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
Arko Barman
Sonja Stieb
Clifton D Fuller
Luca Giancardo

Follow this and additional works at: https://digitalcommons.wustl.edu/oa_4

Part of the Medicine and Health Sciences Commons

Please let us know how this document benefits you.
Novel Autosegmentation Spatial Similarity Metrics Capture the Time Required to Correct Segmentations Better Than Traditional Metrics in a Thoracic Cavity Segmentation Workflow

Kendall J. Kiser1 · Arko Barman2 · Sonja Stieb3 · Clifton D. Fuller3 · Luca Giancardo2

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.


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 otolaryngologists [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
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; theJI; 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

Jaccard Index
\[
\frac{A \cap B}{A \cup B}
\]
where A and B are volumes

Surface Dice Similarity Coefficient
\[
\frac{2(A \cap B)}{A + B}
\]
where A and B are surfaces

Maximum Hausdorff Distance (mm)
\[
\max\{h(A, B), h(B, A)\}
\]
where
\[
h(A, B) = \max_{a \in A} \min_{b \in B} ||a - b||
\]

Average Surface Distance (mm)
\[
\frac{1}{2}(h(A, B) + h(B, A))
\]
where
\[
h(A, B) = \frac{1}{A} \sum_{a \in A} \min_{b \in B} ||a - b||
\]

Added Path Length (pixels)
\[
A - (A \cap B)
\]
where A is the manually corrected segmentation surface and B is the autosegmentation surface

False Negative Path Length (pixels)
\[
A - (A \cap B)
\]
where A is the manually corrected segmentation surface and B is the autosegmentation volume

False Negative Volume (pixels)
\[
A - (A \cap B)
\]
where A is the manually corrected segmentation volume and B is the autosegmentation volume
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, among 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 ($p < 0.001$), 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 ($p < 0.01$). 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; $p \leq 0.006$), and APL (median T4: 69,707 pixels, median T2: 61,553 pixels, median T1: 61,788 pixels; $p \leq 0.002$). Autosegmentations on CT scans with overall stage I 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; $p \leq 0.04$). 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...
IIIB: 0.70; \( p \) values \( \leq 0.004 \); 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 \( \leq 0.004 \); 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 \( \leq 0.001 \). 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. 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
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.

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

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 conformity 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 $\rho$ 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 conformity 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 conformity, surface distance, or pixel metric except the added path length and false negative path length. Other conformity metrics correlated poorly (Jaccard index: $\rho = -0.26, p < 0.001$; volumetric Dice similarity coefficient: $\rho = -0.25, p < 0.001$).
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 data-sets, 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.
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 (R01DE028299); 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 (PS0CA097007); 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.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

**References**


Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.