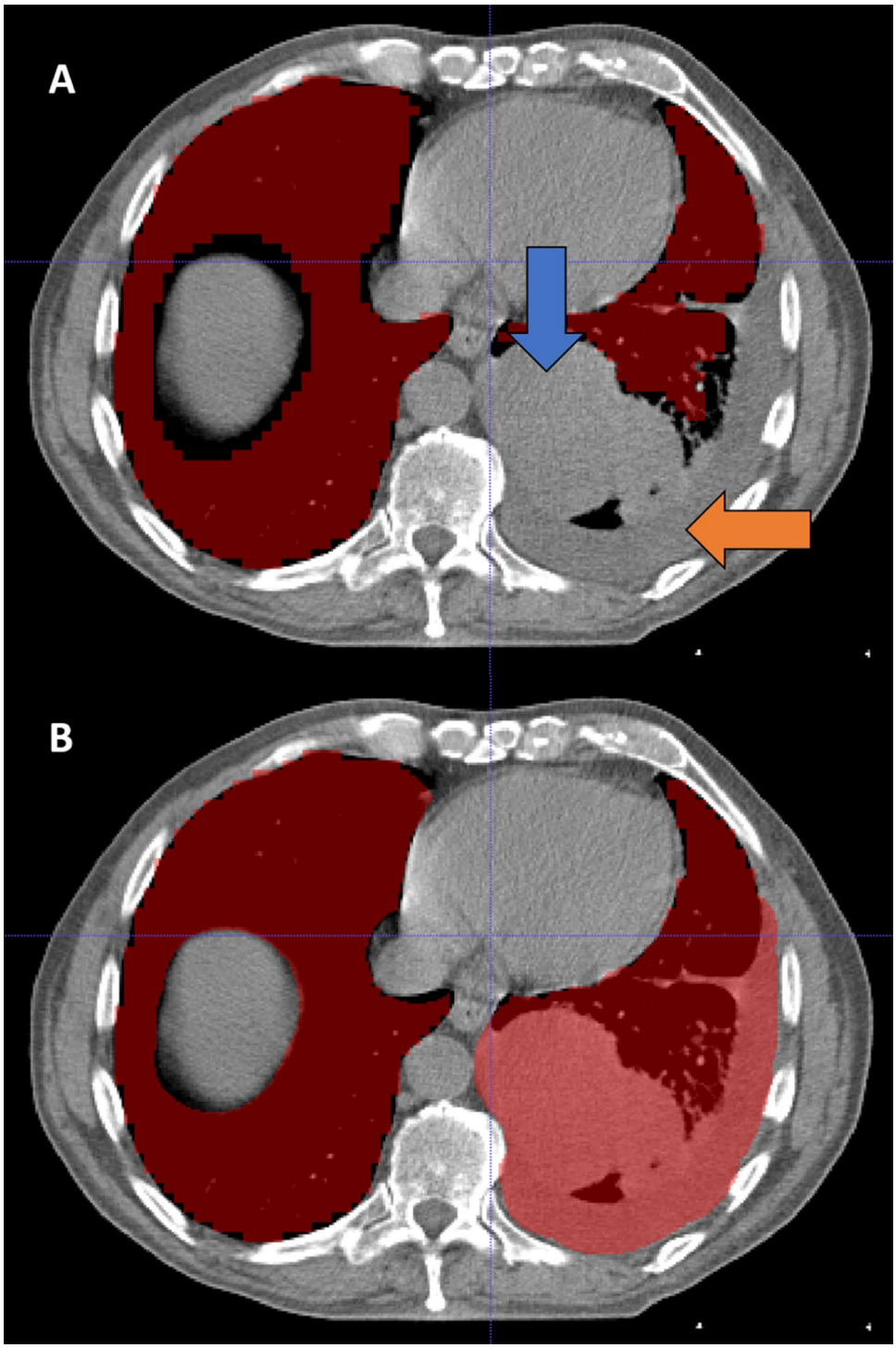


Fig. 2 **A** A deep learning algorithm segmented bilateral thoracic cavity volumes. Accuracy varied in the presence of disease-induced anatomic changes, exemplified by pleural effusion (orange arrow) and primary tumor (blue arrow). **B** A fourth-year medical student corrected the autosegmentations

IIIB: 0.70; p values ≤ 0.004; FNPL median for stage II: 48,262 pixels vs. stage I: 62,698 pixels, stage IIIA: 58,211 pixels, stage IIIB: 58,863 pixels; p values ≤ 0.004; APL median for stage II: 52,260 pixels vs. stage I: 67,446 pixels, stage IIIA: 62,986 pixels, stage IIIB: 67,054 pixels; p values ≤ 0.001). Of the quantitative clinical variables, total thoracic cavity volume was the only significant correlate with correction time ($\rho = 0.19$, $p < 0.001$).

