



A PSC-tailored deep learning model for liver segmentation on fat-saturated T2-weighted MR: Robustness to hepatic dysmorphia and multicentre generalisability

Elisabetta De Bernardi^{a,b,1,*}, Laura Cristoferi^{c,1}, Paolo Niccolò Franco^d, Daphne D'Amato^a, Davide Ippolito^{a,d}, Pietro Invernizzi^{a,c}, Marco Carbone^{a,e}, Cesare Maino^d

^a Medicine and Surgery Department, University of Milano-Bicocca, Monza, Italy

^b Bicocca Bioinformatics Biostatistics and Bioimaging Centre-B4, School of Medicine and Surgery, University of Milan-Bicocca, Monza, Italy

^c Division of Gastroenterology, Center for Autoimmune Liver Diseases, European Reference Network on Hepatological Diseases (ERN RARE-LIVER), IRCCS Fondazione San Gerardo dei Tintori, Monza, Italy

^d Department of Diagnostic Radiology, IRCCS Fondazione San Gerardo dei Tintori, Monza, Italy

^e Department of Hepatology and Gastroenterology, ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy

ARTICLE INFO

Keywords:

Liver segmentation
Autoimmune liver diseases
Radiomics
Quantitative radiology
Artificial intelligence

ABSTRACT

In this study, we aimed to identify an effective and accurate liver segmentation strategy for fat-saturated T2-weighted MRI in PSC patients, enabling multicentre validation of GLRM-RunEntropy, a radiomics feature which showed prognostic value. We evaluated three segmentation tools (Total Segmentator, Philips MR Liver Health, and an in-house trained nnU-Net model) across 183 PSC patients from internal and external cohorts. Ten segmentation algorithms derived from these tools were quantitatively assessed using DSC and HD against reference liver contours. The nnU-Net model trained on PSC-specific T2w-SPIR images (algorithm #10) outperformed all others, achieving a mean DSC of 95 % and mean HD of 25.7 mm. Total Segmentator (algorithm #1) showed strong DSC (90 %) but was more affected by variations in liver morphology. Qualitative evaluation by expert radiologists confirmed the superior inclusion of liver parenchyma and exclusion of non-hepatic tissues by the nnU-Net model, together with a lower sensitivity to liver stiffness and robustness to cohort variability. Our findings highlight the importance of domain-specific training for robust segmentation in the setting of chronic liver disease imaging. The validated nnU-Net segmentation model will be made available to support large-scale multicentre validation of GLRM-RunEntropy as a prognostic biomarker in PSC.

1. Introduction

Radiomics, i.e. the quantitative analysis of medical images, can potentially identify imaging biomarkers with prognostic value. When integrated with clinical data, these biomarkers have the potential to improve patient stratification and management [1]. This approach can be particularly helpful for those diseases in which imaging is integral part to diagnosis, risk stratification, and follow-up, such as primary sclerosing cholangitis (PSC), a rare cholestatic liver disease with highly variable prognosis and no approved medical therapy.

In a recent study involving a cohort of 99 patients with PSC, we demonstrated the prognostic potential value of a liver parenchyma radiomics feature, GLRM-RunEntropy, extracted from Fat-Saturated T2-

Weighted (T2W) Magnetic Resonance (MR) images [2]. This feature demonstrated strong associations with two key clinical biomarkers: the Mayo Risk Score (MRS) [OR = 3.90, 95 % CI 1.46–10.42, $p = 0.007$] and liver stiffness measurement (LSM) [OR = 2.97, 95 % CI 1.33–6.66, $p = 0.008$]. It also demonstrated meaningful predictive power for clinical outcomes, with a hazard ratio of 1.480 per 0.1-unit increase (95 % CI 1.226–1.786) and a C-index of 0.857. These promising results were, however, obtained on a relatively small for a rare disease, monocentric dataset of images acquired on a single MR scanner (Philips Ingenia, Philips Healthcare, Best, The Netherlands), employing a standardized protocol, including Fat Sat T2 Weighted–MultiVane (T2 SPIR) sequence. To propose fat-saturated T2W MRI GLRM-RunEntropy as a robust imaging biomarker to identify high-risk PSC patients, further validation on

* Corresponding author at: Medicine and Surgery Department, University of Milano-Bicocca, via Cadore 48, 20900 Monza, Italy.

