ノート

Evaluation of an automatic contour segmentation software on segmentation of liver in average 4-dimensional computed tomography images

オープンソース自動セグメンテーションソフトによる平均4次元CT画像の肝臓セグメンテーション評価

KOHIGASHI Junya[1], OSADA Shoui[2], SAIYO Nipon[2,3,4], SEGAWA Yuto[2], NOTO Kimiya[5], KOJIMA Hironori[5], KAIDO Ryoto[2,6], TAKEMURA Akihiro[7]* ［* Corresponding Author］

1）Department of Radiology, Toyama Prefectural Central Hospital
2）Division of Health Sciences, Graduate School of Medical Sciences, Kanazawa University
3）Present address: School of Radiological Technology, Faculty of Health Science Technology, Chulabhorn Royal Academy
4）Present address: Personalized Radiotherapy and Imaging in Medicine (PRIME) Research Center, Chulabhorn Royal Academy
5）Department of Radiology, Kanazawa University Hospital, Kanazawa University
6）Radiological Center, University of Fukui Hospital
7）Faculty of Health Sciences, Institute of Medical, Pharmaceutical and Health Sciences, Kanazawa University

Key words: MONAI Auto3DSeg, liver segmentation, average 4DCT image, radiotherapy research

【Abstract】
This study evaluated the accuracy of liver delineation on the MONAI Auto3DSeg (Monai) extension of 3D Slicer software using four-dimensional computed tomography (4DCT) images for radiotherapy treatment planning. Average 4DCT images of 13 patients who underwent radiotherapy were obtained. Liver contours in the average 4DCT images were delineated by the Monai extension (A-Seg) and an inexperienced researcher (M-Seg). A- and M-Segs were compared to reference contours from radiation oncologists. The Dice similarity coefficient for A-Seg (0.93 ± 0.01) was higher than that for M-Seg (0.91 ± 0.02, p<0.01). Hausdorff distance for the A-Seg (28.0 ± 10.0 mm) was shorter than that for the M-Seg (34.1 ± 13.7 mm, p=0.24). Monai delineated the liver in average 4DCT images with high accuracy, comparable to radiation oncologists and superior to an inexperienced researcher.

【要 旨】
放射線治療計画での4次元CTの平均画像（平均4DCT）に対し，3D SlicerのMONAI Auto3DSeg（Monai）による肝臓輪郭描出の精度を検証した．13例の平均4DCTを用い，医師による肝臓輪郭を基準にMonaiおよび臨床経験のない研究者による輪郭（それぞれA-Seg，M-Seg）を比較した．Dice係数はA-Seg 0.93，M-Seg 0.91であった（p<0.01）．ハウスドルフ距離はA-Seg 28.0 mm，M-Seg 34.1 mmであった（p=0.24）．Monaiの肝臓輪郭は医師と同程度で，臨床経験のない研究者より高精度であった．

Introduction

Radiotherapy treatment planning requires accurate contouring of targets and organs. Manual contouring by radiation oncologists and medical physicists is time-consuming and prone to inter-observer variability[1-6]. Automatic contour extraction may reduce inter-operator variability and working time. Recently, automatic contour extraction methods using artificial intelligence (AI) have been developed and their usefulness has been reported[7-9].

Automatic contouring software is effectively evaluates deformable image registration (DIR) owing to its contouring stability and time efficacy. However, the cost and limitations of clinically available software can hinder its used in research settings. Therefore, open-source automatic contour software for medical images is a suitable

小東 純也[1]，**長田 翔聖（学生）**[2]，**SAIYO Nipon（学生）**[2,3,4]，**瀬川 裕斗（学生）**[2]，**能登 公也**[5]，**小島 礼慎**[5]，**街道 亮斗**[2,6]，**武村 哲浩**[7]*

1）富山県立中央病院放射線部
2）金沢大学大学院医薬保健総合研究科保健学専攻
3）現：診療放射線技術学科，医療科学技術学部，チュラポーン・ロイヤル・アカデミー
4）現：個別化放射線治療・医療画像研究センター，チュラポーン・ロイヤル・アカデミー
5）金沢大学附属病院放射線部
6）福井大学医学部附属病院放射線部
7）金沢大学医薬保健研究域保健学系
＊ E-mail: at@mhs.mp.kanazawa-u.ac.jp

Received October 19, 2025; accepted February 4, 2026

option. The MONAI project, an open-source deep-learning framework, provides several models for various purposes in the healthcare field[10]. The MONAI Auto3DSeg (Monai) extension of the 3D Slicer software[11] offers automatic contouring with AI. This is easy to use because the 3D slicer operates at the front end of the Monai Extension. Although previous studies have evaluated the accuracy of the Monai extension on cone-beam CT (CBCT) images[12], no study has assessed its performance using average four-dimensional computed tomography (4DCT) images in radiotherapy treatment planning.