Linear correlations between autosegmentation spatial similarity metrics and correction times were also evaluated. Correction time and metric distribution summary statistics are reported in Table 1. All metrics had statistically significant correlations with correction time (p values < 0.05), but the strength of these correlations varied from strongest to weakest as follows: APL ($\rho = 0.69$, $p < 0.001$), FNPL ($\rho = 0.65$, $p < 0.001$), surface DSC at

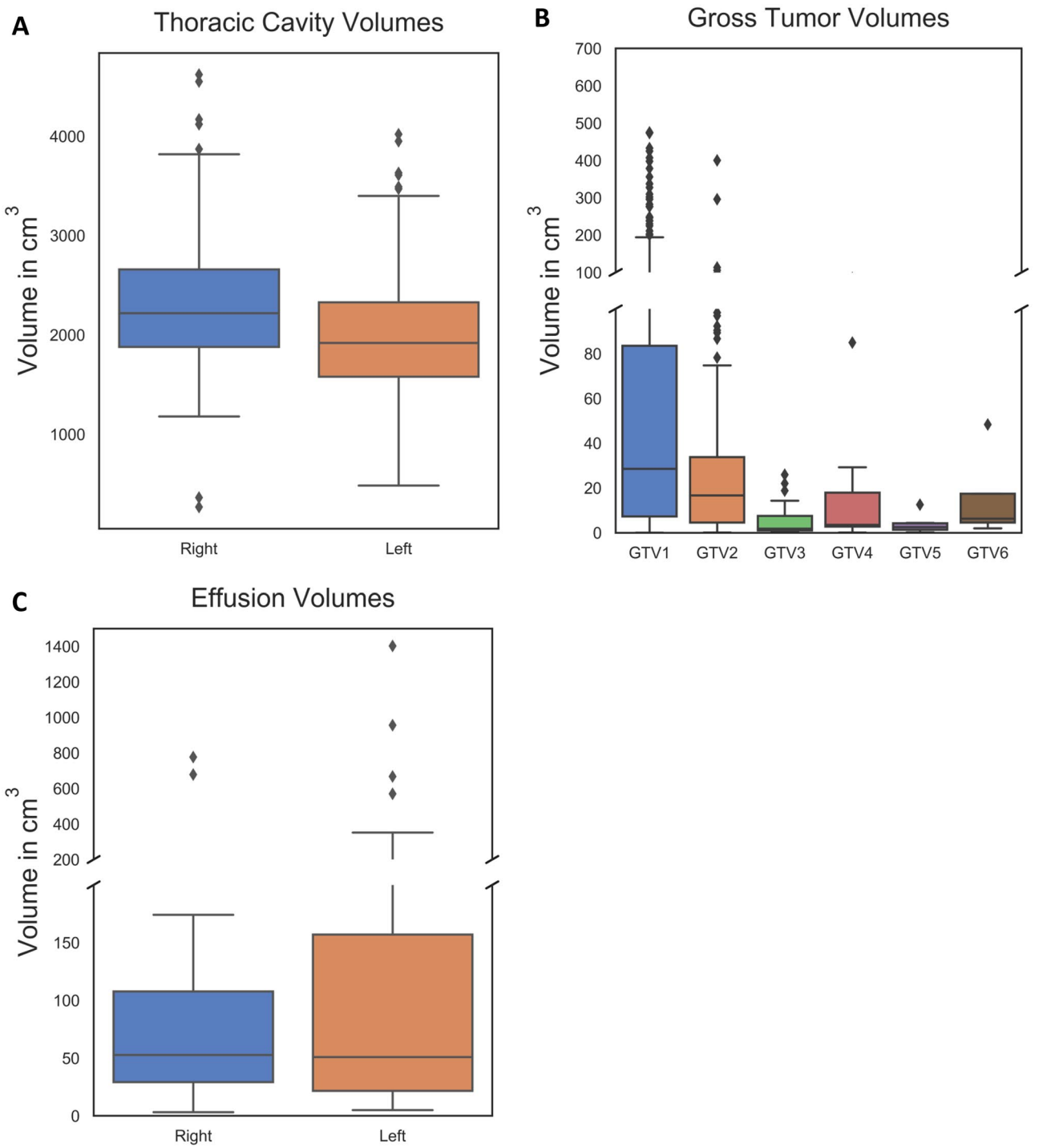


Fig. 3 **A** Right and left thoracic cavity volumes in cases with a recorded autosegmentation correction time ($n = 329$). Volumes were collected after autosegmentation correction by a medical student and subsequent vetting by a physician. **B** Gross tumor volumes as delineated in "RTSTRUCT" segmentation files available from The Cancer Imaging Archive NSCLC-Radiomics data collection [53]. "GTV1" denotes the primary tumor volumes ($n = 328$), whereas "GTV2" through "GTV6" denote secondary tumor volumes that were occasionally present. Usually, the latter were clusters of mediastinal nodes. Because the mediastinum is not part of the lung nor the space healthy lung usually occupies, correlations with tumor volume consider only GTV1, not the sum of GTV1 through GTV6. **C** Right and left pleural effusion volumes in cases with a pleural effusion and a recorded thoracic cavity autosegmentation correction time ($n = 59$). These were delineated de novo by a medical student (rather than corrected from an autosegmentation template) and subsequently vetted by a radiologist

Table 1 Eight spatial similarity metrics were calculated between autosegmented and manually corrected bilateral thoracic cavity segmentations ($n=329$). Two metrics—the surface Dice similarity coefficient and the Hausdorff distance—were calculated with different parameters. Table values are the median, range, and interquartile range for each of these distributions

	Median	Range	Interquartile Range
Contour Correction Time	18.8 min	5.4–51.6 min	7.6 min
Volumetric DSC	0.958	0.354–0.994	0.040
JI	0.919	0.215–0.987	0.072
Surface DSC (0 mm)	0.707	0.108–0.944	0.116
Surface DSC (4 mm)	0.913	0.311–0.987	0.085
Surface DSC (8 mm)	0.954	0.370–0.995	0.062
Surface DSC (10 mm)	0.964	0.391–0.998	0.054
Hausdorff Distance (Max) (mm)	54	15–288	42
Hausdorff Distance (99%) (mm)	29	6–256	41
Hausdorff Distance (98%) (mm)	19	3–246	36
Hausdorff Distance (95%) (mm)	9	1–232	22
Average Surface Distance (mm)	1.5	0.2–57	1.9
Added Path Length (pixels)	64,060	15,658–158,283	23,806
False negative path length (pixels)	58,306	14,555–154,867	21,675
False negative volume (pixels)	102,323	21,482–915,852	76,204

