Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Physica Medica 87 (2021) 115–122

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

Automatic lung segmentation in COVID-19 patients: Impact on quantitative computed tomography analysis

L. Berta [a], F. Rizzetto [b,c], C. De Mattia [a], D. Lizio [a], M. Felisi [a], P.E. Colombo [a], S. Carrazza [d,e], S. Gelmini [d], L. Bianchi [b,c], D. Artioli [b], F. Travaglini [b], A. Vanzulli [b,f], A. Torresin [a,d,*], on behalf of the Niguarda COVID-19 Working Group

[a] Department of Medical Physics, ASST Grande Ospedale Metropolitano Niguarda, Piazza Ospedale Maggiore 3, 20162 Milan, Italy
[b] Department of Radiology, ASST Grande Ospedale Metropolitano Niguarda, Piazza Ospedale Maggiore 3, 20162 Milan, Italy
[c] Postgraduate School of Diagnostic and Interventional Radiology, Università degli Studi di Milano, via Festa del Perdono 7, 20122, Milan, Italy
[d] Department of Physics, Università degli Studi di Milano, via Giovanni Celoria 16, 20133 Milan, Italy
[e] Department of Physics, INFN Sezione di Milano, via Giovanni Celoria 16, 20133 Milan, Italy
[f] Department of Oncology and Hemato-Oncology, Università degli Studi di Milano, via Festa del Perdono 7, 20122, Milan, Italy

ARTICLE INFO

Keywords:
Quantitative imaging
Computed tomography
QCT
Lung segmentation
Segmentation algorithms
COVID-19

ABSTRACT

Purpose: To assess the impact of lung segmentation accuracy in an automatic pipeline for quantitative analysis of CT images.
Methods: Four different platforms for automatic lung segmentation based on convolutional neural network (CNN), region-growing technique and atlas-based algorithm were considered. The platforms were tested using CT images of 55 COVID-19 patients with severe lung impairment. Four radiologists assessed the segmentations using a 5-point qualitative score (QS). For each CT series, a manually revised reference segmentation (RS) was obtained. Histogram-based quantitative metrics (QM) were calculated from CT histogram using lung segmentationsfrom all platforms and RS. Dice index (DI) and differences of QMs (ΔQMs) were calculated between RS and other segmentations.
Results: Highest QS and lower ΔQMs values were associated to the CNN algorithm. However, only 45% CNN segmentations were judged to need no or only minimal corrections, and in only 17 cases (31%), automatic segmentations provided RS without manual corrections. Median values of the DI for the four algorithms ranged from 0.993 to 0.904. Significant differences for all QMs calculated between automatic segmentations and RS were found both when data were pooled together and stratified according to QS, indicating a relationship between qualitative and quantitative measurements. The most unstable QM was the histogram 90th percentile, with median ΔQMs values ranging from 10HU and 158HU between different algorithms.
Conclusions: None of tested algorithms provided fully reliable segmentation. Segmentation accuracy impacts differently on different quantitative metrics, and each of them should be individually evaluated according to the purpose of subsequent analyses.

1. Introduction

In March 2020 the World Health Organization defined SARS-CoV-2 disease (COVID-19) as a pandemic [1]. Since then, it has spread throughout the world causing more than 89 million global cases and nearly 1.9 million deaths [2]. COVID-19 is an infectious disease characterized by a broad spectrum of non-specific clinical manifestations, as fever, cough, dyspnoea and fatigue [3] which can cause from very mild to severe illness, including Acute Respiratory Distress Syndrome (ARDS) [4]. Although respecting the Berlin definition of ARDS [5] there is growing evidence that COVID-19 ARDS has distinctive pathophysiological features responsible for the heterogeneous presentations and responses of the patients [6–8].

Computed tomography (CT) plays a key role in the clinical classification and management of COVID-19 patients [9,10]. Starting from the experience acquired on ARDS [11–13], = numerous studies have

* Corresponding author.
E-mail address: alberto.torresin@unimi.it (A. Torresin).

https://doi.org/10.1016/j.ejmp.2021.06.001
Received 31 March 2021; Received in revised form 5 May 2021; Accepted 4 June 2021
Available online 7 June 2021
1120-1797/© 2021 Associazione Italiana di Fisica Medica. Published by Elsevier Ltd. All rights reserved.

focused on the quantitative analysis of the CT images (QCT) for the extraction, analysis and interpretation of quantitative data about COVID-19. Lung quantitative CT analysis includes threshold measurements that count the number of voxels above or below a specific attenuation value [14,15] features based on the intensity histogram [16,17] and texture metrics that consider the spatial relationship between voxels [18,19] up to the application of complex AI algorithms to create predictive models [20–22]. Most of the research works used a partition of the lung based on threshold measures to quantify the well-aerated and the compromised pulmonary volumes as predictors of the disease severity and its complications, such as the need for oxygenation support or ICU admission or the risk of death [14,15,19].

The first step required for the QCT is volume segmentation, which is the identification and delineation of the region of interest, usually coinciding with the entire lung volume in the case of pneumonia studies [23]. Several software from commercial and research sources may be exploited for this task, as reported in the Internet Analysis Tools Registry [24]. Automatic algorithms are preferable to minimize intra- and inter-operator variability and reduce the time dedicated to the contouring process, thus increasing the number of patients analysed [25]. Especially, the performance of deep learning-based or atlas-based algorithms continues to improve, so that they are expected to swiftly replace the other methods and become the standard [26–28]. However, when dealing with routine data, automatic segmentation still relies on human inspection and manual refinements because of the high diversity of physiological and pathological phenotypes and image data [29].

The lung segmentation task provides a clear example of this issue. Healthy lungs are a low-density high-contrast region where automatic segmentation algorithms work well [30]. Conversely, the presence of abnormalities, like pleural effusion or parenchymal consolidations, which have attenuation characteristics similar to the pleural margin and the thoracic soft tissues, often leads to inaccurate output [31,32]. Therefore, manual contouring, or at least manual correction, remains the actual reference standard for lung segmentation, but such an approach is time-consuming and hardly compatible with large-scale data analyses [29]. Given the pressing need to implement automatic segmentation tools, and considering that the contouring process affects the QCT results [33–35] even when automatic [36] it is crucial to investigate how this source of variability impacts on the QCT of COVID-19 patients.