4DCT is acquired for radiotherapy planning, and the organs and target volumes are delineated on average 4DCT images created from the 4DCT images by averaging all respiratory phase images. The average 4DCT image represents a range of tumor respiratory motion and the range is defined as an internal target volume for radiotherapy planning. The average 4DCT image was blurred because of the averaging of respiration phases. Additionally, in some cases, irregular respiration causes split liver or spleen on a diaphragm. Therefore, whether the Monai extension works well for average 4DCT images remains unclear. Moreover, DIR-related research requires evaluation of the contours of the organs. Previous studies have shown that DIR does not work well for abdominal organs, such as the liver, owing to constant movement during respiration and low and homogeneous image contrast[13-15]. Currently, we are investigating an evaluation method for organs with low and homogeneous contrast, such as the liver. This study evaluated the accuracy of the Monai extension on average 4DCT images for radiotherapy treatment planning, and its potential effectiveness in DIR research.

Materials and Methods

Average 4DCT images of 13 patients (11 men, two women, median age; 75 years [53–85]) who underwent radiotherapy were collected. The patients included seven with liver cancer and six with pancreatic cancer. Images were acquired using a CT simulator (Aquilion Exceed LB, CANON MEDICAL SYSTEMS CORPORATION, Tochigi, Japan). The 4DCT images were obtained using a respiration-gated helical scan (120 kV), 100 mA effective tube current with automatic exposure control, 2.0 mm slice thickness, and an FC05 reconstruction kernel. Respiration waveforms were collected using an optical surface scanning system (Sentinel, C-Rad AB, Uppsala, Sweden). Using the respiration waveform, 10-phase 4DCT images were reconstructed, and phase-averaged images were created from all-phase images. This study was approved by the IRB of our university (#2023-015).

The reference contour of the liver in the average 4DCT images was delineated by radiation oncologists and used for radiotherapy planning and treatment. A researcher with no clinical experience delineated the liver using 4DCT images (M-Seg), as advised by a researcher with 20 years of radiotherapy experience. The M-seg served as the benchmark if delineated by an inexperienced researcher. The Monai extension was applied to the same averaged 4DCT images to create liver contours (A-Seg). The Monai extension offers models of body sites. This study used the Abdominal Organs TS2-Quick (v2.0.0), which is a 3 mm low-resolution model, because of its adequate resolution and low calculation costs. Abdominal Organs TS2-Quick (v2.0.0) uses a SegResNet model[16], which is provided by MONAI project and has been trained using the dataset that TotalSegmentator[17] released. TotalSegmentator is a segmentation model that

has released the model and dataset they used for the model training. The TotalSegmentator dataset comprises diagnostic computed tomography and magnetic resonance images. An alternative option was Abdominal Organ TS2 (v2.0.0), which was the high resolution model; however, it failed to calculate for our cases on our desktop PC with AMD Ryzen 5 3600, 16 GB of memory, and NVIDIA GeForce GTX 1650.

The Dice similarity coefficient (DSC) and Hausdorff distance (HD) were calculated for M- and A-Seg based on the reference contour. The liver volume difference between the M- and A-Seg from the reference contour was measured. DSC and HD were used to evaluate the DIR accuracy[18]. The DSC was calculated using Equation 1.

$$\mathrm{DSC} = \frac{2|X \cap Y|}{|X| + |Y|} \quad \cdots\cdots (1)$$

in which X and Y are the two contour regions to be compared.

HD was calculated using Equation 2.

$$\mathrm{HD} = max_{x \in X}\left\{min_{y \in Y} d(x, y)\right\} \quad \cdots\cdots (2)$$

in which x denotes any boundary point on contour X; y denotes any boundary point on contour Y; and d (x, y) denotes the distance between x and y.

The DSC and HD results for M- and A-Segs were statistically compared using EZR ver. 1.66 (Saitama Medical Center, Jichi Medical University, Saitama, Japan)[19], which is a graphical user interface for R version 4.5.2 (The R Foundation for Statistical Computing, Vienna, Austria).