0 mm tolerance ($\rho=-0.48$, $p<0.001$), FNV ($\rho=0.40$, $p<0.001$), JI ($\rho=-0.26$, $p<0.001$), volumetric DSC ($\rho=-0.25$, $p<0.001$), ASD ($\rho=0.24$, $p<0.001$), surface DSC at 4 mm tolerance ($\rho=-0.23$, $p<0.001$), 95th percentile HD ($\rho=0.20$, $p<0.001$), surface DSC at 8 mm tolerance ($\rho=-0.20$, $p<0.001$), surface DSC at 10 mm tolerance ($\rho=-0.19$, $p<0.001$), 98th percentile HD ($\rho=0.17$, $p=0.002$), 99th percentile HD ($\rho=0.11$, $p=0.04$), and maximum HD ($\rho=0.11$, $p=0.05$). Correction time correlations with conformality metrics (volumetric DSC, JI, and best-performing surface DSC) are visualized in Fig. 4, with surface distance metrics (best-performing HD and ASD) in Fig. 5, and with pixel count metrics (APL, FNPL, FNV) in Fig. 6.

Secondary regression analyses were performed between autosegmentation spatial similarity metrics and correction times after stratifying by clinical variables known to have significant relationships with correction times (i.e., T stage, overall stage, and total thoracic cavity volume; thoracic cavity volume was transformed to a categorical variable by binning volumes by quartile). The APL, FNPL, and surface DSC at 0 mm remained highly significant correlates with correction time in every T stage, overall stage, and thoracic volume quartile subgroup (p values <0.001) except the stage IIIA subgroup, in which only the APL and FNPL (but not the surface DSC) were significant correlates. APL ρ correlation coefficients ranged from 0.60 to 0.80 and were the highest of all metrics in every subgroup except the thoracic

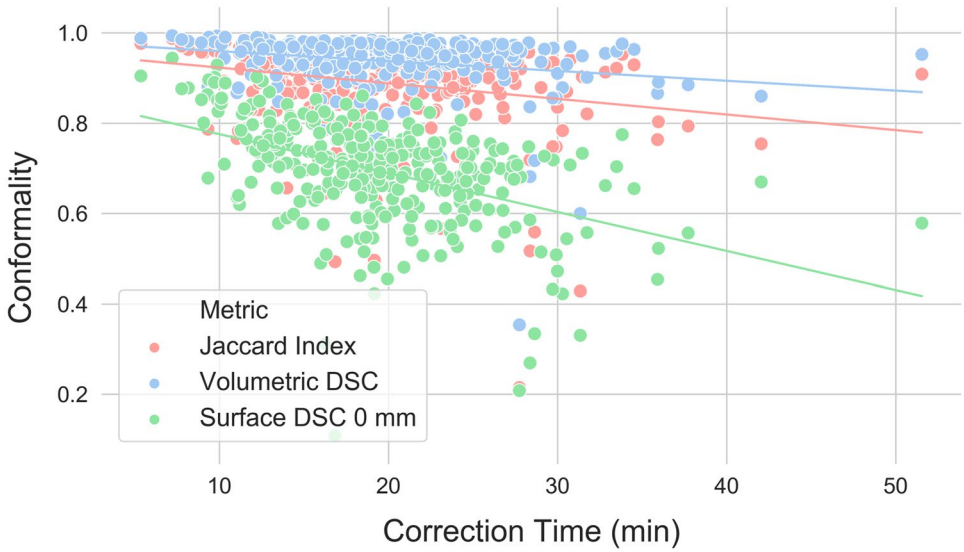


Fig. 4 Correlations between correction time and conformality metrics. The surface Dice similarity coefficient at 0 mm tolerance correlated more strongly with correction time ($\rho=-0.48$, $p<0.001$) than any other conformality, surface distance, or pixel metric except the added path length and false negative path length. Other conformality metrics correlated poorly (Jaccard index: $\rho=-0.26$, $p<0.001$; volumetric Dice similarity coefficient: $\rho=-0.25$, $p<0.001$)

