Artificial Intelligence-powered automatic volume calculation in medical images – available tools, performance and challenges for nuclear medicine

Automatische Volumenberechnung mithilfe künstlicher Intelligenz in der medizinischen Bildgebung – verfügbare Werkzeuge, Performance und Herausforderungen für die Nuklearmedizin

Authors

Thomas Wendler[1, 2, 3], Michael C. Kreissl[4], Benedikt Schemmer[5], Julian Manuel Michael Rogasch[6], Francesca De Benetti[3]

Affiliations

1 Clinical Computational Medical Imaging Research, Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Augsburg, Germany
2 Institute of Digital Medicine, Universitätsklinikum Augsburg, Germany
3 Computer-Aided Medical Procedures and Augmented Reality School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
4 Abteilung für Nuklearmedizin, Universitätsklinikum Magdeburg, Germany
5 Department of Nuclear Medicine, Universitätsklinikum Bonn, Germany
6 Department of Nuclear Medicine, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Germany

Key words

artificial intelligence, machine learning, deep learning, volumetry, volume estimation, volume calculation

received 16.10.2023
accepted 26.10.2023

Bibliography

Nuklearmedizin 2023; 62: 343–353
DOI 10.1055/a-2200-2145
ISSN 0029-5566

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Correspondence

Prof. Thomas Wendler
Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Augsburg, Stenglinstr. 2, 86156 Augsburg, Germany
Tel.: +49/8 21/4 00 24 05
Fax: +49/8 21/4 00 33 12
thomas.wendler@med.uni-augsburg.de

ABSTRACT

Volumetry is crucial in oncology and endocrinology, for diagnosis, treatment planning, and evaluating response to therapy for several diseases. The integration of Artificial Intelligence (AI) and Deep Learning (DL) has significantly accelerated the automatization of volumetric calculations, enhancing accuracy and reducing variability and labor. In this review, we show that a high correlation has been observed between Machine Learning (ML) methods and expert assessments in tumor volumetry; Yet, it is recognized as more challenging than organ volumetry. Liver volumetry has shown progression in accuracy with a decrease in error. If a relative error below 10 % is acceptable, ML-based liver volumetry can be considered reliable for standardized imaging protocols if used in patients without major anomalies. Similarly, ML-supported automatic kidney volumetry has also shown consistency and reliability in volumetric calculations. In contrast, AI-supported thyroid volumetry has not been extensively developed, despite initial works in 3D ultrasound showing promising results in terms of accuracy and reproducibility. Despite the advancements presented in the reviewed literature, the lack of standardization limits the generalizability of ML methods across diverse scenarios. The domain gap, i. e., the difference in probability distribution of training and inference data, is of paramount importance before clinical deployment of AI, to maintain accuracy and reliability in patient care. The increasing availability of improved segmentation tools is expected to further incorporate AI methods into routine workflows where volumetry will play a more prominent role in radionuclide therapy planning and quantitative follow-up of disease evolution.

Introduction

In nuclear medicine, volumetry (i. e., calculating the volume of organs, tumors or other lesions of interest) is essential for various clinical indications. Accurate volume measurements help in diagnosis, prognosis, treatment planning, and response assessment as well as planning and dosimetry for radionuclide therapy to determine the optimal administered radioactivity and estimate the real deposited dose.

In oncology, assessing tumor size and extent is relevant for (a) staging, (b) monitoring changes following treatment and hence to determine the efficacy of the intervention, and (c) to predict the likelihood of metastasis, recurrence, and overall patient outcome. In the special case of selective internal radiotherapy (SIRT), pretherapeutic volumetry of the tumor and the treated liver segments is often used to determine the dose to tumor and normal liver [1]. Thus it provides crucial information to decide if SIRT can be offered as an effective and safe treatment for an individual patient, and in case this is true, to calculate the optimal therapeutic activity.

In endocrinology, volumetry is crucial for assessing thyroid volume in cases of goiter, to calculate the radioactivity to be administered in radioiodine therapy for patients with Graves' disease, or to trigger potential changes in the management strategy for suspicious thyroid nodules.

Although several tools are available that are dedicated to volumetry or include manual or semi-automated segmentation and volume calculation options (e. g., in alphabetic order, DosePlan, Freesurfer, ImFusion, MIM, LIFEx, Satori, among many), they are prone to errors and labor-intensive because they require manual intervention. Among the potential pitfalls are (a) the erroneous transfer of labels between slides, where a manual annotation is wrongly extended to the neighboring slides in particular in cases where the spacing between them is larger than the in-plane resolution, (b) the incorrect definition of margins between structures due to low contrast, such as the boundary between heart and liver, or (c) the creation of artificial "holes" in segmentation masks due to high contrast variation within structures, such as calcifications in thyroid and liver.

Artificial Intelligence (AI) and in particular Deep Learning (DL) can automate volume calculations, thereby reducing intraobserver and interobserver variability, and thus improving the accuracy of static and longitudinal volume comparison. Artificial neural networks can efficiently process 3D images, segment structures of interest, and based on those segmentations, calculate volumes with high accuracy. Some AI techniques are capable of adapting to variations in image quality and anatomical structures, improving consistency and reproducibility in volume measurements across different observers and imaging sessions compared to humans in several studies (e. g., [2–4]). The major impact of AI can be shown in 3D ultrasonography, a growing field given the drop in costs (due to the possibility of reimbursement and the availability of cheaper devices), where the steep learning curve currently makes it difficult for users to exploit its full potential [2, 5].

It is worth noting that even in nuclear medicine applications, volumetry is often not applied to the scintigram, single-photon emission tomography (SPECT), or positron emission tomography (PET) image but to the radiological counterpart, because the limited spatial resolution of nuclear medicine images and partial volume effects complicate the definition of lesion boundaries. Also due to longer acquisition times, these images are more prone to distortion from breathing.

In this review, we therefore summarize existing literature on the performance of these AI methods in computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound (US) or – if suitable – when applied to PET images. We close the paper discussing challenges and perspectives for the near future.

Volumetry

Volumetry, also known as volume calculation or volume estimation, is the procedure of determining the volume of particular structures. A common application in the context of clinical care in nuclear medicine is thyroid volumetry, where the size of the two thyroid lobes of a patient is measured using 2D ultrasound. For this the thickness and width of each of the lobes are measured directly on the ultrasound plane of an axial image, while its length is obtained from the perpendicular sagittal plane. All three measures are then combined with the measures of the contralateral lobe to obtain an approximated thyroid volume using the ellipsoid formula [6, 7]. This approach is clearly not optimal in terms of accuracy, but has become a standard worldwide due to its simplicity. Similar methods are applied for the volumetry of other organs, such as the liver [8].

If volumetry is performed in 3D images, more accurate methods can be applied. Using image segmentation methods, a binary mask of the structure of interest can be obtained. As a subsequent step, the volume is calculated by counting the number of non-zero voxels and multiplying the total sum by the voxel size. The accuracy of such a method relies completely on the quality of the segmentation mask. Most volumetry approaches that incorporate machine learning (ML) follow this approach [9].