In this work, four medical radiologists evaluated the performance of four different image analysis platforms for the automatic lung segmentation applied on a cohort of COVID-19 patients with severe lung involvement. These automatic segmentations were then compared against a manually corrected reference. Finally, quantitative metrics resulting from CT histogram of the lungs were calculated using automatic and manually corrected segmentations in order to understand how differences in segmentations affect quantitative measurements.

2. Material and methods

This retrospective study was approved by the Local Ethics Committee. The need for informed consent was waived owing to the retrospective nature of the study.

2.1. Patient population and CT protocol

To select cases with relevant lung involvement, patients with positive Real-Time Polymerase Chain Reaction for SARS-CoV-2 and positive chest CT scan were randomly selected among those admitted in intensive care unit within the 48 h after the CT scan acquisition in March-April 2020 in Niguarda Hospital.

Chest CT examinations were acquired on 3 Siemens scanners installed at our hospital (Somatom Definition Edge, Somatom Sensation 64, Somatom Definition) with patients in supine position, during inspiratory breath-hold, in keeping with the collaboration status of the patient. For each patient, the unenhanced series reconstructed with slice thickness of 3 or 1 mm and Bl57- or B70 kernels, as available, was considered for subsequent processing.

2.2. Segmentation algorithms

In all chest CT scans the lungs were automatically segmented using four image analysis platforms:

a) Philips IntelliSpace Portal 9.0 (Philips Healthcare SpA), hereafter called ISP: automatic region-growing segmentation of the lungs was obtained using the *Chronic Obstructive Pulmonary Disease (COPD) Analysis* module included in the PACS software suite;
b) 3D Slicer 4.10.2 (https://www.slicer.org), hereafter simply called Slicer: lungs were automatically segmented in two ways, i.e. using:
   1. the *Parenchymal Analysis* function in the *Chest Imaging Platform* extension (Applied Chest Imaging Laboratory; Boston, Massachusetts, USA), which implements a region-growing-like technique specifically tailored for lung segmentation;
   2. a Deep Convolutional Neural Network extension based on a U-Net architecture trained to segment lungs affected by multiple pathologies including COVID-19 (https://github.com/acil-bwh/ChestImagingPlatform/blob/develop/cip_python/dcnn/projects/lung_segmenter/lung_segmenter_dcnn.py);
c) QUIBIM Precision® Platform 2.8 (QUIBIM, Valencia, Spain), hereafter simply called QUIBIM: automatic segmentation of the lungs was provided by means of a deep learning model;
d) the Simultaneous Truth And Performance Level Estimation (STAPLE) approach inside the atlas-based auto-segmentation software package ABAS® (Elekta Oncology Systems, Crawley, UK), hereafter simply called ABAS: this multiatlas-based automatic tool applies several individual atlases to obtain multiple segmentations of the same subject and combine them into a final unique segmentation. The selection of the input required for the best fit in lung segmentation is not straightforward: the larger the number of images and structures used, the longer it will take the tool to search through the atlas to select the best match. Onestudy suggested that a data set of 15 patients may berequired for abdominal organs [37]. For this analysis, different combinations of atlases were tested:
   1. lungs from 18 patients with COVID-19 pneumonia and from 6 patients with no pulmonary involvement;
   2. lungs, liver and heart from 6 patients with COVID-19 pneumonia;
   3. lungs, liver and heart from 12 patients with COVID-19 pneumonia.

In all cases, the segmentations were obtained without the need of user input and no manual corrections were applied to the algorithms output.

2.3. Assessment of segmentation performance

The lung volumes delineated in all CT images with all platforms were evaluated using subjective and quantitative approaches. Furthermore, the effects of different segmentations on QCT were analysed.

2.3.1. Qualitative assessment of segmentations

Two resident radiologists (2 and 3 years of experience) and two senior radiologists (15 and 14 years of experience) evaluated the segmentation performance of the four algorithms. When different methods were tested for a certain platform, the Authors collectively assessed the quality of lung segmentation to choose the best option. Thus, for each CT series the four segmentations representative of the different image platforms were ranked according to quality (1 = best segmentation; 4 = worst segmentation) and scored using the following 5-point Qualitative Score (QS):

- 1 = very poor: segmentation with extensive errors requiring the reader excessive effort to correct them;
- 2 = poor: segmentation with errors that require sizeable and/or time-consuming corrections;
- 3 = acceptable: segmentation with inaccuracies that require limited and/or brief corrections;
- 4 = good: segmentation with small imperfections negligible for the reader;
- 5 = excellent: segmentation corresponding to the ideal result for the reader.

The test was performed using a program developed in JavaScript that automatically opened the same CT series in four separated windows, overlaying each of them with one of the four segmentations to be assessed. The axial slices were synchronized to facilitate comparison, while the patient and segmentation data were not shown. The radiologist selected the window with the best segmentation by clicking on it and assigned the QS to it through a dialog box; afterwards, the window closed. The order in which the windows were selected was tracked to define the segmentation rank. After evaluating the 4 segmentations of the same CT series, the program automatically loaded the data of the next patient.

For each patient, the segmentation with higher QS and higher rank in case of QS tie was manually corrected, when necessary, to obtain a collectively agreed reference segmentation (RS) for the subsequent analysis.

2.3.2. Quantitative assessment of segmentations

Quantitative assessment of segmentations was performed using Dice index (DI), an objective metric that quantifies the spatial overlap between two contours, ranging from 0 for null overlapping to 1 for perfect overlapping [38,39]. Lung volumes were also calculated multiplying the number of voxels included in the segmentations by the voxel size. DI and differences between lung volumes were computed comparing the RS with the segmentations provided by the different image platforms. Differences between lung volumes were expressed as relative values (ΔVol (%)).

2.3.3. Histogram metrics of QCT on segmented lungs

Quantitative metrics (QM) derived from the relative CT lung histogram were calculated using all segmentations for each CT image. Deciles, mean value, skewness and kurtosis of the histogram values were calculated using a software for automatic analysis of CT lung images written in JavaScript code [17]. In addition, a metric representative of well-aerated lung volume (WAVE.f) previously described [17] was calculated. The differences and the absolute values between QM obtained with RS and other segmentations (ΔQM, ΔQM_{abs}) were calculated for all data.

2.4. Data analysis

The data analysis was generated using the Real Statistics Resource Pack software (Release 6.8) (www.real-statistics.com).