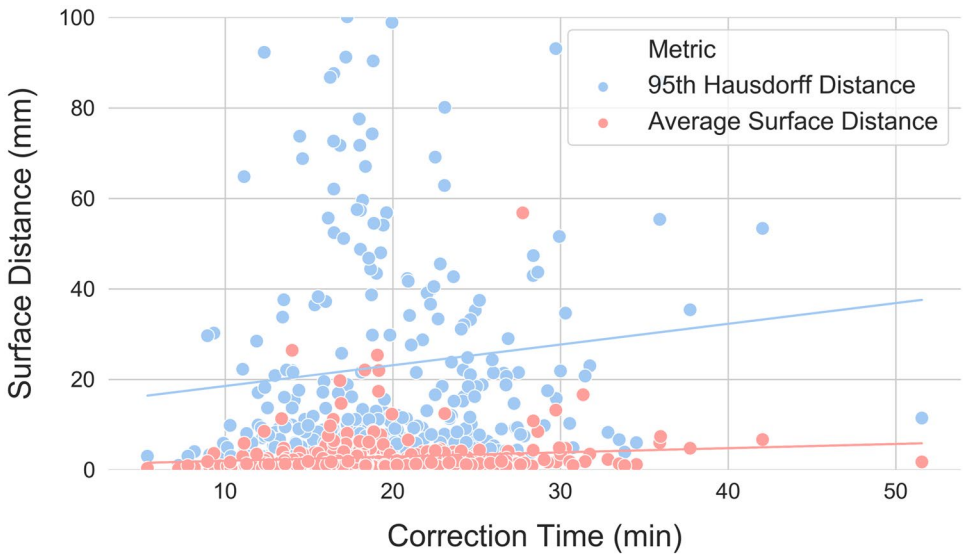


Fig. 5 Correlations between correction time and surface distance metrics. For visual clarity, only the average surface distance ($\rho = 0.24$, $p < 0.001$) and the 95th percentile Hausdorff distance ($\rho = 0.20$, $p < 0.001$) are displayed, which are the two best-performing surface distance metrics. The *y* axis maximum has been limited to better visualize the distributions, excluding ten 95th percentile Hausdorff distance points that exceeded 100 mm. As a class, surface distance metrics were poorer correlates with correction time than conformality or pixel metrics

cavity volume of the 1st quartile subgroup, in which the surface DSC at 0 mm correlation coefficient was slightly stronger ($\rho = -0.62$).

Discussion

Autosegmentation algorithms can assist physicians in an increasing number of clinical tasks, but algorithms are evaluated by spatial similarity metrics that do not necessarily correlate with clinical time savings. The question of which metrics correlate best with time savings has not been thoroughly investigated. To our knowledge, ours is only the second and largest study described for this purpose. In thoracic cavity segmentations delineated on 329 CT datasets, we evaluated correlations between the time required to review and correct autosegmentations and eight spatial similarity metrics. We find the APL, FNPL, and surface DSC to be better correlates with correction times than traditional metrics, including the ubiquitous [4, 6, 10, 11, 16, 22–26, 28, 29, 32–47] volumetric DSC. We find that clinical variables that worsen autosegmentation similarity to manually-corrected references do not necessarily prolong the time it takes to correct the autosegmentations. We also show that APL, FNPL, and surface DSC remain strong correlates with correction time even after controlling for clinical variables that *do* prolong correction time. Using the APL or surface DSC to optimize algorithm training—such as to compute a loss function [69, 70]—may make the algorithms' outputs faster to correct. Using them to assess autosegmentation performance may communicate a more accurate expectation of the time needed to correct the autosegmentations.

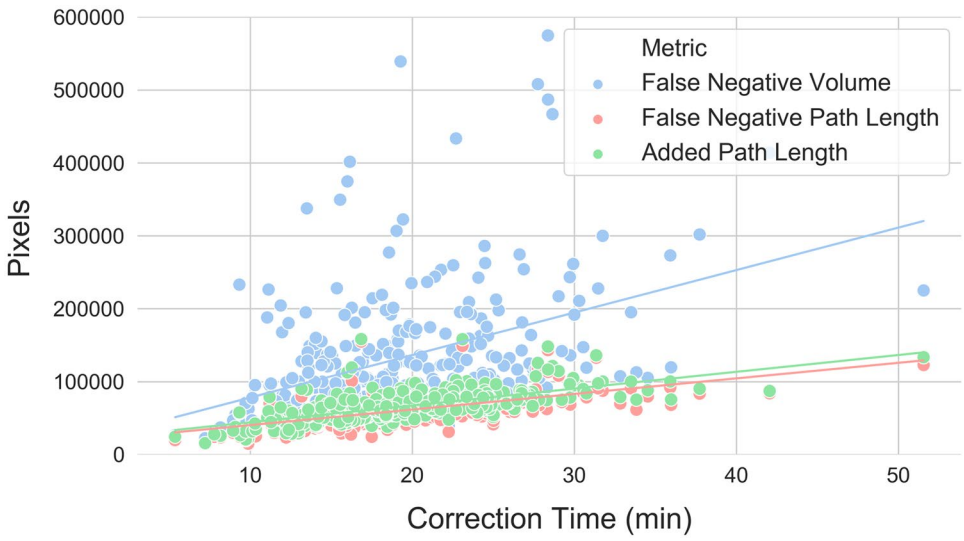


Fig. 6 Correlations between correction time and pixel count metrics. The added path length correlated better with correction time than any other metric ($\rho = 0.69$, $p < 0.001$), while the false negative path length ($\rho = 0.65$, $p < 0.001$) and false negative volume ($\rho = 0.40$, $p < 0.001$) were respectively the second and fourth best performing metrics. The *y* axis maximum has been limited to better visualize the distributions, excluding three false negative volume points between 600,000 and 1,000,000 pixels

Notably, for any comparison of two segmentations where neither can be considered the reference standard, the surface DSC should be preferred to the APL. The surface DSC is directionless, but calculating the APL requires designating one segmentation as a standard.