Alternatively, the volume of the structure of interest can be directly inferred from the images without the need for segmentation. Such an approach has been proposed, e. g., by Hussain et al. [9], yet such works seem to be an exception.

To quantify the quality of volume estimations, typically the relative volume error (RVE) is calculated with respect to an expert. Whenever the volume is calculated from a segmentation mask, the RVE can be calculated as the (absolute) difference between false positive and false negative pixels or voxels divided by the true positive plus the false negative ones. Alternatively, some works report the absolute volume error (AVE) or the correlation, in terms of Pearson's correlation coefficient (also known as interobserver correlation coefficient, ICC), between AI and human experts. Some works evaluate the interobserver variability of the volume calculation, or the interobserver variability of the error on the volume. For the latter, the volume calculated by an expert is assumed as ground truth to calculate the volume error, which is then analyzed in terms of variability between non-expert readers and automatic methods. Metrics like the well-established Sørensen-Dice coefficient (DSC) and the Jaccard-Index (JI), also known as Intersection-over-Union (IoU), are proxies to the RVE

(see below), yet they do not provide direct volumetric information.

In nuclear medicine, volumetry based on radiological imaging is used to determine the total volume of organs and tumors, because CT, MRI or sonography can be regarded as the ground truth for such purposes. However, in certain conditions, volume calculations based on nuclear medicine images can be helpful to quantify the active tumor volume it this is associated with the patient's prognosis (e. g., in neuroendocrine tumors [10], lymphoma [11] or prostate cancer [12, 13]) or if there is a discrepancy between the morphological and the metabolic response to therapy (e. g., in lymphoma after chemotherapy [14]). Another application of volume calculation from nuclear medicine images is the preoperative volume estimation of viable liver tissue before partial hepatectomy.

Methodology

We framed the review by going through literature on volume estimation for clinical applications that are closely related to daily routine in nuclear medicine: tumor, liver and thyroid volumetry. For these means, a search on PubMed was performed combining three elements (1) application, (2) task and (3) tool on September 1st, 2023. For the application, the terms "tumor", "liver" and "thyroid" were used, for the task, we applied the keywords "volumetry", "volume estimation" and "volume calculation", while for the tool "machine learning", "deep learning" or "neural network" were selected. As an example, one possible search was: ("Tumor" AND "Volumetry" AND "Machine learning"). All 27 retrieved combinations were evaluated.

We considered any paper providing volumetry data for this review, including publications on radiological examinations due to their relevance for nuclear medicine procedures, as described above. Additionally, works not including volumetry evaluations, but only segmentation results were removed, as well as papers not using ML approaches.

In total, 32 hits were obtained for tumor, 38 for liver and 4 for thyroid. From those 5, 11 and 4 papers were deemed relevant for this review. Hits were removed in cases of repetitions, phantom or preclinical studies, lack of evaluation of volumetry, or if the hit did not provide clinical information. Two publications on tumor volumetry included data on kidney volume, and we decided to address kidney volumetry in a separate chapter.

Results

Tumor volumetry

Calculating the volume of a tumor can be crucial in different scenarios. Tumor volume (in particular, change in volume) plays a role in staging, therapy selection and follow-up/prognosis.

Gutsche et al. propose to use an nnUnet [15] to segment the metabolic tumor volume in O-(2-[^{18}F]fluoroethyl)-L-tyrosine ([^{18}F]FET) PET images of patients with glioblastoma [16]. They used 399 manually annotated patients for training and 156 patients for testing. To evaluate volumetry, they evaluated the correlation between the automatically obtained volumes and those provided by the experts, which ranged from 0.92 to 0.95 and was considered acceptable.

Vivanti et al. used a pretrained convolutional neural network (CNN) to segment liver tumors inside a region of interest (ROI) from a CT image [17]. They used the volumes derived from manually segmented tumors of 37 patients for comparison. Unfortunately, the CNN design was not described in detail, yet the mean volumetric percentage error was reported to be 16.0 % and 17.9 % with a standard deviation of 10.0 %.

The group of Shapey et al. proposed the use of Unet [18] with a supervised attention module for segmentation and subsequent volumetry of vestibular schwannomas in 246 patients [19]. They evaluated the use of contrast-enhanced T1-weighted MRI or high-resolution T2-weighted MRI as input. They split the dataset in 200 training patients and 46 test patients and evaluated the Unet with and without the supervised attention module, as well as using a hardness-weighted Dice loss or the common Dice loss. The relative volumetric error for their networks is reported to be between 6.96 % ± 5.68 % and 15.98 % ± 14.65 % depending on the sequences used for training, the architecture and the loss function. Their most sophisticated model obtained an average error of 7.03 % ± 5.04 % on both T1 and T2 images.

In the work by Kang et al., the authors evaluated six different Unet architectures for the segmentation and volume estimation of meningiomas in T1-weighted contrast-enhanced MRI [20]. The authors report the average volume as obtained from manually segmented images by two experts for an internal dataset as well as a multi-institution cohort. They trained 2D, 3D and attention Unets [21] and their nnUnet variants with 489 patients. Both evaluation cohorts included 100 patients. The expert tumor volumes were on average 2.46 ml ± 5.18 ml, while the networks yielded 5.77 ml ± 13.90 ml in the internal dataset. In the external cohort, the results were more consistent with 2.71 ml ± 4.94 ml vs. 3.51 ml ± 8.29 ml. The differences between networks were moderate; only the 3D Unet performed significantly worse than the other architectures.

In a different task, Toyonaga et al. used volumetry to evaluate the quality of a CNN-based attenuation map calculation in oncological patients undergoing PET with three tracers ([^{18}F]FDG, [^{68}Ga]Ga-DOTA-octreotate ([^{68}Ga]Ga-DOTATATE) and [^{18}F]Fluciclovine) [22]. The PET images were reconstructed using the synthetic attenuation maps as well as the real low-dose CT. A pretrained CNN [23] was used to automatically extract the metabolic tumor volume of 246 patients. The resulting segmentation masks were compared with the ones manually delineated by two experts. The correlation between volumes was between 0.99 and 1.00, and the volumetric percentage error assuming the manual segmentations to be true was on average between 1.8 %-6.4 % with a standard deviation of 15.0 %.

▶ **Table 1** summarizes the findings from the publications in this section. In conclusion, tumor volumetry using the sequential approach of segmentation followed by volume calculation yields average relative errors below 20 %. From a clinical perspective, 20 % is on the upper bound of acceptance, yet the major advantage of automatic ML methods is the fact that they can better deal with shapes that differ from ellipsoids and take significantly

▶ **Table 1** Overview of results for tumor volumetry. ML = machine learning, ce = contrast-enhanced, hr = high resolution, CNN = convolutional neural network, ICC = interobserver correlation coefficient, RVE = relative volumetric error (%).