Quantitative data were tested with Shapiro-Wilk test for normality and Levene test for homogeneity of variance. The comparisons of paired and non-paired quantitative data among multiple groups were evaluated using Friedman test and Kruskal-Wallis test, respectively; comparison of paired data between two groups were evaluated using paired *t*-test and Wilcoxon signed rank test, as appropriate.

Statistical significance was established at the $p < 0.05$ level, applying Bonferroni's correction for multiple comparisons when appropriate.

2.4.1. Data analysis of qualitative assessment of segmentations

Categorical variables were expressed as counts and percentage. Results of continuous variables were reported as median values with 25th and 75th percentiles of their distribution.

The chance-corrected inter-reader agreement for the QS was tested using Gwet's second-order agreement coefficient (AC2) with ordinal weights [40]. AC2 was chosen to correct for the partial agreement occurring when comparing ordinal variables with multiple readers and because it is less affected by prevalence and marginal distribution [41,42]. A level of agreement > 0.8 was considered very good, following Altman's interpretation [43]. Weighted percentage agreement was also reported [44].

To account for the role of heterogeneity in COVID-19 lung lesions on the segmentation task, the CT images were grouped in 3 classes as follows. CT images with at least one automatic segmentation judged not worthy of manual correction (QS ≥ 4) by all the readers were considered "*easy*" to segment. Conversely, images were considered "*critical*" when all readers judged all outputs of the automatic platforms sub-optimal (QS ≤ 3). The remaining cases were labelled as "*challenging*".

Furthermore, according to the scores given by the four readers, each segmentation for all platforms was labelled as *"optimal"*, if all readers agreed with a QS ≥ 4, *"sub-optimal"*, if at least one reader judged necessary to manually correct the segmentation (QS ≤ 3), and *"unsuitable"* if all readers gave a QS < 4.

2.4.2. Data analysis on quantitative assessment of segmentations and histogram metrics

For each patient, DI, lung volumes and the results of QCT metrics obtained from each of the four segmentation masks were compared with the corresponding results from the RS to understand how different segmentations affected QCT results.

3. Results

According to the inclusion criteria, a cohort of 55 patients (41 males, 14 females; median age 56 years, range 33–74 years) was enrolled for the study. Example of segmentation outputs and subsequent analysis is reported in Fig. 1.

3.1. Results of qualitative assessment of segmentations

After a preliminary assessment, the *Parenchymal Analysis* function in the *Chest Imaging Platform* extension for Slicer and the combination of the atlases from 12 patients with COVID-19 pneumonia for ABAS were chosen to be tested with the QS together with QUIBIM and ISP platforms. Thus, a total of 220 segmentations were obtained and scored by the four readers.

Since the inter-reader agreement for the QS was very good, with an AC2 = 0.88 (95 %CI: 0.86–0.90; $p < 0.001$) and weighted percentage agreement was 95%(with perfect agreement between the four readers in 75 (34%) cases and at least three readers giving the same score in 205 (93%) cases), no further tests on reproducibility of readers' scoring were performed. In 59% of the cases where the readers showed perfect agreement, a QS = 3 was assigned to the segmentation. Averaging over the readers, 45%, 20%, 37% and 7% of the scores assigned to QUIBIM, Slicer, ABAS and ISP segmented lungs, respectively, were positive, indicating no need no or only minimal corrections (QS ≥ 4), as reported in Table 1. The QS differences observed between the four platforms were significant ($p < 0.001$) according to Friedman test. For 17 CT scans, automatic segmentations provided RS without any manual correction.

As regards the classification of cases according to the difficulty of automatic segmentation, lung images of 20 CT scans resulted *"easy to segment"*, 10 resulted *"critical to segment"* and 25 were considered *"challenging"*.

On the other hand, regarding the classification of each individual segmentation, 37 of them resulted *"optimal"* (QUIBIM = 18; Slicer = 10; ABAS = 7; ISP = 2), 61 resulted "*suboptimal*" (QUIBIM = 15; Slicer = 7; ABAS = 33; ISP = 6) and 122 resulted "*unsuitable*" (QUIBIM = 22; Slicer = 38; ABAS = 15; ISP = 47).

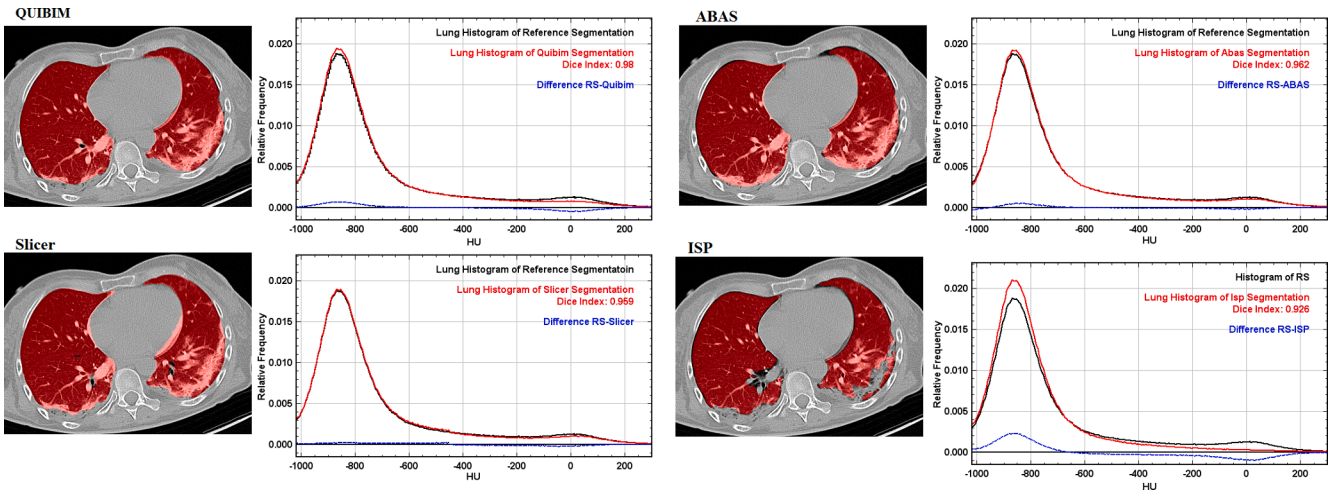

Fig. 1. Example of the outputs of the four automatic segmentation tools tested on a COVID-19 CT scan. Automatic segmentations are reported as a red overlay. CT lung histograms calculated from reference (black curve) and automatic (red curve) segmentations are reported on the same graphs with their difference (blue curve). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Results of the Qualitative Score (QS) given by four radiologists to the automatic segmentations of 55 COVID-19 lungs obtained with four image platforms. Data are averaged over the 55 segmentations and the four readers. A QS ≥ 4 was given when the segmentation was assessed as needing no or only minimal corrections.