Autosegmentations that are optimized to save clinicians time may facilitate faster urgent and emergent interventions [1, 2]. They may decrease intraoperative overhead costs [31]. They may be especially beneficial for treatment paradigms that demand daily image segmentation. For example, in an online adaptive MRI-guided radiotherapy workflow, autosegmentations for various anatomic structures are generated every day. Segmentation review occurs while the patient remains in full-body immobilization [30, 71]. This creates a need for a metric to generate a "go/no-go" decision for real-time manual segmentation [72]. Computing the APL between autosegmentations-of-the-day and the physician-approved segmentations from the previous day could signal to the radiation oncologist whether re-segmentation is likely feasible within the time constraints of online fractionation, or whether offline corrections are needed given patient time-in-device. Furthermore, optimized autosegmentation algorithms are foundational to unlocking the benefits of artificial intelligence in radiology; indeed, the Radiological Society of North America, National Institutes of Health, and American College of Radiology identify improved autosegmentation algorithms among their research priorities [73]. These benefits include clinical implementation of radiomics-based clinical decision support systems. While not the only obstacle preventing implementing of these systems, region-of-interest segmentation is currently the rate-limiting step [74].

We corroborate the findings of Vaassen et al., [28] who likewise reported the APL and surface DSC to be superior correlates with correction time. Importantly, our methodology differs from Vaassen et al. in that we used an autosegmentation algorithm that was not optimized to segment thoracic cavity volumes in CT scans from patients with NSCLC, whereas Vaassen et al. used a commercial atlas-based tool and a commercial prototype deep learning tool. The good correlation between the APL and surface DSC and correction time in our study suggests that these metrics may be robust even when evaluating autosegmentation tools that are not highly optimized for their tasks. In contrast, other metrics may degrade in this circumstance. For example, surface distance metrics performed dramatically worse in our study than in Vaassen et al. The maximum, 99th, and 98th percentile HDs were worse correlates with correction time than the surface DSC even at an impractically high error tolerance (10 mm). Given the popularity of the HD as a measure of autosegmentation goodness, this alone is an informative result.

Autosegmentations have achieved unprecedented spatial similarity to reference segmentations [29, 35, 36, 51, 70] and improved computational efficiency [37, 43, 47, 75] since deep learning's [76] emergence in 2012 [77]. Deep learning algorithms should be trained on data representing the spectrum of clinical variation, but the practical consequences of deploying algorithms that are not trained on diverse data remains an outstanding question. Our methodology permits an interesting case study in the time-savings value of deep learning autosegmentation tools that are deployed on classes of data that are underrepresented in the algorithms' training data, since our autosegmentation algorithm was not trained on CTs from patients with NSCLC. We expected that autosegmentation spatial similarity losses due to unseen, cancer-induced anatomic variation would prolong the time required to correct autosegmentations. Rather, we made the interesting observation that clinical variation did not always cost time. Presumably, manual segmentation tools such as adaptable brush sizes and segmentation interpolation were enough to buffer similarity losses.

It is a limitation of this study that autosegmentation corrections were delineated by a fourth-year medical student, but all medical student segmentations underwent subsequent vetting by a radiation oncologist or radiologist and showed very high agreement with physician-corrected segmentations. Furthermore, we acknowledge that our conclusions are limited to the context of thoracic cavity segmentation and should be replicated for clinical autosegmentation tasks across medical domains.

Conclusion

Deep learning algorithms developed to perform autosegmentation for clinical purposes should save clinicians time. It follows that the metrics used to optimize an algorithm ought to correlate closely with clinician time spent correcting the algorithm's product. In this study, we report that three novel metrics—the added path length, the false negative path length, and the surface Dice similarity coefficient—each captured the time-saving benefit of thoracic cavity autosegmentation better than traditional metrics. They correlated strongly with autosegmentation correction time even after controlling for confounding clinical variables. Nevertheless, most algorithms are developed with traditional metrics that we find to be inferior correlates with correction time (most prominently the volumetric Dice similarity coefficient). The findings in this study provide preliminary evidence that novel spatial similarity metrics may be preferred for optimizing and evaluating autosegmentation algorithms intended for clinical implementation.

Acknowledgements The authors thank Dr. Femke Vaassen for the correspondence clarifying the calculation of the added path length.

Funding SS is funded by a grant from the Swiss Cancer League (BIL KLS-4300-08-2017). CDF has received funding and salary support unrelated to this project from: National Institutes of Health (NIH) National Institute for Dental and Craniofacial Research Establishing Outcome Measures Award (1R01DE025248/R56DE025248) and an Academic Industrial Partnership Grant (R01DE028290); National Cancer Institute (NCI) Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program (1R01CA218148); an NIH/NCI Cancer Center Support Grant (CCSG) Pilot Research Program Award from the UT MD Anderson CCSG Radiation Oncology and Cancer Imaging Program (P30CA016672) and an NIH/NCI Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award (P50CA097007); National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant (NSF 1557679); NSF Division of Civil, Mechanical, and Manufacturing Innovation (CMMI) standard grant (NSF 1933369) a National Institute of Biomedical Imaging and Bioengineering (NIBIB) Research Education Programs for Residents and Clinical Fellows Grant (R25EB025787-01); the NIH Big Data to Knowledge (BD2K) Program of the NCI Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (1R01CA214825). CDF has also received direct industry grant support, honoraria, and travel funding from Elekta AB. LG is supported in part by a Learning Healthcare Award funded by the UTHealth Center for Clinical and Translational Science (CCTS), an NIH grant (UL1TR003167), and a Cancer Prevention and Research Institute of Texas grant (RP 170668).

Declarations

Competing Interests The authors declare no competing interests.