Author	Modality	Train	Test	Tumor Entity	ML	Criterion	Value
Gutsche	[^{18}F]FET PET	156	399	Gliobastoma	nnUnet	ICC	0.92–0.95
Vivanti	CT			Liver	CNN	RVE	16.0 %–17.9 % ± 10.0 %
Shapey	MR ceT1, hrT2	200	46	Vestibular schwannoma	2.5 D Unet with attention	RVE	6.96 %–15.98 % ± 14.65 %
Kang	MR ceT1	459	200	Meningioma	nnUnet and Unet (2D, 3D, attention)	Volume cohort 1 (manual vs. ML)	2.46 ml ± 5.18 ml vs. 5.77 ml ± 13.90 ml
						Volume cohort 2 (manual vs. ML)	2.71 ml ± 4.94 ml vs. 3.51 ml ± 8.29 ml
Toyonaga	[^{68}Ga]Ga-DOTATATE PET, [^{18}F]Fluciclovine PET, [^{18}F]FDG PET	120	246	Several	3D Unet	ICC	0.99–1.00
						RVE	1.8–6.4 % ± 15.0 %

less time than manual delineation. In all but one reviewed article, the tumor was directly segmented without the need of the determination of a region-of-interest (ROI) or the input of a seed prompt by the user.

Overall, the reviewed works have reported a high correlation between the ML methods and experts (0.92–1.00), while the relative volumetric error ranges between 6.96 % and 17.9 %. When judging these results, it is worth noting that tumor volumetry is a more difficult task than organ volumetry – even for experts.

Liver volumetry

Liver volumetry plays a role in surgical scenarios, such as planning a partial hepatectomy or in liver transplantation, yet it is also of relevance in the monitoring of liver diseases and in the context of nuclear medicine for selective internal radiotherapy (SIRT).

Several of the works listed below have used data from different medical image processing challenges as training dataset or as benchmark, such as the MICCAI-Sliver07 dataset [24], the 3Dircadb1 dataset [25] or the Liver Tumor Segmentation Benchmark (LiTS) dataset [26].

Takamoto et al. evaluated the performance of the commercial software Synapse 3D version 6 by Fujifilm to calculate the volume of the liver and liver segments in contrast-enhanced CT [27]. They used the volumes derived from the manual segmentation performed by a surgical specialist and reviewed by an expert as ground truth. The liver volumes showed a correlation of 0.98 (with the 95 % confidence interval (CI) between 0.98 and 0.99), while the average volumetric error was 69.3 ml ± 46.5 ml (95 % CI, 61.6 ml – 77.0 ml). This corresponds to a percentage error of 6.70 % ± 4.49 % (95 % CI, 5.95 % – 7.43 %).

Lu et al. propose the sequential use of a 3D CNN and graph cuts for refining the segmentation of the liver in contrast-enhanced CTs [28]. They trained on data of 68 in-house patients and 10 patients from the MICCAI-Sliver07 dataset. They benchmarked their results using 10 other patients from the MICCAI-Sliver07 dataset and 20 patients from the 3Dircadb1 dataset. The mean relative volumetric error was reported to be 2.70 % and 0.97 % for the two test sets. Combining both datasets (10 + 20 patients), the authors report an ICC of 0.931.

In a study of 197 in-house patients and 131 patients from the LiTS challenge, Marinelli et al. used a 3D-Unet to estimate the volume of the liver for multiple pathologies in contrast-enhanced CT scans [29]. They used natural language processing tools to extract the liver volume from reports, which they regarded as the ground truth. They compared the absolute error of four models, one trained on the LiTS dataset and three including either five “difficult” in-house patients where the LiTS model underestimated or overestimated the volume, or using all 10 hard cases. In these cases, they generated the masks manually. They report an improvement of the error from 231 mL (95 % CI: 202, 260) to 183 mL (95 % CI: 160, 206), 216 mL (95 % CI: 186, 247), and 176 mL (95 % CI: 154 198), respectively, when compared to the volumes obtained from the medical reports.

Koitka et al. used a multiresolution 3D Unet (MultiResUNet) [30] to segment and estimate the volume of each lobe of liver from contrast-enhanced CT [31]. The training set consisted of 100 in-house patients and the test set of additional 30 patients from a different hospital. The RVE was 2.19 %± 1.40 %, and the AVE 32.12 ± 19.40 ml. For both lobes the RVE was reported to be 2.22 %± 2.30 % and 2.07 %± 5.52 % (right and left).

The team of Shin et al. took volumetry for patients with autosomal dominant polycystic kidney disease under consideration [32]. In a first stage, they trained a multiorgan segmentation model based on Vnet [33] with 153 native CTs and 22 contrast-enhanced ones from five hospitals. The correlation for the AI-calculated combined volume of liver and kidney was 0.9997 with respect to three experts, and in only 5.1 % of the 39 test cases the RVE was above 5 %. They then evaluated the ICC of their AI versus 11 experts. Except for one human rater the ICC ranged between 0.966 and 0.999, while the AI rated 0.992. The one expert showing an ICC of 0.897, a clear outlier from the data, had 9 years of experience reporting on patients with the same pathology. This points out at the fact that human experts do not necessarily provide the ultimate ground truth.

In a similar application, Cayot et al. evaluated the automatic segmentation and volume estimation of polycystic livers training on 64 patients and evaluating on additional 24 patients [34]. They used a 3D Unet approach using as ground truth the delineation by an expert getting an RVE of 1.1 %± 4.0 % for the AI. The intra- and interobserver variability for the experts was assessed to be 0.6 % ± 2.1 ± and 2.8 %± 3.8 % respectively. The concordance correlation coefficient for the AI, the same expert (intra) and another expert (inter) were 0.995 (95 % CI, 0.991–0.997), 0.998 (95 % CI, 0.997–0.999) and 0.994 (95 % CI, 0.990–0.997).

Ng et al. evaluated both a Gaussian mixture model (GMM) and a Unet for the liver segmentation and subsequent volumetry in contrast-enhanced images from a spectral detector CT (SDCT) [35]. They trained their model using 30 patients with healthy livers using a 5-fold cross validation scheme. They compared volume estimation also using the spectral data or not (i. e., using only the Hounsfield units image). The GMM showed better results than the Unet with respect to the manual annotation by experts in both SDCT and ceCT (9.2 %± 3.4 % and 5.9 %± 12.8 % vs. 10.2 % ± 131.% and 15.3 %± 16.1 % respectively). The size of the dataset is limited such that a strong statement is not possible, yet the contribution of the spectral information seems to be beneficial for the Unet while it did not contribute to the GMM.

For MRI, Chelus et al. propose a semi-automatic approach that combines a CNN with manual corrections [36]. They use an in-house dataset of 83 patients where dynamic contrast-enhanced (DCE) MRI images were acquired as part of the SIRT of primary or metastatic liver cancer. Sixty-two patients were used for training and validation, and 21 patients for testing. Three 2D-CNNs for axial, coronal and sagittal images were combined by majority voting. The manual annotations of an expert were used as ground truth, while the method was evaluated with and without manual corrections and against a radiologist and two residents. Without corrections, the CNN ensemble yielded a relative error of 4.5 % ± 3.5 % while the human readers achieved 3.6 %–5.8 %± 5.0 %. The collaborative approach (i. e., ML with human correction) reduced the error to 3.1 %-3.5 %± 2.2 %. The interobserver variability of the error of the human rater was 2.8 %± 1.4 %, which was higher than with the collaborative approach (0.7 %± 0.9 %).