E-mail address: elisabetta.debernardi@unimib.it (E. De Bernardi).

¹ Co-first authors.

internal and external cohorts, ideally acquired on diverse MRI systems, is essential. The computation of the GLRM-RunEntropy within a liver Region Of Interest (ROI) can be efficiently and repeatably performed using Pyradiomics, a free IBSI-compliant software [3]. Our previous work demonstrated that the prognostic value of GLRM-RunEntropy depends on the ROI covering the whole liver, with major vessels and large bile ducts excluded. Validating this approach at scale therefore requires an efficient and accurate method for segmenting the liver on fat-saturated T2W MR images.

Current state-of-the-art techniques for medical image segmentation are based on Convolutional Neural Networks (CNNs) and related deep learning (DL) architectures [4]. The research community provides both pretrained and customizable (i.e. trainable) segmentation tools; commercial vendors are also progressively incorporating AI-based segmentation tools into clinical workstations. Pretrained models are ready-to-use; however, their efficacy may be compromised if the dataset characteristics (e.g., pathology, image contrast) diverge from those of the training dataset. This limitation is particularly relevant in PSC, where chronic biliary inflammation and progressive fibrosis lead to hepatic dysmorphism (including lobar atrophy, compensatory hypertrophy, and parenchymal heterogeneity), that deviates substantially from the liver morphology encountered in more common conditions such as hepatocellular carcinoma or metabolic-associated steatosis liver disease, on which most available segmentation models have been trained. Moreover, most liver segmentation tools available to the scientific community are specific to CT images or contrast-enhanced T1W MRI, and none has been developed or validated for fat-saturated T2W in PSC patients. Conversely, trainable models necessitate additional expertise and resources but can be tailored to specific pathologies or imaging protocols.

Based on these considerations, this study aims to identify an effective and accurate liver segmentation strategy for fat-saturated T2W MR images in PSC patients, with the goal of recommending it for use in the multicenter validation of GLRM-RunEntropy. The ideal strategy must capture the entire hepatic parenchyma, including areas affected by morphological variations or varying degrees of fibrosis, to ensure these regions contribute appropriately to the radiomic feature calculation. At the same time, it must reliably exclude non-parenchymal structures to avoid artifacts in the computation of GLRM-RunEntropy.

We evaluated three distinct tools: 1) Total Segmentator [5], a pretrained model developed on diverse MR sequences and pathologies (i.e., not specific to fat-saturated T2W MRI or PSC); 2) Philips MR Liver Health (Philips Healthcare, Best, The Netherlands), a vendor-provided liver segmentation tool designed for contrast-enhanced images, thus necessitating image registration to segment fat-saturated T2W images; 3) nnU-Net [6], a trainable model ensemble, trained in-house on a PSC fat-saturated T2W MRI cohort.

2. Materials and methods

2.1. Patient cohorts

Patient characteristics. A total of 183 patients followed at the Center for Autoimmune Liver Diseases of Fondazione IRCCS San Gerardo dei Tintori were included in the study. Selected patients had a PSC diagnosis established according to European Association for the Study of the Liver (EASL) guidelines [7] and age ≥ 18 years. Exclusion criteria were: 1) small duct PSC, 2) known hepatic comorbidities, 3) previous surgery on biliary tree, 4) liver transplantation, 5) hepatic or biliary neoplasms, and 6) decompensated cirrhosis before or at the time of MRI.