QS	QUIBIM		Slicer		ABAS®		ISP	
1	0%		5%		0%		10%	
2	5%	55%	29%	80%	4%	63%	45%	93%
3	49%		47%		59%		37%	
4	29%	45%	16%	20%	33%	37%	7%	7%
5	16%		4%		4%		0%	

3.2. Results of quantitative assessment of segmentations and histogram metrics

Median (25th percentile, 75th percentile) values of the DI were 0.994 (0.979, 1), 0.952 (0.921, 0.971), 0.959 (0.945, 0.963) and 0.904 (0.848, 0.937) for QUIBIM, Slicer, ABAS and ISP, respectively. The differences between results were significant ($p < 0.001$) according to Friedman test. Values of DI and ΔVol(%) are graphically reported in Fig. 2.

The results of the ΔQM are reported in Table 2. CT lung histogram percentiles calculated using automatic segmentations were always significantly different from QM obtained with RS, albeit the extent of the differences could be very limited. For example, both QUIBIM and ABAS segmentations showed a median volume discrepancy of −1% compared with the RS. Also, an increase in the extent of the differences was observed as the percentiles(i.e. the CT voxel density) increased. In particular, the histogram 90th percentiles was the most unstable metrics, with median values ranging from 10HU and 158HU betweendifferent algorithms.

Notably, while for QUIBIM, Slicer and ISP these differences were negative, indicating the exclusion of areas with higher density, ABAS algorithm showedan opposite behaviour in that the differences were positive.

Moreover, the median difference in WAVE.f between automatic segmentations and RS were within 2%, except for ISP where it rose to 7.9%.

Median values of ΔQM_{abs} for each histogram percentiles are graphically reported in Fig. 3 for the three classes of patients (“easy”, “challenging” and “critical” to segment) and for the platforms used. The gap between histogram metrics increased according to the class of segmentation difficulty, except for ABAS results, which were more homogeneous and independent of the class of patients.

Regarding the subdivision of the segmentations based on the QS received, the distribution of values of all quantitative metrics between the three classes (“optimal”, “sub-optimal” and “unsuitable”) are summarized in boxplots in Fig. 4. Such differences were all significative ($p < 0.001$) according to Kruskal-Wallis test and showed how the QS assigned by the radiologists is reflected in different discrepancies compared with the RS. In particular, wider ranges of values were observed when moving from the “optimal” to the “unsuitable” category.

Notably, for the 122 “unsuitable” segmentations (QUIBIM = 22,

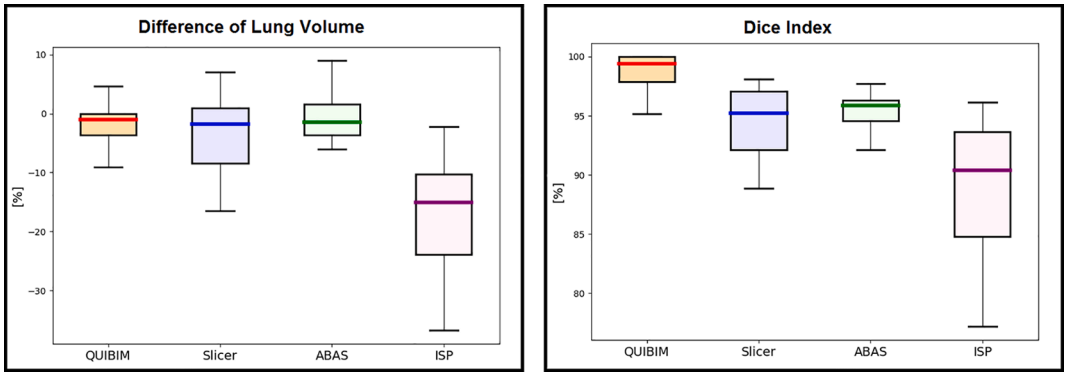

Fig. 2. Boxplots of Dice Index values and differences of volumes calculated between the automatic segmentations and the reference segmentations revised by the radiologists.

Table 2
Results of QCT calculated from the automatic lung segmentation obtained with four different automatic platforms. Values are expressed as differences to the equivalent metrics calculated from the Reference Standard and reported as the three quartiles of the distribution of the 55 cases. The p values adjusted after Bonferroni's correction were reported. [n]p: [n]th percentile; Mean.H: density histogram mean; Skew.H: density histogram skewness; Kurt.H: density histogram kurtosis; W.fit %: well-aerated lung volume estimation[17]; ΔVol%: percentual difference of volumes.

	QUIBIM				Slicer				ABAS				ISP			
	Quartile			p	Quartile			p	Quartile			p	Quartile			p
	1st	2nd	3rd		1st	2nd	3rd		1st	2nd	3rd		1st	2nd	3rd	
10p	−3	−1	0	<0.001	−7	−3	0	<0.001	1	1	2	<0.001	−19	−10	−5	<0.001
20p	−4	−1	0	<0.001	−10	−3	0	<0.001	0	2	3	<0.001	−31	−17	−8	<0.001
30p	−6	−2	0	<0.001	−15	−4	0	<0.001	0	2	4	<0.001	−46	−22	−12	<0.001
40p	−9	−3	0	<0.001	−21	−5	0	<0.001	0	2	6	<0.001	−62	−30	−16	<0.001
50p	−13	−3	0	<0.001	−33	−8	0	<0.001	−1	4	8	<0.001	−87	−40	−22	<0.001
60p	−18	−5	0	<0.001	−55	−14	−1	<0.001	0	5	12	<0.001	−120	−57	−27	<0.001
70p	−20	−7	0	<0.001	−69	−23	−3	<0.001	1	9	17	<0.001	−142	−87	−40	<0.001
80p	−23	−9	0	<0.001	−86	−28	−6	<0.001	1	12	27	<0.001	−203	−118	−59	<0.001
90p	−33	−10	0	<0.001	−86	−24	−3	<0.001	2	18	52	<0.001	−245	−158	−93	<0.001
Mean.H	−14	−5	0	<0.001	−45	−15	−3	<0.001	1	7	15	<0.001	−100	−61	−37	<0.001
Skew.H	0.00	0.01	0.06	<0.001	0.03	0.10	0.22	<0.001	−0.05	−0.01	0.02	0.023	0.06	0.17	0.41	<0.001
Kurt.H	0.00	0.03	0.14	<0.001	0.09	0.22	0.62	<0.001	−0.26	−0.07	−0.01	<0.001	0.33	0.73	1.52	<0.001
W.fit %	0.0	0.5	1.6	<0.001	0.1	2.0	5.5	<0.001	−1.5	−0.9	0.0	<0.001	4.3	7.9	12.6	<0.001
ΔVol%	−4	−1	0	<0.001	−8	−2	1	0.003	−4	−1	1	0.045	−24	−15	−10	<0.001