Winther et al. studied the ICC of the liver volume of a 3D Unet with respect to experts in a dataset of 100 patients undergoing contrast-enhanced T1-weighted volumetric interpolated breath-hold examination (VIBE) MRI [32]. They trained with 75 patients and evaluated in 4-folds cross validation setup. The human inter-reader volumetry had an ICC of 0.973, while the 3D Unet obtained 0.987. Other metrics used such as the DSC also were favorable for the AI-method.

In the scope of autosomal dominant polycystic kidney disease (ADPKD), Woznicki et al. trained an nnUnet on a large dataset of 327 patients from multiple hospitals in Germany and the Netherlands to segment liver and kidneys. They verified their model on a separate internal 93-patient dataset, as well as an external one including 323 patients scanned with multiple devices and varied image protocols, yet only including kidneys. The training input was MRI images of different T2-weighted sequences (turbo-spin echo (TSE), spectral presaturation with inversion recovery (SPIR), mapping, half-fourier-acquired single-shot turbo spin echo (HASTE) and true fast imaging with steady-state free precession (TRUFI)) in 2D (coronal and axial) as well as 3D. From the 323 multiple vendor dataset, the authors used only T2-weighted images including sequences that were not used during training. In the internal datasets, the ICC for kidney and liver was between 0.996 and 0.999 with an RVE of 0.5 % (limits of agreement, LoA -8.7 %, 8.2 %). In the external kidney dataset the RVE was 1.3 % (LoA -15.9 %, 13.2 %) using both planes.

Trying to understand which MRI sequence is best for AI-supported segmentation and volume calculation, Saunders et al. trained multiple 3D Unets [37]. 42 obese adolescents underwent an MRI scan with multiple sequences (water, fat, T2*, true fast imaging with steady state precession (TrueFISP), HASTE) and the liver was segmented manually. The Unets were trained using 5-fold cross-validation and consisted of five single channel Unets (for all 5 sequences), one two-channel sequence (water+fat) and one three-channel one (water, fat, T2*). They reported the normalized root mean squared error (NRMSE) to range between 4.23 % for the two-channel 3D Unet to 6.82 % for the T2*-Unet. The AVE ranged from 7.7 ml for the three-channel Unet to 41.5 ml for the TrueFISP-Unet.

In an approach to offer a multimodal liver segmentation method, Wang et al. used a 2D-Unet trained on multiecho spoiled gradient-echo (SPGR) MRI (both 2D and 3D for variable T2-weighting), contrast-enhanced T1-weighted MRI (ceT1) and contrast-enhanced CT (ceCT) [38]. They followed a sequential training starting with 300 SPGR MRI patients, and then refined the model with additional 30 SPGR MRI, 20 ceT1 and 10 ceCT datasets. For evaluation, they used 230 CT (also unenhanced) and 100 ceT1 patients. As ground truth, experts were involved yielding a correlation of 0.95 for the test CT dataset with an average absolute error of –58.1 mL (95 % CI -298, 180), i. e., a relative error of –3.5 % (95 % CI –17.8, 10.7). For the ceMRI, the same metrics were 0.98, –89 mL (95 % CI –358, 180), and -4.0 % (95 % CI –16.1, 8.1).

Liver volumetry seems to be a widely investigated topic with initial public challenges dating as far back as 2007. For example, the results of the LiTS challenge show a relative volumetric error in the range of 0 % to 49.6 % for a total of 228 participants at the time this review was written [26]. Bilic et al. show that with increasing quality of annotations and improved methods, the performance in the latest challenge from 2018 has already improved significantly; the best three teams achieved an error below 2 % [39]. Such improvement in performance can also be observed when analyzing the submissions to the challenges over the years. However, a lack of harmonized data, wrong annotations (e. g., [32]) or missing information with respect to the scanner or imaging protocols that were used or on patient demographics limit the possibility to include such information in the ML methods, making it harder to obtain generalizable models that could be applied to a wide range of scenarios.

In summary, the papers on liver volumetry reviewed for this work reported absolute (and relative) errors ranging between 58 and 231 mL (0.97 and 15.4 %). All but one of the articles yield a percentage error below 10 %. However, in the only publication that reported relative errors above 10 %, the authors did not include manually annotated volumes but derived them from clinical

▶ **Table 2** Overview of results for liver volumetry. ML = machine learning, ceCT = contrast-enhanced CT, nCT = native CT, ceT1 = contrast-enhanced T1-weighted MRI, ceVIBE = contrast-enhanced T1-weighted VIBE MRI, DCE = dynamic contrast-enhancement, SPGR = multiecho spoiled gradient-echo, mCRC = metastatic colorectal cancer, ADPKD = autosomal dominant polycystic kidney disease. CNN = convolutional neural network, GMM = Gaussian Mixture Model, AVE = absolute volumetric error in mL, RVE = relative volumetric error in percentage, CCC = concordance correlation coefficient, CI = 95 % confidence interval. * RVE is not reported, but calculated from AVE assuming an average liver volume of 1500 mL.

Author	Modality	Train	Test	Indication	ML	Criterion	Value
Takamoto	ceCT	n/a	144	mCRC, Cholangio-carcioma, other liver metastases	Synapse 3D	ICC	0.98 (CI 0.98, 0.99)
						RVE	6.70 % ± 4.49 % (CI 5.95 %, 7.43 %)
Marinelli	CT	141	187	Various liver diseases	3D Unet	AVE	183–231 mL
						RVE*	12.2–15.4 %
Lu	ceCT	78	30	Various, non-tumorous anomalies and healthy	3D-CNN + graph cut	ICC	0.931
						RVE	0.97–2.70 %
Shin	nCT, ceCT	153	39	ADPKD	Vnet	ICC	0.9997
			50			ICC (AI vs. 11 experts)	0.992 vs. 0.897–0.999
Cayot	nCT, ceCT	64	24	Polycystic liver, hepato-renal polycystic disease	3D Unet	RVE	1.1 % ± 4.0 %
						CCC	0.995 (CI, 0.991–0.997)
Ng	ceSDCT, ceCT	25	5	Healthy volunteers	2D Unet	RVE	10.2–15.3 % ± 16.1 %
					GMM		5.9–9.2 % ± 12.8 %
Chlebus	DCE MRI	62	21	Primary and metastatic liver cancer	Ensemble 2D-CNN + manual correction	RVE	3.1–3.5 %± 2.2 %
						Interobserver variability	0.7 ± 0.9 %
Winther	ceVIBE	75	25	Various, also non oncological	3D Unet	ICC	0.973 vs. 0.987
Saunders	5 MRI sequences	32	10	Obese adolescents	3D Unet	NRMSE	4.23–6.82 %
						AVE	7.7–41.5 mL
Woznicki	Several T2-MRI	327	93	ADPKD	nnUnet	ICC	0.996–0.999
						RVE	–0.5 % (LoA –8.7 %, 8.2 %)
Wang	SPGR MRI, ceT1 MRI, ceCT, nCT	360	330	Various, also non oncological	2D Unet	ICC	0.95–0.98
						AVE	58–89 mL
						RVE	3.5–4.0 %

reports [29]. A high ICC with experts was reported (0.93 to 0.99), and interobserver variability was as low as 0.7 % (▶ **Table 2**).