References

1. Sheth SA, Lopez-Rivera V, Barman A, et al: Machine Learning-Enabled Automated Determination of Acute Ischemic Core From Computed Tomography Angiography. *Stroke.* 2019;50(11):3093-3100. https://doi.org/10.1161/STROKEAHA.119.026189
2. Gillebert CR, Humphreys GW, Mantini D: Automated delineation of stroke lesions using brain CT images. *Neuroimage Clin.* 2014;4:540-548. https://doi.org/10.1016/j.nicl.2014.03.009
3. Pena-Nogales O, Ellmore TM, de Luis-Garcia R, Suescun J, Schiess MC, Giancardo L: Longitudinal Connectomes as a Candidate Progression Marker for Prodromal Parkinson's Disease. *Front Neurosci.* 2018;12:967. https://doi.org/10.3389/fnins.2018.00967
4. Chen H, Sprengers AMJ, Kang Y, Verdonschot N: Automated segmentation of trabecular and cortical bone from proton density weighted MRI of the knee. *Medical & Biological Engineering & Computing.* 2018;57(5):1015-1027. https://doi.org/10.1007/s11517-018-1936-7
5. van Eijnatten M, van Dijk R, Dobbe J, Streekstra G, Koivisto J, Wolff J: CT image segmentation methods for bone used in medical additive manufacturing. *Med Eng Phys.* 2018;51:6-16. https://doi.org/10.1016/j.medengphy.2017.10.008
6. Rao TJN, Girish GN, Kothari AR, Rajan J: Deep Learning Based Sub-Retinal Fluid Segmentation in Central Serous Chorioretinopathy Optical Coherence Tomography Scans. Paper presented at: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 23–27 July, 2019. https://doi.org/10.1109/EMBC.2019.8857105
7. Kapoor R, Whigham BT, Al-Aswad LA: Artificial Intelligence and Optical Coherence Tomography Imaging. *Asia Pac J Ophthalmol (Phila).* 2019;8(2):187-194. https://doi.org/10.22608/APO.201904
8. Chakravarthy U, Goldenberg D, Young G, et al: Automated Identification of Lesion Activity in Neovascular Age-Related Macular Degeneration. *Ophthalmology.* 2016;123(8):1731-1736. https://doi.org/10.1016/j.ophtha.2016.04.005
9. Moccia S, Foti S, Routray A, et al: Toward Improving Safety in Neurosurgery with an Active Handheld Instrument. *Ann Biomed Eng.* 2018;46(10):1450-1464. https://doi.org/10.1007/s10439-018-2091-x
10. Zaffino P, Pernelle G, Mastmeyer A, et al: Fully automatic catheter segmentation in MRI with 3D convolutional neural networks: application to MRI-guided gynecologic brachytherapy. *Phys Med Biol.* 2019;64(16):165008. https://doi.org/10.1088/1361-6560/ab2f47
11. Fehling MK, Grosch F, Schuster ME, Schick B, Lohscheller J: Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. *PLoS One.* 2020;15(2):e0227791. https://doi.org/10.1371/journal.pone.0227791
12. Aerts HJ, Velazquez ER, Leijenaar RT, et al: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014;5:4006. https://doi.org/10.1038/ncomms5006
13. Zwanenburg A, Vallieres M, Abdalah MA, et al: The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology.* 2020;295(2):328-338. https://doi.org/10.1148/radiol.2020191145
14. Kuhl CK, Truhn D: The Long Route to Standardized Radiomics: Unraveling the Knot from the End. *Radiology.* 2020;295(2):339-341. https://doi.org/10.1148/radiol.2020200059
15. Walker GV, Awan M, Tao R, et al: Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. *Radiother Oncol.* 2014;112(3):321-325. https://doi.org/10.1016/j.radonc.2014.08.028
16. van Heeswijk MM, Lambregts DM, van Griethuysen JJ, et al: Automated and Semiautomated Segmentation of Rectal Tumor Volumes on Diffusion-Weighted MRI: Can It Replace Manual Volumetry? *Int J Radiat Oncol Biol Phys.* 2016;94(4):824-831. https://doi.org/10.1016/j.ijrobp.2015.12.017
17. Miles EA, Clark CH, Urbano MT, et al: The impact of introducing intensity modulated radiotherapy into routine clinical practice. *Radiother Oncol.* 2005;77(3):241-246. https://doi.org/10.1016/j.radonc.2005.10.011
18. Sardanelli F, Quarenghi M, Di Leo G, Boccaccini L, Schiavi A: Segmentation of cardiac cine MR images of left and right ventricles: interactive semiautomated methods and manual contouring by two readers with different education and experience. *J Magn*