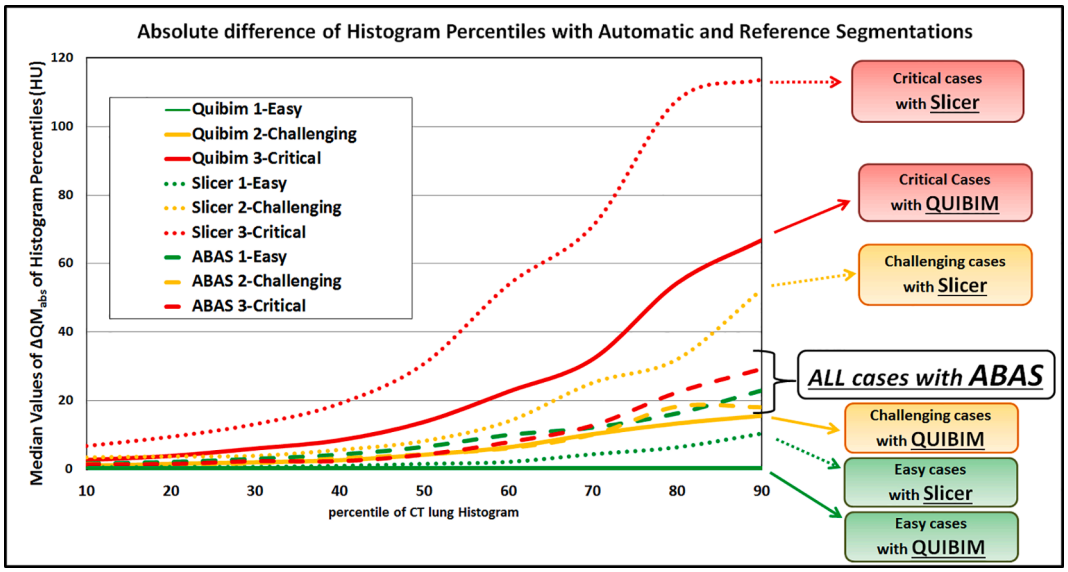

Fig. 3. Absolute difference in percentile values of the CT histogram calculated between automatic and reference segmentations. The results were divided according to the classification of patients based on the segmentation difficulty ("easy", "challenging" and "critical"). Values obtained from ISP were not shown because out-of-scale.

Slicer = 38, ABAS = 15, ISP = 47) median (25thpercentile, 75th percentile) value of DI and 90th percentile resulted in 0.930 (0.890, 0.957) and 88 (29,159)HU. Also, median (25thpercentile, 75th percentile) difference for WAVE.f was 0.31 (0.00, 0.91)%, 0.88 (0.31, 1.9)% and 3.00 (1.30, 5.65)% for "optimal", "sub-optimal" and "unsuitable" segmentations, respectively.

4. Discussion

In this work we compared four different tools for the automatic segmentation of lung in CT images of COVID-19 patients with the aim of assessing accuracy and suitability in a fully integrated workflow of image analysis. There are many papers in the literature on automatic segmentations applied to radiotherapy field [28] and, to the best of our knowledge, this is the first work published on these topics in COVID-19 diagnostic imaging field.

Quantitative metrics from imaging could allow to overcome the current lack of effective biomarkers for diagnosis, prognosis and outcome prediction in COVID-19 patients. However, quantitative image analysis still relies on manual intervention, which is time-consuming and often not compatible with daily clinical workloads. Since a fully automatic process would represent a paradigm shift in the application of quantitative imaging, it is crucial to understand the degree of inaccuracy of these metrics entailed by the use of unsupervised automatic segmentations.

A critical issue in dealing with performance evaluation of automatic segmentation is the dependency of output scoring on individual subjectivity. In our study, the agreement between the four readers involved was high, especially considering their different experience. This suggests that the biases related to the qualitative score were limited, in particular in separating the "easy" and "critical" cases and the "optimal" and "unsuitable" segmentations as defined in the *Materials and Methods* section.

Following the readers' evaluation, the automatic segmentations of only 17 out of 55 images of COVID-19 patients with severe lung impairment were taken directly as reference. Instead, for the other 38