Kidney volumetry

Several diseases, such as polycystic kidney disease or renal tumors, result in an abnormal size and shape of the kidneys. In certain conditions, the total size of the kidney can be a good indicator of disease progression or recurrence. In the context of nuclear medicine, kidney volumetry can be relevant for dosimetry in peptide receptor radionuclide therapy [40].

In a work of 2022, Hsiao et al [41] used a Unet with a ResNet-41 as encoder to calculate the total kidney volume (TKV), i. e., the sum of the size of both kidneys including tumors and cysts. They trained their networks on 210 publicly available cases from the Kidney and Kidney Tumor Segmentation Challenge (KiTS) dataset [42] using 5-fold cross-validation. To evaluate its performance, they compared it to the conventional ellipsoid formula and a ground truth (manual annotations) in 10 cases. Their method showed an average percentage error of 1.43 % (95 % CI: 0.40, 2.47) vs. the 10.5 % (95 % CI: 6.6, 14.3) error of the ellipsoid method.

In a different approach, Hussain et al. presented a cascaded regression neural network (a CNN-guided Mask-RCNN) for segmentation-free volume estimation [9]. Their network was trained/validated on 160/15 patients of the KiTS dataset, or alternatively 65/10 from an in-house dataset. They used the remaining 35 KiTS and 25 in-house patients for performance evaluation. The mean percentage error was 4.80 %± 3.89 % for the in-house dataset and 7.26 %± 6.80 % for the 35 KiTS patients. When evaluating the correlation between their method and the ground truth, they report ICC of 0.964 and 0.971 (Student's t-test: p-values of 0.904 and 0.752 for both datasets).

Interestingly, the authors derive an empirical formula to relate the well-established Sørensen-Dice coefficient (DSC) to the relative volume error, namely:

$$RVE\ (\%) \approx \left| \frac{2}{DSC} - 2 \right| \times 100$$

Assuming that this formula would prove to be sufficiently accurate, most of the results on organ segmentation from the literature could be quantified in terms of volume. However, such an endeavor is out of the scope of this review.

In this review, no explicit search was performed for ML papers focusing on kidney volumetry (▶ **Table 3**). However, we regard especially the work of Hussain et al. [9] as relevant because it proposes a different approach and provides an approximation for converting Dice scores to relative volumetric error.

The range of error for kidney volume calculation of 1.43 % and 7.26 % that we derived from this small subset of articles is similar to the error reported for liver volumetry. Correlation with experts is also documented to be high (0.964–0.971).

Thyroid volumetry

The thyroid volume is estimated often as part of the clinical routine for the diagnosis, treatment and monitoring of thyroid diseases. Conventionally, the thyroid is imaged using ultrasound (US) because of its limited costs and the lack of ionizing radiation. However, the interpretation of US images is challenging and poses some limitations to the accuracy of thyroid volumetry showing errors ranging between 1.1 % to up to 22.7 % [43, 44].

Chang et al. [45] proposed a combination of a radial basis function neural network (RBF-NN) [46] and the particle swarm (PSO) algorithm [47, 48] to perform thyroid volumetry from 2D B-mode US. First, they acquire 2D US, and the image quality is enhanced using conventional image processing methods. During training, the RBF-NN learns to classify small regions of interest extracted from the US images as “thyroid” or “non-thyroid”. At the test phase, by predicting the class of the patches, the RBF-NN returns an approximated thyroid segmentation which is then refined using a region growing method. Finally, they use the particle swarm optimization (PSO) algorithm to optimize the parameters of the thyroid volume estimation formula, so that its result is as close as possible to the volume computed from a CT scan. They report an AVE of 0.69 ± 0.61 ml, i. e., an RVE of 3.74 ± 2.00 % for a test cohort of five patients.

Kumar et al. [49] employed a multi-prong convolutional neural network (MP-CNN) [50] to segment the thyroid gland, thyroid nodules and cysts in 2D B-mode US. They reported a DSC of 78 % for the three anatomies and an RVE in the estimation of the volume of the nodule of 7.47 %, where the volume calculated by a board certified radiologist was used as reference.

In 2018, Poudel et al. compared three non-automatic methods, namely active contours without edges (ACWE) [51], graph cut (GC) [52] and pixel-based classifier (PBC) [45], and two automatic ones, namely a random forest classifier (RFC) [53, 54] and a Unet [18] CNN, for thyroid segmentation. A manually annotated label map was used as reference. They reported DSC of 0.80, 0.77 and 0.67 for ACWE, GC and PBC, respectively, which are below the required accuracy for clinical practice. On the other hand, the automatic methods, namely RFC and Unet, resulted in a DSC of 0.86 and 0.87. The non-automatic methods used 2D B-mode US images, whereas the RFC and the Unet used a 3D B-mode US volume. The segmentation resulting from the non-automatic methods were then compounded using ImFusion and MeVisLab to evaluate the volumetry. Consistently with the results for the DSC, ACWE was reported to be the best method with an RVE of 17.90 %.

Krönke et al. [55] presented a comparison between thyroid volumetry based on 2D US and 3D US in 28 healthy volunteers. They also acquired MR (T1 VIBE) to compute the reference volume of the thyroid. The volumetry on 2D US was performed using the ellipsoid formula [56], whereas 3D US were segmented by a NN (namely QuickNat) [57]. The label map was used to derive the volume of the thyroid. They reported an intraobserver variability of 16.67 %± 6.66 % and an interobserver variability of 4.86 %± 2.83 % (over three 3D US acquisitions between three physicians). The 3D-US-based volumetry was reported to be more accurate than the 2D-US-based approach, with a percentage error of 4.14 % ± 7.32 % versus 26.95 %± 14.95 % if compared to the MR-based volumetry. The difference between the volumes computed with 3D US and MRI was not significant.

To summarize the results on thyroid volumetry, the best performing method with respect to the DSC was the CNN (DSC = 0.87), and the lowest performance was obtained with the GC (DSC = 0.67), both reported by Poudel et al. In terms of RVE in 2D US, we observed a high variability ranging from a RVE of 3.74 ± 2.00 % (Chang et al.) to a RVE of 26.95 ± 14.95 % (Krönke et al.) (▶ **Table 4**). 3D US appears to improve the accuracy of thyroid volumetry (RVE < 5 %) which has been also confirmed in an experimental robotic ultrasound setup [58]. However, the current literature is too limited to draw a definite conclusion.

Discussion

Volumetry constitutes a critical tool for various clinical applications, aiding in diagnosis, prognostication, treatment planning, and evaluation of response to therapy. It holds particular significance in oncology for staging, evaluating treatment efficacy, and predicting patient outcomes. In particular, volumetry has shown good agreement with manual metrics such as Response Evaluation Criteria in Solid Tumors (RECIST) [59, 60] as well as a prognostic relevance if combined with uptake intensities in nuclear medicine images (e. g., [10]). This opens the possibility to evaluate disease progression in a more reproducible and consistent way [61]. In the realm of radionuclide therapy, volumetry has played a relevant role in dosimetry, and it will likely gain importance with the increasing relevance of personalized dose planning [62, 63]. Finally, in endocrinology, thyroid volumetry is expected to continue to be an integral part of the diagnostic workflow of goiter and Graves’ disease [64].