Results

The DSC for the A-Seg (0.93 ± 0.01) was significantly better than that for the M-Seg

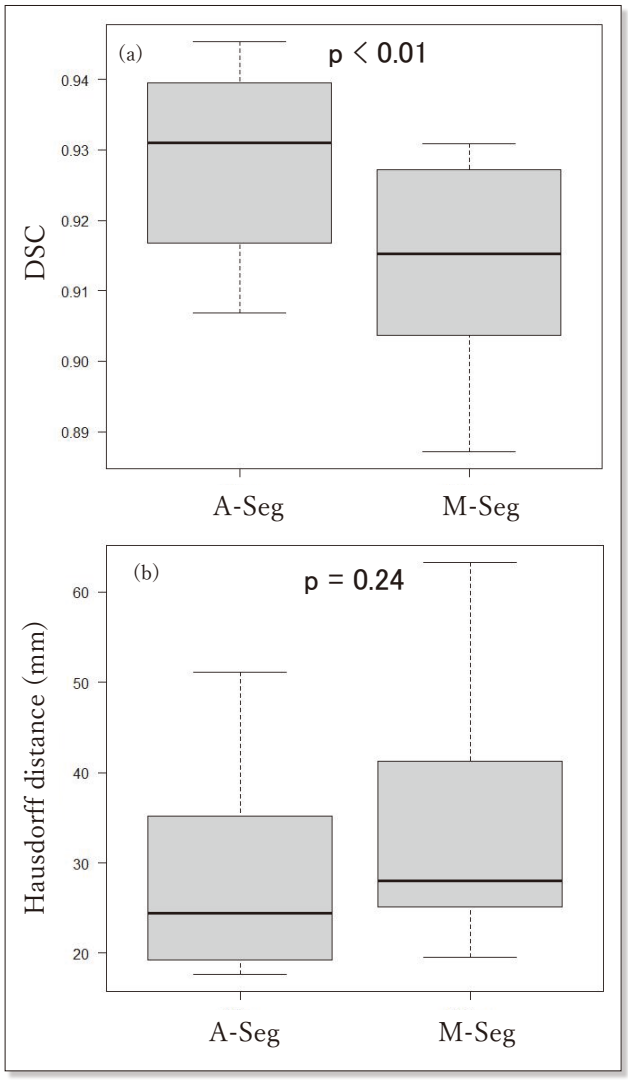

Fig. 1 Comparison of (a) DSC and (b) HD for A- and M-Segs

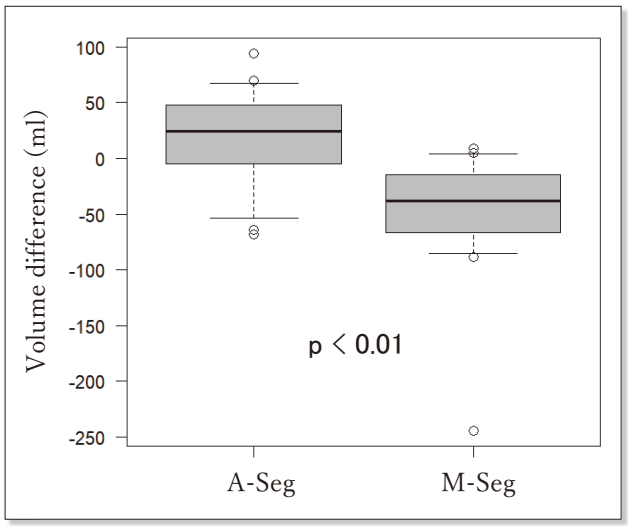

Fig. 2 Comparison of volume difference of segmented liver

Volume difference for A-Seg was significantly smaller than that for M-Seg ($p<0.01$) tested by Wilcoxon signed-rank exact test

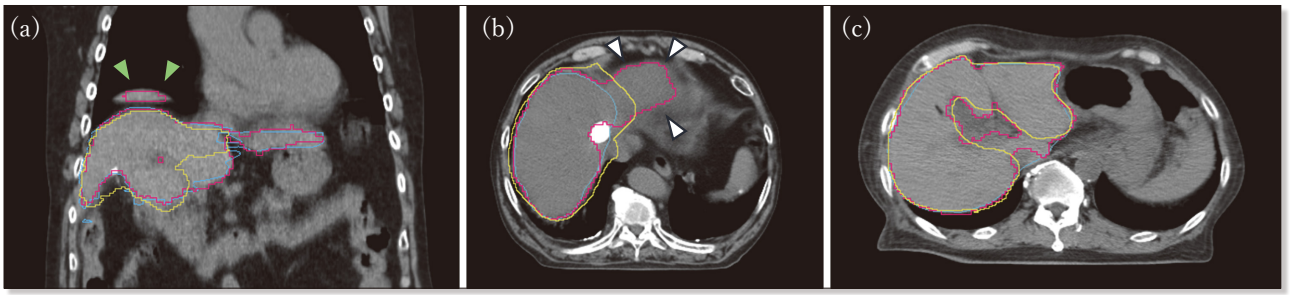

Fig. 3 Liver segmentation results

Contours of the reference, A-Seg, and M-Seg are shown in blue, red, and yellow, respectively. Monai could correctly recognize doubled liver top (green triangle) (a). Monai failed to segment the lipiodol region and misrecognized the bottom of the heart as the top of the liver (white triangles) (b). However, the right lobe of liver could be recognized better by Monai than by the unexperienced researcher. (c) The worst DSC case.