135 (73.8 %) of the patients included in the study underwent MRI examinations at Fondazione IRCCS San Gerardo dei Tintori using a 1.5 T Philips Ingenia MRI scanner (Philips Healthcare, Best, The Netherlands) with a standardized protocol and in line with European Society of Gastrointestinal and Abdominal Radiology (ESGAR) [8]. These patients constituted the internal dataset. By contrast, 48 (26.2 %) patients underwent 1.5 T MRI at 45 different external institutions, with different

MRI scanner, and constituted the external dataset.

For each patient, liver stiffness measurement (LSM) performed within 3 months before or after MRI were collected. LSM was collected using transient elastography (FibroScan 502 Touch, Echosens, Paris, France) performed by trained gastroenterologists.

Internal cohort. From the standardized MRI protocol, MRI sequences enumerated in Table 1 were considered for the present study.

The 135 MRI studies were divided chronologically into three groups: Group A (n = 62, 45.9 %, training), Group B (n = 50, 37 %), and Group C (n = 23, 17.1 %). As it will be detailed in the following, Group A studies were used for model training (first and second assessment); Group B studies for testing (first assessment) and for training (second assessment); Group C studies for testing (second assessment).

External cohort. The 48 external cohort studies were acquired at 45 different MR centres, 24 (50 %) on Philips scanners, 6 (12.5 %) on GE scanners, and 18 (37.5 %) on Siemens scanners. Axial fat-saturated T2 weighted sequences were considered. Echo time (TE) ranged from 76 msec to 123 msec, with an average of 88 msec; repetition time (TR) ranged from 303 msec to 11,250 msec, with an average of 3069 msec. Details are provided in Supplementary Materials.

2.2. Reference liver segmentation

For the 112 studies included in Groups A and B of the internal cohort, reference liver contours for the T2W-SPIR sequence were generated through a multi-step annotation workflow. Details are provided in Supplementary Materials.

2.3. Segmentation tools

Three segmentation tools were considered in this study. 1) Total Segmentator, an open-source tool that segments most major anatomical structures in CT or MR image [5]. Total Segmentator was trained on a wide range of CT and MR images (e.g. different scanners, institutions, protocols) and is not specific to any MR sequence or pathology. 2) MR Liver Health, a proprietary tool implemented in Philips Advanced Visualization Workspace 15, which provides a dedicated workflow for automatic liver segmentation on MR images using an AI model trained on contrast-enhanced T1-weighted (T1W) sequences. 3) nnU-Net, an open-source segmentation tool that automatically configure itself (pre-processing, network architecture, training and post-processing) for any given task and dataset [6].

2.4. Explored algorithms

Utilizing the three above-reported segmentation tools, we finally evaluated a total of 10 segmentation algorithms, as detailed in Table 2.

Spatial registrations were all performed with the ANTsPy library with mutual information as optimization metric. Spatial transformations able to warp Out Phase T1W and T1W venous images to T2-SPIR were estimated and subsequently applied to corresponding liver segmentations.

2.5. First assessment – quantitative evaluation of the ten algorithms

In the first assessment, the ten segmentation algorithms were quantitatively evaluated on the 50 patients in the Group B of the internal cohort. The nnU-Net model was trained automatically using data from the 62 patients in the Group A of the internal cohort. Segmentation results were quantitatively compared to the reference ground truth using the Dice Similarity Coefficient to assess spatial overlap (DSC, expressed as percentage, range 0 % – 100 %), and the Hausdorff Distance to evaluate boundary accuracy (HD, expressed in mm). For both metrics, minimum, maximum, mean, standard deviation and median values were computed.

Table 1

MRI sequences considered for the internal cohort. T1W venous and T1W HBP were acquired after the injection of 0.1 mL/kg body weight Gd-EOB-DTPA (Primovist, Bayer HealthCare Pharmaceuticals).