Reson Imaging. 2008;27(4):785-792. https://doi.org/10.1002/jmri.21292
19. Altman MB, Kavanaugh JA, Wooten HO, et al: A framework for automated contour quality assurance in radiation therapy including adaptive techniques. *Phys Med Biol.* 2015;60(13):5199-5209. https://doi.org/10.1088/0031-9155/60/13/5199
20. Vinod SK, Jameson MG, Min M, Holloway LC: Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiother Oncol.* 2016;121(2):169-179. https://doi.org/10.1016/j.radonc.2016.09.009
21. Vinod SK, Min M, Jameson MG, Holloway LC: A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. *J Med Imaging Radiat Oncol.* 2016;60(3):393-406. https://doi.org/10.1111/1754-9485.12462
22. van der Veen J, Gulyban A, Nuyts S: Interobserver variability in delineation of target volumes in head and neck cancer. *Radiother Oncol.* 2019;137:9-15. https://doi.org/10.1016/j.radonc.2019.04.006
23. Joskowicz L, Cohen D, Caplan N, Sosna J: Automatic segmentation variability estimation with segmentation priors. *Med Image Anal.* 2018;50:54-64. https://doi.org/10.1016/j.media.2018.08.006
24. Schreier J, Genghi A, Laaksonen H, Morgas T, Haas B: Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam CT and CT. *Radiother Oncol.* 2019;145:1-6. https://doi.org/10.1016/j.radonc.2019.11.021
25. Gambacorta MA, Boldrini L, Valentini C, et al: Automatic segmentation software in locally advanced rectal cancer: READY (REsearch program in Auto Delineation sYstem)-RECTAL 02: prospective study. *Oncotarget.* 2016;7(27):42579-42584. https://doi.org/10.18632/oncotarget.9938
26. Bi N, Wang J, Zhang T, et al: Deep Learning Improved Clinical Target Volume Contouring Quality and Efficiency for Postoperative Radiation Therapy in Non-small Cell Lung Cancer. *Front Oncol.* 2019;9:1192. https://doi.org/10.3389/fonc.2019.01192
27. Fu Y, Mazur TR, Wu X, et al: A novel MRI segmentation method using CNN -based correction network for MRI -guided adaptive radiotherapy. *Med Phys.* 2018;45(11):5129-5137. https://doi.org/10.1002/mp.13221
28. Vaassen F, Hazelaar C, Vaniqui A, et al: Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology.* 2020;13:1-6. https://doi.org/10.1016/j.phro.2019.12.001
29. van der Veen J, Willems S, Deschuymer S, et al: Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiother Oncol.* 2019;138:68-74. https://doi.org/10.1016/j.radonc.2019.05.010
30. Henke LE, Olsen JR, Contreras JA, et al: Stereotactic MR-Guided Online Adaptive Radiation Therapy (SMART) for Ultracentral Thorax Malignancies: Results of a Phase 1 Trial. *Adv Radiat Oncol.* 2019;4(1):201-209. https://doi.org/10.1016/j.adro.2018.10.003
31. Sinha P, Skolnick G, Patel KB, Branham GH, Chi JJ: A 3-Dimensional-Printed Short-Segment Template Prototype for Mandibular Fracture Repair. *JAMA Facial Plast Surg.* 2018;20(5):373-380. https://doi.org/10.1001/jamafacial.2018.0238
32. Ayyalusamy A, Vellaiyan S, Subramanian S, et al: Auto-segmentation of head and neck organs at risk in radiotherapy and its dependence on anatomic similarity. *Radiat Oncol J.* 2019;37(2):134-142. https://doi.org/10.3857/roj.2019.00038
33. Gordaliza PM, Munoz-Barrutia A, Abella M, Desco M, Sharpe S, Vaquero JJ: Unsupervised CT Lung Image Segmentation of a Mycobacterium Tuberculosis Infection Model. *Sci Rep.* 2018;8(1):9802. https://doi.org/10.1038/s41598-018-28100-x
34. Lambert Z, Petitjean C, Dubray B, Ruan S: SegTHOR: Segmentation of Thoracic Organs at Risk in CT images. arXiv:1912.05950. Published 12 Dec 2019. Accessed 01 May 2020.
35. Men K, Zhang T, Chen X, et al: Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. *Phys Med.* 2018;50:13-19. https://doi.org/10.1016/j.ejmp.2018.05.006
36. Men K, Dai J, Li Y: Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. *Med Phys.* 2017;44(12):6377-6389. https://doi.org/10.1002/mp.12602
37. Rhee DJ, Cardenas CE, Elhalawani H, et al: Automatic detection of contouring errors using convolutional neural networks. *Med Phys.* 2019;46(11):5086-5097. https://doi.org/10.1002/mp.13814
38. Roth HR, Lu L, Lay N, et al: Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. *Med Image Anal.* 2018;45:94-107. https://doi.org/10.1016/j.media.2018.01.006
39. Tao CJ, Yi JL, Chen NY, et al: Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: A multi-institution clinical study. *Radiother Oncol.* 2015;115(3):407-411. https://doi.org/10.1016/j.radonc.2015.05.012
40. Thomson D, Boylan C, Liptrot T, et al: Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. *Radiat Oncol.* 2014;9:173. https://doi.org/10.1186/1748-717X-9-173
41. Simmat I, Georg P, Georg D, Birkfellner W, Goldner G, Stock M: Assessment of accuracy and efficiency of atlas-based autosegmentation for prostate radiotherapy in a variety of clinical conditions. *Strahlenther Onkol.* 2012;188(9):807-815. https://doi.org/10.1007/s00066-012-0117-0
42. Vandewinckele L, Willems S, Robben D, et al: Segmentation of head-and-neck organs-at-risk in longitudinal CT scans combining deformable registrations and convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization.* 2019:1–10. https://doi.org/10.1080/21681163.2019.1673824
43. Tong N, Gou S, Yang S, Ruan D, Sheng K: Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Med Phys.* 2018;45(10):4558-4567. https://doi.org/10.1002/mp.13147
44. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al: Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. *Sci Rep.* 2017;7(1):5301. https://doi.org/10.1038/s41598-017-05728-9
45. Voet PW, Dirkx ML, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJ: Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. *Radiother Oncol.* 2011;98(3):373-377. https://doi.org/10.1016/j.radonc.2010.11.017
46. Wardman K, Prestwich RJ, Gooding MJ, Speight RJ: The feasibility of atlas-based automatic segmentation of MRI for H&N radiotherapy planning. *J Appl Clin Med Phys.* 2016;17(4):146-154. https://doi.org/10.1120/jacmp.v17i4.6051
47. Yang J, Veeraraghavan H, Armato SG, et al: Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. *Med Phys.* 2018;45(10):4568-4581. https://doi.org/10.1002/mp.13141
48. Dice LR: Measures of the Amount of Ecologic Association Between Species. *Ecology.* 1945;26(3):297. https://doi.org/10.2307/1932409
49. Rockafellar RT, Wets RJB: *Variational Analysis.* Springer-Verlag Berlin Heidelberg; 1998.
50. Gooding MJ, Smith AJ, Tariq M, et al: Comparative evaluation of autocontouring in clinical practice: A practical method using the