(0.91 ± 0.02, $p<0.01$, Figure 1a) by paired t-test. The hypothesis that the DSC results for A- and M-Segs were normal was not rejected by the Shapiro–Wilk normality test ($p=0.543$), and the homogeneity of variance between these results was not rejected ($p=0.561$). Contrastingly, the HD for the A-Seg (28.0 ± 10.0 mm) was shorter than that for the M-Seg (34.1 ± 13.7 mm). However, no significant difference in HD was noted between the A- and M-Seg groups ($p=0.244$, Figure 1b). Normality of HD was rejected using the Shapiro–Wilk test ($p=0.017$).

The volume difference for A-Seg was significantly smaller than that for M-Seg by Wilcoxon signed-rank exact test ($p<0.01$, Figure 2). Shapiro–Wilk normality test rejected the normality of the results ($p=0.393$). Average volumes difference were 17.7 ± 46.2 mL for A-Seg and -51.8 ± 63.0 mL for M-Seg.

Examples of the results were shown in Figure 3. Monai could delineate doubled liver top (Fig. 3a); however, it sometimes failed to correctly contour the boundary between the liver and heart, which had low intensity difference (Fig. 3b) and excluded a lipiodol region. The worst case was shown in Fig. 3c and major difference was whether vessels in the liver were excluded. The reference contour of the liver included these regions; however, A- and M-Segs excluded these regions, and Monai excluded more exactly these regions.

Discussions

According to the DSC results, the Monai extension was delineated significantly better than by the inexperienced researchers. In the DIR evaluation, the DSC value was 0.8, indicating good agreement[18]. Bilic et al. summarized the liver and liver tumor segmentation challenge results[20]. The challenges were competitions for automatic segmentation methods to segment the liver and liver tumors in 2016–2018. Almost all methods were achieved with the DSC of ≥0.92, and the results was similar to the DSC result of Monai extension. In the competitions, many methods employed a U-net based model. Monai extension with Abdominal Organs TS2 employed the SegResNet model, which was a U-net like model. Thus, Monai extension could achieve similar DSC results.

The mean DSC for A-Seg was superior to that value (0.93); therefore, the Monai extension should have the same accuracy for liver delineation in the average 4DCT images as the radiation oncologists. Additionally, no significant difference was observed between the HD for A- and M-Segs. Monai extension attempted to delineate the liver, excluding portal veins and hepatic arteries (Fig. 3c). Compared with contour boundaries, HD had a

relatively large variance. However, the HD for A-Segs was smaller than that for M-Segs. Thus, Monai extension delineated the liver with the same accuracy as radiation oncologists and more accurately than an inexperienced researcher. Monai extension delineated larger areas than the radiation oncologists. The volume of the A-Seg was slightly larger than that of the reference contour, with a difference of only 17.7 mL (1.5%). The volume of the M-Seg was slightly smaller than that of the reference contour, with a -51.8 mL (4.4%) difference. The volume difference between A- and M-Segs was significant ($p<0.01$), and the volume of A-Seg was similar to the reference contour delineated by radiation oncologists.

The average 4DCT image sometimes exhibits artifacts, such as the diaphragm and liver appearing double-illuminated owing to irregular breathing during 4DCT scanning. The Monai extension was trained with the same dataset as TotalSegmentator was trained. The dataset excluded average 4DCT images for radiotherapy planning. Moani extension had difficulty segmenting our data. Despite having these artifacts, Monai extension could accurately delineate the liver (Fig. 3a). This demonstrates the adaptability of Monai extension to the average 4DCT images, which are used for radiotherapy planning.

The Monai extension is useful for research that requires organ segmentation in CT images, such as DIR. The Monai extension and 3D slicer have not been approved for clinical use. However, we were able to run the Monai extension on a general performance desktop PC in the laboratory. The Monai extension provided segmentation results comparable to radiation oncologists and superior to an inexperienced researcher, despite supervision by an experienced researcher. This finding suggests that the Monai extension can be used by radiation oncologists in preliminary studies.