Sequence (Acronym)	TR (msec)	TE (msec)	FA (°)	FOV (mm x mm)	Slice Thickness (mm)	Matrix
Axial fat-saturated T2 weighted SPIR Multivane (T2-SPIR)	2352	80	90	350x350	5	252x252
Axial Out Phase T1 weighted (Out Phase T1W)	160	2.3–4.6	80	360x320	5	256x172
Axial dynamic T1 weighted (T1W) 3D GRE in portal-venous (T1W venous) and in hepatobiliary (20 min) phases (T1W HBP)	4	1.94	10	360x320	5	256x202

Table 2

Detail of the explored segmentation algorithms.

Algorithm #	Tool	MRI sequence for tool application	Spatial registration to T2-SPIR
1	Total Segmentator	T2W-SPIR	/
2	Total Segmentator	Out Phase T1W	No
3	Total Segmentator	Out Phase T1W	Rigid
4	Total Segmentator	Out Phase T1W	Affine
5	Total Segmentator	Out Phase T1W	Elastic
6	MR Liver Health	T1W portal-venous	No
7	MR Liver Health	T1W portal-venous	Rigid
8	MR Liver Health	T1W portal-venous	Affine
9	MR Liver Health	T1W portal-venous	Elastic
10	nnU-Net	T2W-SPIR	/

2.6. Second assessment – qualitative evaluation of two best performing algorithms

In the second assessment, the two best performing algorithms (#1 and #10, as detailed in the Results section) were applied to the 23 patients in the Group C of the internal cohort (internal test cohort) and to the 48 patients of the external cohort (external test cohort). For this evaluation, the nnU-Net model was retrained using a combined dataset of 112 patients from Groups A and B.

Two board-certified abdominal radiologists with 10 and 6 years of experience in body MR imaging (C.M. and P.N.F., respectively) independently evaluated MR images and segmentations in a blinded fashion, in two separate reading sessions. Firstly, for each study, they evaluated the global image quality, considering liver intensity homogeneity, liver margin definition, perceived blur, perceived noise, efficacy and homogeneity of fat suppression, presence of artifacts. Successively, segmentations were reviewed slice-by-slice with specific attention to anatomically and clinically relevant regions. In particular, radiologists assessed: 1) the accuracy of liver contour delineation at the hepatic margins, distinguishing between acceptable minor contour inaccuracies and clinically relevant under- or over-segmentation; 2) the correct inclusion or exclusion of extra-hepatic tissues, with particular focus on perihepatic fat, gallbladder, and the inferior vena cava (suprahepatic, intrahepatic, and intrahepatic segments); 3) the segmentation of the portal vein at the hepatic hilum, evaluating whether the main portal trunk was erroneously included within the liver parenchyma; 4) the handling of major intrahepatic vascular and biliary structures, specifically the principal intrahepatic portal branches and bile ducts, which should be excluded from the parenchymal region of interest; 5) the delineation of the hepatic dome and the interface with adjacent structures, including the diaphragm and the heart, which are frequent sources of contouring errors in superior liver segments; 6) the exclusion of the spleen, especially in patients with anatomical variants such as

“beaver tail liver”, in whom splenic tissue may be closer to or partially overlapping the left hepatic lobe. These criteria were used by both readers to assign 5-point Likert-scale scores for liver parenchyma inclusion and for exclusion of non-hepatic tissues, thus ensuring that the qualitative assessment reflected clinically meaningful segmentation performance.

2.7. Second assessment – qualitative evaluation of two best performing algorithms – statistical analysis

As to image quality, inter-rater agreement was statistically assessed using the quadratic weighted Cohen’s kappa. Scores assigned to internal test cohort and external test cohort were compared with the non-parametric Wilcoxon rank-sum test. As to segmentations, for each rater, the scores assigned to the two algorithms were compared using the non-parametric Wilcoxon signed-rank test with normal approximation due to tied pairs. Inter-rater agreement was statistically assessed using the quadratic weighted Cohen’s kappa. Subsequently, a mixed-effects model was fitted to evaluate, for both algorithms, the dependence of segmentation quality on the degree of fibrosis (quantified by LSM) and on the patient cohort (internal or external). This modelling approach was deliberately adopted to de-confound these two factors, given that the distribution of LSM values differed between the two test cohorts (see Results section). The model is defined in equation (1):