Turing test. *Med Phys.* 2018;45(11):5105-5115. https://doi.org/10.1002/mp.13200
51. Nikolov S, Blackwell S, Mendes R, et al: Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv:1809.04430. Published 12 Sep 2018. Accessed 07 Mar 2019.
52. Kiser KJ, Ahmed S, Stieb S, et al. PleThora: Pleural effusion and thoracic cavity segmentations in diseased lungs for benchmarking chest CT processing pipelines. *Med Phys.* 2020;47(11):5941-5952. https://doi.org/10.1002/mp.14424
53. Aerts HJWL, Wee L, Rios Velazquez E, et al: Data from NSCLC-Radiomics . In: *The Cancer Imaging Archive*. 2019. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI
54. Li X, Morgan PS, Ashburner J, Smith J, Rorden C: The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J Neurosci Methods.* 2016;264:47-56. https://doi.org/10.1016/j.jneumeth.2016.03.001
55. Li X, Morgan PS, Ashburner J, Smith J, Rorden C: *dcm2niix.exe* [computer program]. Version v1.0.201811142020. Accessed January 23, 2019. Available from: https://www.nitrc.org/plugins/mwiki/index.php/dcm2nii:MainPage
56. Pesiuk V: *Lung Segmentation (3D)* [computer program]. GitHub 2017. Accessed December 15, 2018. Available from: https://github.com/imlab-uiip/lung-segmentation-3d
57. Yushkevich PA, Piven J, Hazlett HC, et al: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage.* 2006;31(3):1116-1128. https://doi.org/10.1016/j.neuroimage.2006.01.015
58. Taha AA, Hanbury A: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging.* 2015;15:29. https://doi.org/10.1186/s12880-015-0068-x
59. Oliphant TE: Python for Scientific Computing. *Computing in Science & Engineering.* 2007;9(3):10-20. https://doi.org/10.1109/MCSE.2007.58
60. Abraham A, Pedregosa F, Eickenberg M, et al: Machine learning for neuroimaging with scikit-learn. *Front Neuroinform.* 2014;8:14. https://doi.org/10.3389/fninf.2014.00014
61. McKinney W: Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference (SCIPY 2010).* 2010;445:51–56.
62. Roesch J, Panje C, Sterzing F, et al: SBRT for centrally localized NSCLC - What is too central? *Radiat Oncol.* 2016;11(1):157. https://doi.org/10.1186/s13014-016-0732-5
63. Chang JY, Bezjak A, Mornex F, Committee IART: Stereotactic ablative radiotherapy for centrally located early stage non-small-cell lung cancer: what we have learned. *J Thorac Oncol.* 2015;10(4):577-585. https://doi.org/10.1097/JTO.0000000000000453
64. Wilk MB, Shapiro SS: An analysis of variance test for normality (complete samples). *Biometrika.* 1965;52(3-4):591-611. https://doi.org/10.1093/biomet/52.3-4.591
65. Spearman Rank Correlation Coefficient: In: *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York; 2008:502–505.
66. Mann HB, Whitney DR: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics.* 1947;18(1):50-60. https://doi.org/10.1214/aoms/1177730491
67. Kruskal-Wallis Test: In: *The Concise Encyclopedia of Statistics.* New York, NY: Springer New York; 2008:288–290.
68. Dunn OJ: Multiple comparisons among means. *Journal of the American Statistical Association.* 1961;56(293):52. https://doi.org/10.2307/2282330
69. Zhu W, Huang Y, Zeng L, et al: AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys.* 2018. https://doi.org/10.1002/mp.13300
70. Cardenas CE, McCarroll RE, Court LE, et al: Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function. *Int J Radiat Oncol Biol Phys.* 2018;101(2):468-478. https://doi.org/10.1016/j.ijrobp.2018.01.114
71. Kiser KJ, Smith BD, Wang J, Fuller CD: "Apres Mois, Le Deluge": Preparing for the Coming Data Flood in the MRI-Guided Radiotherapy Era. *Front Oncol.* 2019;9:983. https://doi.org/10.3389/fonc.2019.00983
72. Hunt A, Hansen VN, Oelfke U, Nill S, Hafeez S. Adaptive Radiotherapy Enabled by MRI Guidance. *Clin Oncol (R Coll Radiol).* 2018;30(11):711-719. https://doi.org/10.1016/j.clon.2018.08.001
73. Langlotz CP, Allen B, Erickson BJ, et al. A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology.* 2019;291(3):781-791. https://doi.org/10.1148/radiol.2019190613
74. Ibrahim A, Vallieres M, Woodruff H, et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. *Semin Nucl Med.* 2019;49(5):438-449. https://doi.org/10.1053/j.semnuclmed.2019.06.005
75. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys.* 2017;44(2):547-557. https://doi.org/10.1002/mp.12045
76. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436-444. https://doi.org/10.1038/nature14539
77. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Communications of the ACM.* 2017;60(6):84-90. https://doi.org/10.1145/3065386