A limitation of this study is that only the liver contour was focused. Other abdominal organs, such as the kidneys and spleen, were excluded. DSC affects the organ volume. The liver is the largest organ; therefore, its DSC value should be better than those of other organs. Thus, the Monai extension may not provide accurate results for other organs.

Conclusions

This study evaluated the Monai extension regarding the accuracy of liver segmentation using 4DCT images for radiotherapy planning. The Monai extension can delineate the liver with the same accuracy as radiation oncologists. Additionally, this tool may be beneficial for DIR researchers with limited preliminary clinical experience.

Conflict of interest

The authors have no conflict of interest.

Acknowledgements

We would like to thank Editage (www.editage.jp) for English language editing.

図の説明

Fig.1　A-SegおよびM-SegのDSCとHDの結果
Fig.2　肝臓輪郭の体積差比較
A-Segの体積差はM-Segより小さく，ウィルコクソン符号順位検定により有意があった（$p<0.01$）.
Fig.3　肝臓抽出結果
基準となる医師の輪郭, A-Seg, M-Segをそれぞれ青・赤・黄色の線で表している．(a) Monaiは分離した肝臓も正しく認識できた（緑色矢印）．(b) Monaiはリピオドールが入った領域を肝臓と認識できず，また心臓底面を肝臓と認識した．しかし，(c) 肝臓右葉は経験の少ない研究者と比較し，より良い結果となった．この画像がDSCが最も悪かった症例である.

References

1) Joskowicz L, et al.: Inter-observer variability of manual contour delineation of structures in CT. Eur Radiol, 29, 1391-1399, 2019.
2) Guzene L, et al.: Assessing Interobserver Variability in the Delineation of Structures in Radiation Oncology:

A Systematic Review. Int J Radiat Oncol Biol Phys, 115(5), 1047-1060, 2023.
3) Patrick HM, et al.: Reduction of inter-observer contouring variability in daily clinical practice through a retrospective, evidence-based intervention. Acta Oncol, 60, 229-236, 2021.
4) Corrao G, et al.: Intra- and inter-observer variability in breast tumour bed contouring and the controversial role of surgical clips. Med Oncol, 36, 51, 2019.
5) Li F, et al.: Inter-Observer and Intra-Observer Variability in Gross Tumor Volume Delineation of Primary Esophageal Carcinomas Based on Different Combinations of Diagnostic Multimodal Images. Front Oncol, 12, 817413, 2022.
6) Louie AV., et al.: Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era. Radiother Oncol, 95, 166-171, 2010.
7) Heilemann G, et al.: Clinical Implementation and Evaluation of Auto-Segmentation Tools for Multi-Site Contouring in Radiotherapy. Phys Imaging Radiat Oncol, 28, 100515, 2023.
8) Doolan PJ, et al.: A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. Front Oncol, 13, 1213068, 2023.
9) Yang C, et al.: Deep learning in CT image segmentation of cervical cancer: a systematic review and meta-analysis. Radiat Oncol, 17, 175, 2022.
10) Cardoso MJ, et al.: MONAI: An open-source framework for deep learning in healthcare. arxiv.org, 2022. Available from: https://arxiv.org/abs/2211.02701
11) Fedorov A, et al.: 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging, 30, 1323-1341, 2012.
12) Süküt Y, et al.: Accuracy of deep learning-based upper airway segmentation. J Stomatol Oral Maxillofac Surg, 126, 102048, 2025.
13) Brock KK: Results of a Multi-Institution Deformable Registration Accuracy Study (MIDRAS). Int J Radiat Oncol Biol Phys, 76, 583-596, 2010.
14) Hoffmann C, et al.: Accuracy quantification of a deformable image registration tool applied in a clinical setting. J Appl Clin Med Phys, 15, 237-245, 2014.
15) Schreibmann E, et al.: A measure to evaluate deformable registration fields in clinical settings. J Appl Clin Med Phys, 13, 126-139, 2012.
16) Myronenko A: 3D MRI brain tumor segmentation using autoencoder regularization. arXiv:1810.11654v3, 2018.
17) Wasserthal J. et al.: TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiology: Artificial Intelligence, 5, e230024, 2023.
18) Brock KK, et al.: Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. Med Phys, 44, e43-e76, 2017.
19) Kanda Y: Investigation of the freely-available easy-to-use software “EZR” (Easy R) for medical statistics. Bone Marrow Transplant, 48, 452-458, 2013.
20) Bilic P. et al.: The Liver Tumor Segmentation Benchmark (LiTS). Med Image Analys, 84, 102680, 2023.