$$\begin{aligned} \text{Quality} = & \beta_0 + \beta_1 \text{LSM} + \beta_2 \text{Cohort} + \beta_3 \text{Algorithm} + \beta_4 (\text{LSM} \times \text{Algorithm}) \\ & + \beta_5 (\text{Cohort} \times \text{Algorithm}) + u_{\text{patient}} + u_{\text{rater}} + \varepsilon \end{aligned} \quad (1)$$

Fixed effects included liver stiffness, cohort (internal vs external), and algorithm, along with algorithm–stiffness and algorithm–cohort interaction terms. The algorithm–stiffness interaction was used to assess segmentation quality dependence on liver fibrosis, while the algorithm–cohort interaction evaluated segmentation quality robustness to scanner and protocol changes. Random effects included patient and rater, to account for paired measurements and inter-rater variability. In the model implementation algorithm, algorithm #1 is the baseline algorithm and the internal cohort is the baseline cohort, i.e. Algorithm = [0 for alg #1, 1 for alg#10], Cohort = [0 for internal cohort, 1 for external cohort]. LSM is normalized to average = 0. Therefore, the meaning of the coefficients $\beta_0 \dots \beta_5$ is as reported in Table 3.

Table 3

Meaning of the mixed effect model coefficients.

Model coefficient	Meaning
β_0	Baseline quality when cohort = internal, algorithm=#1
β_1	LSM slope (how quality changes per unit of LSM) for algorithm #1 and internal cohort
β_2	Cohort effect (difference in average quality between external and internal cohort) for algorithm #1
β_3	Algorithm effect (difference in average quality between algorithm #10 and algorithm #1) on the internal cohort
β_4	Change in LSM slope for algorithm #10
β_5	Change in cohort effect for algorithm #10

3. Results

3.1. Quantitative assessment of the ten segmentation algorithms

Results of the quantitative assessment of the 10 segmentation algorithms detailed in Table 1 on Group B patients of the internal cohort are presented in Table 4.

DSC and HD results show considerable variability in segmentation performance across the ten algorithms. The optimal performance corresponds to higher DSC and lower HD values. Algorithm #10 (nnU-Net applied to T2-SPIR) clearly outperforms all the others, achieving the highest mean DSC (95 %) and the lowest mean HD (25.7 mm), with consistently strong performance across all statistical descriptors. This makes it the top-performing method overall. The second-best DSC performance is observed with algorithm #1 (Total Segmentator applied to T2-SPIR), which reaches a mean DSC of 90 %, close to that of algorithm #10. However, its HD metrics are considerably worse, with a mean of 50.0 mm, indicating segmentation inaccuracies in spatial distance compared to algorithm #10. When examining the HD metrics alone, no single algorithm clearly emerges as the second-best. Therefore, we selected algorithms #10 and #1 for further assessment: algorithm #10 for its superior and consistent performance across all metrics, algorithm #1 for its strong DSC performance, indicating good overlap accuracy.

3.2. Qualitative assessment of the two best performing algorithms – internal test cohort

Fig. 1 shows the qualitative evaluation by the two independent raters of liver segmentations generated by the two selected algorithms (#1 and #10) in the internal test cohort (Group C).

With respect to both liver parenchyma inclusion and exclusion of non-hepatic tissues, both raters judged algorithm #10 as superior to algorithm #1, with significant differences (Table 5). Inter-rater agreement was almost perfect in the assessment of liver parenchyma inclusion for algorithm #1, moderate in the assessment of liver parenchyma inclusion for algorithm #10, substantial in the assessment of non-hepatic tissue exclusion for both algorithms.