The advent of AI, and more specifically DL, has brought forth advancements in automating and thus significantly speeding volume calculations, thereby mitigating variability, reducing labor and enhancing accuracy. The adaptability and efficiency of AI

▶ **Table 3** Overview of results for kidney volumetry. ML = machine learning, ceCT = contrast-enhanced CT, ccRCC = clear cell renal cell carcinoma, ADPKD = autosomal dominant polycystic kidney disease, CNN = convolutional neural network, AVE = absolute volumetric error in mL, RVE = relative volumetric error (%), ICC = interobserver correlation coefficient, 95 % CI = 95 % confidence interval, LoA = limits of agreement.

Author	Modality	Train	Test	Indications	ML	Criterion	Value
Hsiao	ceCT	200	10	Renal malignancies, mainly ccRCC	Res-Unet	AVE	6.49 ml (CI: 1.80, 11.19)
						RVE	1.43 % (CI: 0.40, 2.47)
Hussain	ceCT	250	60	Renal malignancies and healthy	CNN-guided Mask-RCNN	ICC	0.964–0.971 (p-value 0.904–0.752)
						RVE	4.80 %-7.26 %± 6.80 %
Woznicki	Several T2-MRI	327	93	ADPKD	nnUnet	RVE	−1.3 % (LoA −15.9 %, 13.2 %)

▶ **Table 4** Overview of results for thyroid. ML = machine learning, AVE = absolute volumetric error in mL, RVE = relative volumetric error in percentage, RBF-NN = Radial base function neural network, PSO = particle swarm algorithm, MP-CNN = multi-prong convolutional neural network, QuickNAT = Quick segmentation of NeuroAnaTomy, ACWE = active contours without edges, GC = graph cut, PBC = pixel-based classifier, RFC = random forest classifier, CNN = convolutional neural network.

Author	Modality	Train	Test	Indications	ML	Structure	Criterion	Value
Chang	2D US	not reported	5	Thyroid nodule patients	RBF-NN + PSO	Thyroid gland	AVE vs. CT	0.69 ± 0.61 ml
							RVE vs. CT	3.74 ± 2.00 %
Kumar	2D US	186	48	Thyroid nodule patients	MP-CNN	Thyroid gland	DSC vs. expert	0.78
			5			Nodule	AVE vs. expert	1.40 ± 1.27 ml
							RVE vs. expert	8.13 ± 7.11 %
Krönke	2D US	15	13	Healthy volunteers	(Ellipsoid formula)	Thyroid gland	RVE vs. MRI	26.95 ± 14.95 %
							Intraobserver variability	15.33 ± 3.21 %
							Interobserver variability	12.92 ± 6.32 %
	3D US				QuickNAT		RVE vs. MRI	4.14 ± 7.32 %
							Intraobserver variability	16.67 ± 6.66 %
							Interobserver variability	4.86 ± 2.83 %
Poudel	2D US	not reported	1416 (2D US images)		ACWE	Thyroid gland	DSC vs. expert	0.8
							AVE vs. expert	2.33 ml
							RVE vs. expert	17.09 %
					GC		DSC vs. expert	0.77
							AVE vs. expert	2.61 ml
							RVE vs. expert	19.19 %
					PBC		DSC vs. expert	0.67
							AVE vs. expert	4.09 ml
							RVE vs. expert	30.01 %
					RFC		DSC vs. expert	0.86
					CNN		DSC vs. expert	0.87

have proven superior in several comparative studies presented in this review.

The surveyed literature for tumor volumetry demonstrates a high correlation (0.92–1.00) between ML methods and expert assessments, with a relative volume error spanning 6.96 % to 17.9 %. Given the small number of studies explicitly evaluating tumor volumetry, and the heterogeneity of them, it is difficult to extrapolate the results to other applications. However, tumor volumetry seems to be a more challenging task than organ volumetry both for experts and ML methods based on the results that we obtained from articles on liver, kidney or thyroid volumetry. Considering the importance of the (metabolic) tumor volume as a

biomarker to predict outcome and select therapies (e. g., in diffuse large B-cell lymphoma [65]), we assume that more groups will focus on AI-based volumetry. Since hybrid scanners for single photon emission computed tomography (SPECT)/CT, PET/CT and PET/MRI have become the clinical standard, nuclear medicine physicians might be increasingly exposed to organ or lesion volumetry in the future.

In contrast to tumor volume estimation, liver volumetry has been extensively studied, revealing a progression in accuracy and a decrease in volumetric error over time. Bilic et al. reported a significant improvement in performance, with the best algorithms achieving an error below 2 % in the 2018 LiTS challenge [39]. Our review indicates a high ICC (0.93 to 0.98) with expert assessments and a consistent decrease in errors over time, mostly below 10 %. However, a lack of standardized patient selection and image acquisition protocols, harmonized data and specific scanner and patient information limits the generalizability of ML methods across diverse scenarios. This problem is examined in a dedicated review on harmonization and standardization by Fuchs et al. in this same Special Issue. Still, if a relative error below < 10 % is deemed acceptable, ML-based liver volumetry can be considered reliable for applications like SIRT planning or routine follow-up of non-oncological hepatic pathologies. However, routine clinical use of such methods should only be considered following extensive evaluation/validation with in-house data to ensure that there is no relevant domain gap between the (external) training data and the (local) clinical data.

Our systematic search for publications did not specifically cover ML applications for kidney volumetry. However, based on the two articles that we examined, the observed error range (1.43 %–7.26 %) was comparable to that obtained in liver volumetry, suggesting consistency in volumetric calculations across organs. A high correlation (0.964–0.971) with expert evaluations further substantiates the reliability of the ML methods in kidney volumetry.

A particularly relevant finding is the approximation formula to estimate relative volumetric error from the Dice score. Using this formula, the majority of reports on organ segmentation could be quantified in terms of volume, and as result, the volumetric performance could be estimated when selecting segmentation models.

Despite the high frequency of such examinations in routine clinical care, it seems that AI-supported thyroid gland and nodule volumetry has not yet been extensively developed or evaluated. The availability of low-priced 3D US systems (e.g., piur imaging's tUS) and the possibility of reimbursing them will most likely change this situation in the midterm as the potential gains in labor and reproducibility are evident.

As a summary, volumetry is an integral source of information for nuclear medicine experts when diagnosing, evaluating and planning therapies in oncology and endocrinology. AI can provide fast and reproducible volume estimations, and we assume that such methods will be incorporated more and more in the routine workflows as a result of the availability of ever improving segmentation tools.

However, the wide range of performance in volumetry reported in the reviewed articles underlines the paramount relevance of the domain gap (i. e., the difference in probability distribution of training and inference data) before clinical deployment. After all, AI should not only streamline clinical work, but also keep it accurate and reliable to ensure a true improvement in patient care.