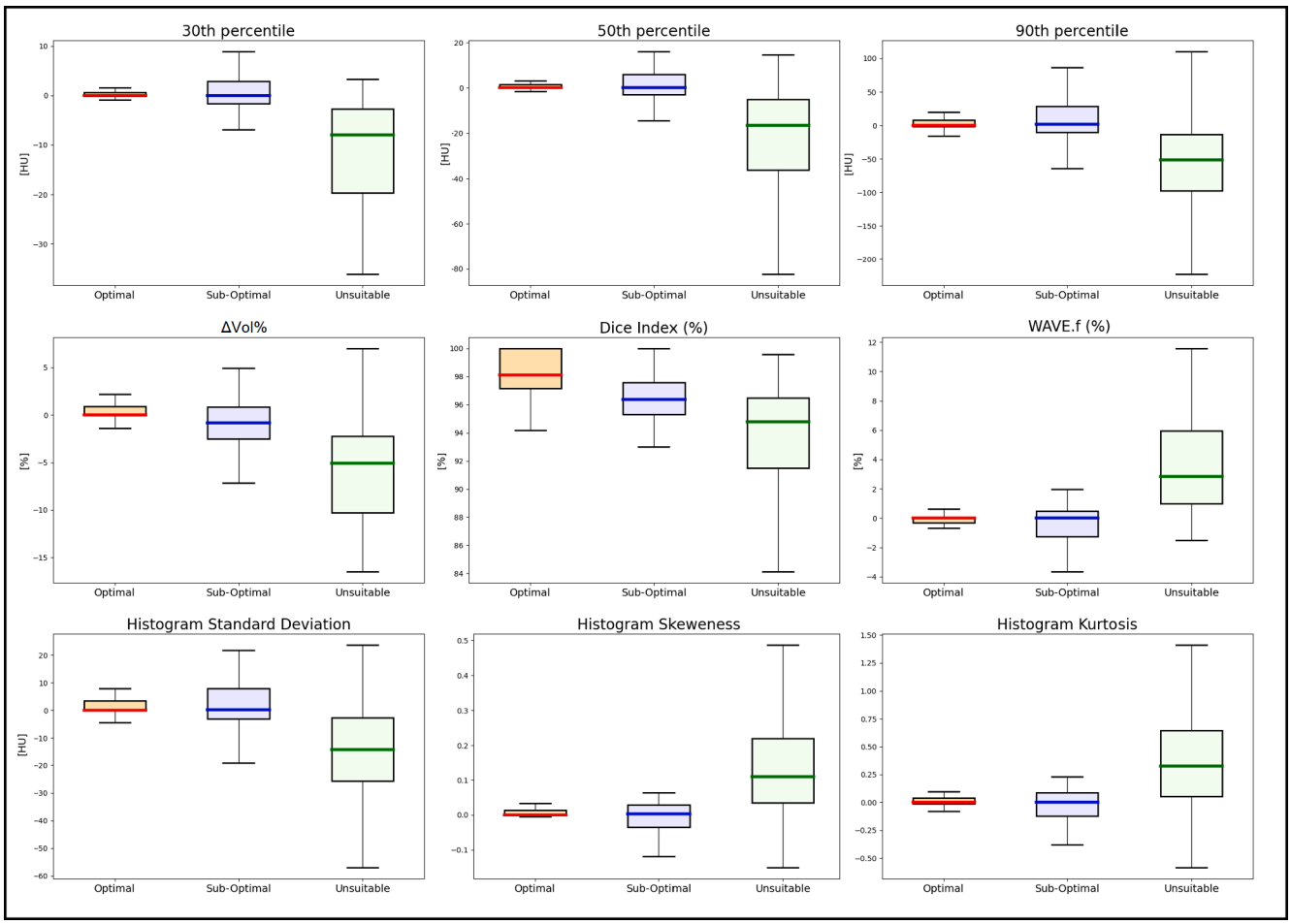

Fig. 4. Boxplots of the differences in CT histogram metrics calculated between the automatic segmentations and the reference segmentations, divided according to Qualitative Score (QS) assigned by the radiologists. "Optimal": segmentations with all QS $\geq$ 4; "sub-optimal": segmentations with at least 1 QS $<$ 4; "unsuitable": segmentations with all QS $\leq$ 3.

patients manual adjustments were introduced due to a certain degree of error. In the best case, <50% of the automatic outputs were judged by the readers to be accurate enough not to require corrections, but the percentage dropped much lower for some of the analysed tools.

Interestingly, median DI values of 98.11, 96.35 and 93.00 were found for "optimal", "sub-optimal" and "unsuitable" segmentations, respectively. In other words, DI indicated a great overlap between automatic segmentation and RS even when the software output was negatively judged by the readers. This is explained by the fact that automatic tools commonly fail to segment the areas of pulmonary consolidations, especially when they are adjacent to the thoracic wall. These segmentation errors may have a low impact on the DI values because of the large total volume of the lungs, but they can lead to a serious underestimation of the extent of the pulmonary damage. For example, in our study the tools based on region growing algorithm, which rely on contrast homogeneity and are more prone to cut off the peripheral opacities, received the worst evaluation by the readers. Since overlap-based similarity indices do not take into account the anatomical and clinical relevance of the segmented regions, even DI values up to 0.90, usually interpreted as optimal in other contexts, may be misleading when large volume structures has to be compared, as lung in COVID-19 patients [28].

The impact of this issue on the QCT is clear observing how, for all algorithms, differences in quantitative metrics increased in the higher percentiles of the HU distribution calculated from the segmentations. Notably, 3D Slicer and also QUIBIM produced the segmentations with the highest QS on average, but both provided quantitative results with a consistently different gap for "critical", "challenging" and "easy" cases. By contrast, the discrepancies of quantitative metrics observed for ABAS were confined in a limited range independently of the class of segmentation difficulty assigned to the CT images. This is explained by the fact that ABAS works with an algorithm based on "image deformation" rather than "region growing", which may be less accurate on average but also less susceptible to the problems caused by low-contrast interfaces. Therefore, algorithms based on morphology rather than grey level thresholds or image contrast could be more practical when the heterogeneity and the diffusion of disease is high.

The downside of atlas-based algorithms is that part of the heart or liver adjacent to the lungs can be included in the segmentation as well. Indeed, the differences between HU percentiles were almost always positive for ABAS compared with RS, meaning that it included larger volume of more dense tissues in the lung segmentation. Nevertheless, it must be pointed out that, differently from the other three platforms, ABAS could be used with an arbitrary number of inputs. In this work, we run the automatic lung segmentation using a total of 12 atlases with four structures each: the two lungs, the heart, and the hepatic dome. More accurate results could be assumed if a higher number of atlases would be used, albeit at the cost of more calculation time.

As expected, significant differences in QCT results were observed between "optimal", "sub-optimal" and "unsuitable" segmentations, with most of the histogram metrics showing the largest discrepancies from RS when "*unsuitable*" segmentations were used. This result shows how qualitative scoring and quantitative metrics are strongly related, with inaccurate segmentations resulting in lower subjective scores and larger

QCT discrepancies compared to the reference. However, the actual error in the results of quantitative metrics may be limited even when the readers gave negative judgements. For example, the WAVE.f, the biomarker we previously described [17] to estimate the well-aerated pulmonary volume from the CT histogram of the lungs, differed by few percentage points from the reference for most algorithms and segmentation classes. Recent works [45,46] have reported positive correlations between CT severity scores, based on the visual assessment of lung involvement by expert radiologists, and clinical outcome like COVID-19 patient prognosis. If severity score and WAVE.f correlates, as it is reasonable to assume, a reliable assessment of lung involvement could be possible in a fully automatic way.

In general, when automatic segmentations are integrated into a fully automatic pipeline, some degree of error is inevitably introduced into the QCT, which may vary between different metrics. The magnitude of these differences compared with the reference standard should be assessed for each metric, and the decision to tolerate it or not should be made in relation to the accuracy required for the purposes of subsequent analysis and discussed individually.