Funding

This article was partially funded by the Deutsche Forschungsgemeinschaft NA 620/51–1 grant.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Riveira-Martin M, Akhavanallaf A, Mansouri Z. et al. Predictive value of 99mTc-MAA-based dosimetry in personalized 90Y-SIRT planning for liver malignancies. EJNMMI Res 2023 13: (1): 63

[2] Krönke M, Eilers C, Dimova D. et al. Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. PLOS ONE 2022 17: (7): e0268550

[3] Sunoqrot MRS, Selnæs KM, Sandsmark E. et al. The Reproducibility of Deep Learning-Based Segmentation of the Prostate Gland and Zones on T2-Weighted MR Images. Diagn Basel Switz 2021 11: (9): 1690

[4] Ding J, Cao P, Chang HC. et al. Deep learning-based thigh muscle segmentation for reproducible fat fraction quantification using fat–water decomposition MRI. Insights Imaging 2020 11: (1): 128

[5] Seifert P, Ullrich SL, Kühnel C. et al. Optimization of Thyroid Volume Determination by Stitched 3D-Ultrasound Data Sets in Patients with Structural Thyroid Disease. Biomedicines 2023 11: (2): 381

[6] Brunn J, Block U, Ruf G. et al. [Volumetric analysis of thyroid lobes by real-time ultrasound (author's transl)]. Dtsch Med Wochenschr 1946 1981 106: (41): 1338–1340

[7] Campennì A, Avram AM, Verburg FA. et al. The EANM guideline on radioiodine therapy of benign thyroid disease. Eur J Nucl Med Mol Imaging [Internet] 2023 doi:10.1007/s00259-023-06274-5

[8] Muggli D, Müller MA, Karlo C. et al. A simple method to approximate liver size on cross-sectional images using living liver models. Clin Radiol 2009; 64 (7): 682–689

[9] Hussain MA, Hamarneh G, Garbi R. Cascaded Regression Neural Nets for Kidney Localization and Segmentation-free Volume Estimation. IEEE Trans Med Imaging 2021; 40 (6): 1555–1567

[10] Weber M, Telli T, Kersting D. et al. Prognostic Implications of PET-Derived Tumor Volume and Uptake in Patients with Neuroendocrine Tumors. Cancers 2023; 15 (14): 3581

[11] Mikhaeel NG, Heymans MW, Eertink JJ. et al. Proposed New Dynamic Prognostic Index for Diffuse Large B-Cell Lymphoma: International Metabolic Prognostic Index. J Clin Oncol Off J Am Soc Clin Oncol 2022; 40 (21): 2352–2360

[12] Gafita A, Calais J, Grogan TR. et al. Nomograms to predict outcomes after 177Lu-PSMA therapy in men with metastatic castration-resistant prostate cancer: an international, multicentre, retrospective study. Lancet Oncol 2021; 22 (8): 1115–1125

[13] Seifert R, Herrmann K, Kleesiek J. et al. Semiautomatically Quantified Tumor Volume Using 68Ga-PSMA-11 PET as a Biomarker for Survival in Patients with Advanced Prostate Cancer. J Nucl Med Off Publ Soc Nucl Med 2020; 61 (12): 1786–1792

[14] Hutchings M, Barrington SF. PET/CT for therapy response assessment in lymphoma. J Nucl Med Off Publ Soc Nucl Med 2009; 50 (Suppl. 1): 21S–30S

[15] Isensee F, Jaeger PF, Kohl SAA. et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021; 18 (2): 203–211

[16] Gutsche R, Lowis C, Ziemons K. et al. Automated Brain Tumor Detection and Segmentation for Treatment Response Assessment Using Amino Acid PET. J Nucl Med Off Publ Soc Nucl Med 2023. doi:10.2967/jnumed.123.265725

[17] Vivanti R, Szeskin A, Lev-Cohain N. et al. Automatic detection of new tumors and tumor burden evaluation in longitudinal liver CT scan studies. Int J Comput Assist Radiol Surg 2017; 12 (11): 1945–1957

[18] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, et al. Medical Image Computing and Computer-Assisted Intervention – MICCAI. Cham: Springer International Publishing; 2015: 234–241

[19] Shapey J, Wang G, Dorent R. et al. An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced T1-weighted and high-resolution T2-weighted MRI. J Neurosurg 2019; 134 (1): 171–179

[20] Kang H, Witanto JN, Pratama K. et al. Fully Automated MRI Segmentation and Volumetric Measurement of Intracranial Meningioma Using Deep Learning. J Magn Reson Imaging JMRI 2023; 57 (3): 871–881

[21] Oktay O, Schlemper J, Folgoc LL. Attention U-Net: Learning Where to Look for the Pancreas. In: Conference Book [Internet]. Amsterdam: Radboud University Medical Center; 2018 [cited 2023 Sep 21]. 15 Available from: https://openreview.net/pdf?id=Skft7cijM

[22] Toyonaga T, Shao D, Shi L. et al. Deep learning-based attenuation correction for whole-body PET – a multi-tracer study with 18F-FDG, 68 Ga-DOTATATE, and 18F-Fluciclovine. Eur J Nucl Med Mol Imaging 2022; 49 (9): 3086–3097

[23] Hirata K, Furuya S, Huang SC. et al. A semi-automated method to separate tumor from physiological uptakes on FDG PET-CT for efficient generation of training data targeting deep learning. J Nucl Med 2019; 60 (Suppl. 1): 1213–1213

[24] Van Ginneken B, Heimann T, Styner M. 3D segmentation in the clinic: A grand challenge. In: MICCAI workshop on 3D segmentation in the clinic: a grand challenge. 2007: 7–15

[25] Soler L, Hostettler A, Agnus V et al. 3D image reconstruction for comparison of algorithm database. 2010. Available from: https://www.ircad.fr/research/data-sets/liver-segmentation-3d-ircadb-01

[26] CodaLab – Competition [Internet]. [cited 2023 Sep 25]. Available from: https://competitions.codalab.org/competitions/17094#results

[27] Takamoto T, Ban D, Nara S. et al. Automated Three-Dimensional Liver Reconstruction with Artificial Intelligence for Virtual Hepatectomy. J Gastrointest Surg Off J Soc Surg Aliment Tract 2022; 26 (10): 2119–2127

[28] Lu F, Wu F, Hu P. et al. Automatic 3D liver location and segmentation via convolutional neural network and graph cut. Int J Comput Assist Radiol Surg 2017; 12 (2): 171–182

[29] Marinelli B, Kang M, Martini M. et al. Combination of Active Transfer Learning and Natural Language Processing to Improve Liver Volumetry Using Surrogate Metrics with Deep Learning. Radiol Artif Intell 2019; 1 (1): e180019

[30] Ibtehaz N, Rahman MS. MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw Off J Int Neural Netw Soc 2020; 121: 74–87

[31] Koitka S, Gudlin P, Theysohn JM. et al. Fully automated preoperative liver volumetry incorporating the anatomical location of the central hepatic vein. Sci Rep 2022; 12 (1): 16479

[32] Shin TY, Kim H, Lee JH. et al. Expert-level segmentation using deep learning for volumetry of polycystic kidney and liver. Investig Clin Urol 2020; 61 (6): 555–564

[33] Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In IEEE; 2016: 565–571

[34] Cayot B, Milot L, Nempont O. et al. Polycystic liver: automatic segmentation using deep learning on CT is faster and as accurate compared to manual segmentation. Eur Radiol 2022; 32 (7): 4780–4790

[35] Ng YS, Xi Y, Qian Y. et al. Use of Spectral Detector Computed Tomography to Improve Liver Segmentation and Volumetry. J Comput Assist Tomogr 2020; 44 (2): 197–203

[36] Chlebus G, Meine H, Thoduka S. et al. Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections. PloS One 2019; 14 (5): e0217228

[37] Saunders SL, Clark JM, Rudser K. et al. Comparison of automatic liver volumetry performance using different types of magnetic resonance images. Magn Reson Imaging 2022; 91: 16–23

[38] Wang K, Mamidipalli A, Retson T. et al. Automated CT and MRI Liver Segmentation and Biometry Using a Generalized Convolutional Neural Network. Radiol Artif Intell 2019; 1 (2): 180022

[39] Bilic P, Christ P, Li HB. et al. The Liver Tumor Segmentation Benchmark (LiTS). Med Image Anal 2023; 84: 102680

[40] Sundlöv A, Sjögreen-Gleisner K, Svensson J. et al. Individualised 177Lu-DOTATATE treatment of neuroendocrine tumours based on kidney dosimetry. Eur J Nucl Med Mol Imaging 2017; 44 (9): 1480–1489

[41] Hsiao CH, Sun TL, Lin PC. et al. A deep learning-based precision volume calculation approach for kidney and tumor segmentation on computed tomography images. Comput Methods Programs Biomed 2022; 221: 106861

[42] Heller N, Sathianathen N, Kalapara A et al. The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes [Internet]. arXiv; 2020 [cited 2023 Sep 21]. Available from: http://arxiv.org/abs/1904.00445

[43] Licht K, Darr A, Opfermann T. et al. 3D ultrasonography is as accurate as low-dose CT in thyroid volumetry. Nukl Nucl Med 2014; 53 (3): 99–104

[44] Reinartz P, Sabri O, Zimny M. et al. Thyroid volume measurement in patients prior to radioiodine therapy: comparison between three-dimensional magnetic resonance imaging and ultrasonography. Thyroid Off J Am Thyroid Assoc 2002; 12 (8): 713–717

[45] Chang CY, Lei YF, Tseng CH. et al. Thyroid segmentation and volume estimation in ultrasound images. IEEE Trans Biomed Eng 2010; 57 (6): 1348–1357

[46] Ham FM, Kostanic I. Principles of Neurocomputing for Science and Engineering. McGraw Hill 2000: 680

[47] Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings of ICNN'95 – International Conference on Neural Networks [Internet]. Perth, WA, Australia: IEEE; 1995 [cited 2023 Sep 25]. 1942–1948 Available from: http://ieeexplore.ieee.org/document/488968/

[48] Eberhart R, Kennedy J. A new optimizer using particle swarm theory. MHS95 Proc Sixth Int Symp Micro Mach Hum Sci [Internet]. 1995: 39–43 Available from: http://ieeexplore.ieee.org/document/494215/

[49] Kumar V, Webb J, Gregory A. et al. Automated Segmentation of Thyroid Nodule, Gland, and Cystic Components From Ultrasound Images Using Deep Learning. IEEE Access Pract Innov Open Solut 2020; 8: 63482–63496

[50] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions [Internet]. arXiv; 2016 [cited 2023 Sep 25]. Available from: http://arxiv.org/abs/1511.07122

[51] Chan TF, Vese LA. Active contours without edges. IEEE Trans Image Process [Internet]; 2001; 10 (2): 266–277. http://ieeexplore.ieee.org/document/902291/

[52] Rother C, Kolmogorov V, Blake A. 'GrabCut': interactive foreground extraction using iterated graph cuts. ACM Trans Graph [Internet]; 2004; 23 (3): 309–314. doi:10.1145/1015706.1015720

[53] Breiman L. Random Forests. Mach Learn [Internet]; 2001; 45 (1): 5–32. doi:10.1023/A:1010933404324

[54] Criminisi A, Shotton J. Decision Forests for Computer Vision and Medical Image Analysis. Springer Science & Business Media 2013: 367

[55] Krönke M, Eilers C, Dimova D. et al. Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. PloS One 2022; 17 (7): e0268550

[56] Dietlein M, Grünwald F, Schmidt M. et al. Radioiodtherapie bei benignen Schilddrüsenerkrankungen (Version 5)*. Nukl – Nucl [Internet]; 2016; 55 (6): 213–220. doi:10.3413/Nukmed-0823-16-04

[57] Guha RoyA, Conjeti S, Navab N. et al. QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage [Internet] 2019; 186: 713–727. https://www.sciencedirect.com/science/article/pii/S1053811918321232

[58] Zielke J, Eilers C, Busam B. et al. RSV: Robotic Sonography for Thyroid Volumetry. IEEE Robot Autom Lett 2022; 7 (2): 3342–3348

[59] Yu SH, Choi SJ, Noh H. et al. Comparison of CT Volumetry and RECIST to Predict the Treatment Response and Overall Survival in Gastric Cancer Liver Metastases. Taehan Yongsang Uihakhoe Chi 2021; 82 (4): 876–888

[60] Hofmann FO, Heinemann V, D'Anastasi M. et al. Standard diametric versus volumetric early tumor shrinkage as a predictor of survival in metastatic colorectal cancer: subgroup findings of the randomized, open-label phase III trial FIRE-3/AIO KRK-0306. Eur Radiol 2023; 33 (2): 1174–1184

[61] Siegel MJ, Ippolito JE, Wahl RL. et al. Discrepant Assessments of Progressive Disease in Clinical Trials between Routine Clinical Reads and Formal RECIST 1.1 Interpretations. Radiol Imaging Cancer 2023; 5 (5): e230001

[62] Garin E, Tselikas L, Guiu B. et al. Personalised versus standard dosimetry approach of selective internal radiation therapy in patients with locally advanced hepatocellular carcinoma (DOSISPHERE-01): a randomised, multicentre, open-label phase 2 trial. Lancet Gastroenterol Hepatol 2021; 6 (1): 17–29

[63] Pacilio M, Conte M, Frantellizzi V. et al. Personalized Dosimetry in the Context of Radioiodine Therapy for Differentiated Thyroid Cancer. Diagn Basel Switz 2022; 12 (7): 1763

[64] Mariani G, Tonacchera M, Grosso M. et al. The Role of Nuclear Medicine in the Clinical Management of Benign Thyroid Disorders, Part 1: Hyperthyroidism. J Nucl Med Off Publ Soc Nucl Med 2021; 62 (3): 304–312

[65] van Heek L, Weindler J, Gorniak C. et al. Prognostic value of baseline metabolic tumor volume (MTV) for forecasting chemotherapy outcome in early-stage unfavorable Hodgkin lymphoma: Data from the phase III HD17 trial. Eur J Haematol 2023; 111 (6): 881–887