The main limit of this study is that four specific image analysis platforms were considered, but the results could vary using other tools, especially those implementing more tailored algorithms. However, we chose to compare tools based on different segmentation techniques and that were commercially licensed or freely accessible and largely used in literature, like 3D Slicer, to give a more realistic picture of the tools currently available. Also, we used a cohort of patients with severe COVID-19 pneumonia, but the discrepancies between the automatic segmentations and the RS are expected to reduce if patients with milder lung involvement are considered.

5. Conclusions

None of the tested imaging platforms fully provided reliable automatic segmentation in COVID-19 patients with severe lung involvement. Since the qualitative score anticipated the extent of differences in quantitative metrics, quality assurance programs for automatic image analysis pipeline may include subjective scoring of a random sample of segmented images. However, the inaccuracy in quantitative metrics due to segmentations should be always weighted according to the purpose of subsequent analyses and the accuracy they required.

Acknowledgements

This work has been developed within the project:“CoviLAKE: a Datalake of images and clinical data of COVID19 patients – Data aggregation for radiomic and artificial intelligence analysis of CT Chest and clinical data of COVID-19 patients”. The project had no financial support and it was approved by ethical committee.

References

[1] World Health Organization. Coronavirus disease 2019 (COVID-19) Situation Report – 51. Geneva (Switzerland); 2020.
[2] Johns Hopkins University. COVID-19 Map. Johns Hopkins Coronavirus Resour Cent; 2021. https://coronavirus.jhu.edu/map.html [accessed March 20, 2021].
[3] Lechien JR, Chiesa-Estomba CM, Place S, Van Laethem Y, Cabaraux P, Mat Q, et al. Clinical and epidemiological characteristics of 1420 European patients with mild-to-moderate coronavirus disease 2019. J Intern Med 2020;288:335–44. https://doi.org/10.1111/joim.13089.
[4] Zhang JJY, Lee KS, Ang LW, Leo YS, Young BE. Risk Factors for Severe Disease and Efficacy of Treatment in Patients Infected With COVID-19: A Systematic Review, Meta-Analysis, and Meta-Regression Analysis. Clin Infect Dis 2020;71:2199–206. https://doi.org/10.1093/cid/ciaa576.
[5] Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, Fan E, et al. Acute respiratory distress syndrome: The Berlin definition. JAMA - J Am Med Assoc 2012. https://doi.org/10.1001/jama.2012.5669.
[6] Chiumello D, Busana M, Coppola S, Romitti F, Formenti P, Bonifazi M, et al. Physiological and quantitative CT-scan characterization of COVID-19 and typical ARDS: a matched cohort study. Intensive Care Med 2020;46:2187–96. https://doi.org/10.1007/s00134-020-06281-2.
[7] Grasselli G, Tonetti T, Protti A, Langer T, Girardis M, Bellani G, et al. Pathophysiology of COVID-19-associated acute respiratory distress syndrome: a multicentre prospective observational study. Lancet Respir Med 2020;8:1201–8. https://doi.org/10.1016/S2213-2600(20)30370-2.
[8] Gattinoni L, Chiumello D, Caironi P, Busana M, Romitti F, Brazzi L, et al. COVID-19 pneumonia: different respiratory treatments for different phenotypes? Intensive Care Med 2020:6–9. https://doi.org/10.1007/s00134-020-06033-2.
[9] Chung M, Bernheim A, Mei X, Zhang N, Huang M, Zeng X, et al. CT imaging features of 2019 novel coronavirus (2019-NCoV). Radiology 2020;295:202–7. https://doi.org/10.1148/radiol.2020200230.
[10] Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, et al. The Role of Chest Imaging in Patient Management During the COVID-19 Pandemic. Chest 2020;158:106–16. https://doi.org/10.1016/j.chest.2020.04.003.
[11] Gattinoni L, Presenti A, Torresin A, Baglioni S, Rivolta M, Rossi F, et al. Adult respiratory distress syndrome profiles by computed tomography. J Thorac Imaging 1986;1:25–30. https://doi.org/10.1097/00005382-198607000-00005.
[12] Ichikado K, Suga M, Muranaka H, Gushima Y, Miyakawa H, Tsubamoto M, et al. Prediction of Prognosis for Acute Respiratory Distress Syndrome with Thin-Section CT: Validation in 44 Cases. Radiology 2006;238:321–9. https://doi.org/10.1148/radiol.2373041515.
[13] Nishiyama A, Kawata N, Yokota H, Sugiura T, Matsumura Y, Higashide T, et al. A predictive factor for patients with acute respiratory distress syndrome: CT lung volumetry of the well-aerated region as an automated method. Eur J Radiol 2020;122:108748. https://doi.org/10.1016/j.ejrad.2019.108748.
[14] Lanza E, Muglia R, Bolengo I, Santonocito OG, Lisi C, Angelotti G, et al. Quantitative chest CT analysis in COVID-19 to predict the need for oxygenation support and intubation. Eur Radiol 2020;30:6770–8. https://doi.org/10.1007/s00330-020-07013-2.
[15] Colombi D, Bodini FC, Petrini M, Maffi G, Morelli N, Milanese G, et al. Well-aerated Lung on Admitting Chest CT to Predict Adverse Outcome in COVID-19 Pneumonia. Radiology 2020;201433. https://doi.org/10.1148/radiol.2020201433.
[16] Ash SY, Harmouche R, Vallejo DLL, Villalba JA, Ostridge K, Gunville R, et al. Densitometric and local histogram based analysis of computed tomography images in patients with idiopathic pulmonary fibrosis. Respir Res 2017;18:1–11. https://doi.org/10.1186/s12931-017-0527-8.
[17] Berta L, De Mattia C, Rizzetto F, Carrazza S, Colombo PE, Fumagalli R, et al. A patient-specific approach for quantitative and automatic analysis of computed tomography images in lung disease: Application to COVID-19 patients. Phys Medica 2021;82:28–39. https://doi.org/10.1016/j.ejmp.2021.01.004.
[18] Wei W, Hu X, wen,, Cheng, Q,, Zhao, Y, ming,, Ge, Y, qiong.. Identification of common and severe COVID-19: the value of CT texture analysis and correlation with clinical characteristics. Eur Radiol 2020. https://doi.org/10.1007/s00330-020-07012-3.
[19] Shen C, Yu N, Cai S, Zhou J, Sheng J, Liu K, et al. Quantitative computed tomography analysis for stratifying the severity of Coronavirus Disease 2019. J Pharm Anal 2020;10:123–9. https://doi.org/10.1016/j.jpha.2020.03.004.
[20] Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput Biol Med 2020. https://doi.org/10.1016/j.compbiomed.2020.103795.
[21] Huang L, Han R, Ai T, Yu P, Kang H, Tao Q, et al. Serial Quantitative Chest CT Assessment of COVID-19: A Deep Learning Approach. Radiol Cardiothorac Imaging 2020. https://doi.org/10.1148/ryct.2020200075.
[22] Lessmann N, Sánchez CI, Beenen L, Boulogne LH, Brink M, Calli E, et al. Automated Assessment of COVID-19 Reporting and Data System and Chest CT Severity Scores in Patients Suspected of Having COVID-19 Using Artificial Intelligence. Radiology 2021;298:E18–28. https://doi.org/10.1148/radiol.2020202439.
[23] Mascalchi M, Camiciottoli G, Diciotti S. Lung densitometry: Why, how and when. J Thorac Dis 2017;9:3319–45. https://doi.org/10.21037/jtd.2017.08.17.
[24] Kennedy DN, Haselgrove C. The Internet Analysis Tools Registry: A Public Resource for Image Analysis. Neuroinformatics 2006;4:263–70. https://doi.org/10.1385/NI:4:3:263.
[25] Withey DJ, Koles ZJ. A Review of Medical Image Segmentation: Methods and Available Software. IjbemOrg 2008;10:125–48.
[26] Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in Auto-Segmentation. Semin Radiat Oncol 2019;29:185–97. https://doi.org/10.1016/j.semradonc.2019.02.001.
[27] Xie W, Jacobs C, Charbonnier J-P, van Ginneken B. Relational Modeling for Robust and Efficient Pulmonary Lobe Segmentation in CT Scans. IEEE Trans Med Imaging 2020;39:2664–75. https://doi.org/10.1109/TMI.2020.2995108.
[28] Maffei N, Fiorini L, Aluisio G, D'Angelo E, Ferrazza P, Vanoni V, et al. Hierarchical clustering applied to automatic atlas based segmentation of 25 cardiac substructures. Phys Medica 2020;69:70–80. https://doi.org/10.1016/j.ejmp.2019.12.001.
[29] Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur Radiol Exp 2020;4:50. https://doi.org/10.1186/s41747-020-00173-2.
[30] Doel T, Gavaghan DJ, Grau V. Review of automatic pulmonary lobe segmentation methods from CT. Comput Med Imaging Graph 2015. https://doi.org/10.1016/j.compmedimag.2014.10.008.
[31] Mansoor A, Bagci U, Foster B, Xu Z, Papadakis GZ, Folio LR, et al. Segmentation and image analysis of abnormal lungs at CT: Current approaches, challenges, and future trends. Radiographics 2015;35:1056–76. https://doi.org/10.1148/rg.2015140232.

[32] Kiser KJ, Ahmed S, Stieb S, Mohamed ASR, Elhalawani H, Park PYS, et al. PleThora: Pleural effusion and thoracic cavity segmentations in diseased lungs for benchmarking chest CT processing pipelines. Med Phys 2020;47:5941–52. https://doi.org/10.1002/mp.14424.
[33] Rizzetto F, Calderoni F, De Mattia C, Defeudis A, Giannini V, Mazzetti S, et al. Impact of inter-reader contouring variability on textural radiomics of colorectal liver metastases. Eur Radiol Exp 2020;4:62. https://doi.org/10.1186/s41747-020-00189-8.
[34] Haarburger C, Müller-Franzes G, Weninger L, Kuhl C, Truhn D, Merhof D. Radiomics feature reproducibility under inter-rater variability in segmentations of CT images. Sci Rep 2020;10:12688. https://doi.org/10.1038/s41598-020-69534-6.
[35] Pavic M, Bogowicz M, Würms X, Glatz S, Finazzi T, Riesterer O, et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncol (Madr) 2018;57:1070–4. https://doi.org/10.1080/0284186X.2018.1445283.
[36] Lim H, Weinheimer O, Wielpütz MO, Dinkel J, Hielscher T, Gompelmann D, et al. Fully Automated Pulmonary Lobar Segmentation: Influence of Different Prototype Software Programs onto Quantitative Evaluation of Chronic Obstructive Lung Disease. PLoS ONE 2016;11:e0151498. https://doi.org/10.1371/journal.pone.0151498.
[37] Hwee J, Louie AV, Gaede S, Bauman G, D'Souza D, Sexton T, et al. Technology Assessment of Automated Atlas Based Segmentation in Prostate Bed Contouring. Radiat Oncol 2011;6:110. https://doi.org/10.1186/1748-717X-6-110.
[38] Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology 1945;26:297–302. https://doi.org/10.2307/1932409.
[39] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging 2015;15:29. https://doi.org/10.1186/s12880-015-0068-x.
[40] Tran D, Dolgun A, Demirhan H. Weighted inter-rater agreement measures for ordinal outcomes. Commun Stat Simul Comput 2020;49:989–1003. https://doi.org/10.1080/03610918.2018.1490428.
[41] Quarfoot D, Levine RA. How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution? Am Statist 2017. https://doi.org/10.1080/00031305.2016.1141708.
[42] Vial A, Assink M, Stams GJJM, van der Put C. Safety and Risk Assessment in Child Welfare: A Reliability Study Using Multiple Measures. J Child Fam Stud 2019;28:3533–44. https://doi.org/10.1007/s10826-019-01536-z.
[43] Altman D. Practical statistics for medical research. London: Chapman and Hall; 1991.
[44] Gwet KL. On The Krippendorff`s Alpha Coefficient; 2011.
[45] Wu J, Pan J, Teng D, Xu X, Feng J, Chen Y-C. Interpretation of CT signs of 2019 novel coronavirus (COVID-19) pneumonia. Eur Radiol 2020;30:5455–62. https://doi.org/10.1007/s00330-020-06915-5.
[46] Francone M, Iafrate F, Masci GM, Coco S, Cilia F, Manganaro L, et al. Chest CT score in COVID-19 patients: correlation with disease severity and short-term prognosis. Eur Radiol 2020;30:6808–17. https://doi.org/10.1007/s00330-020-07033-y.