Fig. 2 presents the paired slope plot of the average ratings across the two raters. For liver parenchyma inclusion, all segmentations showed improvement when moving from algorithm #1 to algorithm #10, except for one segmentation that received the same rating for both algorithms. For the exclusion of non-hepatic tissue, only one segmentation exhibited a worse rating, while all others demonstrated either improved or unchanged ratings.

3.3. Qualitative assessment of the two best performing algorithms – external test cohort

Fig. 3 shows the qualitative evaluation by the two independent raters of liver segmentations generated by the two selected algorithms (#1 and

#10) in the external test cohort.

On the external dataset, both raters judged algorithm #10 to be significantly superior to algorithm #1 in terms of both liver parenchyma inclusion and exclusion of non-hepatic tissues. Inter-rater agreement was substantial in all the assessments (Table 6).

Fig. 4 presents the paired slope plot of the average ratings across the two raters. For liver parenchyma inclusion, only one segmentation showed a deterioration when moving from algorithm #1 to algorithm #10; for the exclusion of non-hepatic tissue, three segmentations exhibited a worse rating for algorithm #10.

3.4. Qualitative assessment of the two best performing algorithms – robustness towards fibrosis grade and patient cohort

In the assessment of image quality, the two raters substantially agree (Cohen's Kappa = 0.74). The distribution of mean Likert scores does not significantly differ between the two test cohorts (Wilcoxon rank-sum test P-value = 0.08).

Fig. 5 represents the distribution of LSM values in the internal and external test cohorts. The distributions differ significantly (Wilcoxon rank-sum test, $p = 0.017$), supporting the use of a mixed-effects model to evaluate the effects of liver stiffness and patient cohort on the segmentation quality of the two algorithms.

Table 7 reports the results of the fitting of the mixed-effects models for liver parenchyma inclusion and exclusion of non-hepatic tissues.

LSM impacts the capacity of algorithm #1 to encompass the entire liver parenchyma in its segmentation process. In contrast, the ability of algorithm #10 to include the entire liver parenchyma in segmentation is less influenced by LSM. LSM does not significantly affect the capability of either algorithm to exclude non-hepatic tissues.

Algorithm #1 demonstrates a superior ability to exclude non-hepatic tissues in the external cohort compared with the internal cohort, whereas Algorithm #10 achieves comparable performance across the two cohorts for both the aspects considered, inclusion of liver parenchyma and exclusion of non-hepatic tissues,

In Fig. 6, the segmentation quality change with liver stiffness is represented for the two algorithms.

Algorithms #1 and #10 perform equivalently in patients with near-normal liver morphology and texture (i.e., early-stage PSC); approach #10 outperforms approach #1 in patients characterized by distorted liver morphology and extensive fibrosis. Four representative examples of patients in which approach #10 outperforms #1 are shown in Fig. 7. Contours obtained with approach #1 are represented in green, those obtained with approach #10 are represented in red.

4. Discussion

This study aimed to identify an effective strategy for segmenting the liver on fat-saturated T2W (T2W-SPIR) MR images in patients with PSC, with the objective of supporting colleagues interested in computing and

Table 4

Quantitative assessment of the ten segmentation algorithms on Group B patients of the internal cohort. For each metric column, the two best values are highlighted in bold.

Algorithm #	DSC					HD (mm)				
	Min	Max	Mean	Std/Mean	Median	Min	Max	Mean	Std/Mean	Median
1	80 %	95 %	90 %	4 %	91 %	17.3	134.6	50.0	59 %	34.2
2	3 %	94 %	83 %	17 %	87 %	15.0	137.6	50.2	57 %	38.0
3	38 %	94 %	82 %	15 %	86 %	20.5	120.6	50.1	53 %	38.6
4	45 %	93 %	83 %	11 %	86 %	21.1	122.6	50.2	51 %	41.0
5	44 %	93 %	85 %	10 %	87 %	20.4	121.8	49.7	51 %	40.6
6	46 %	94 %	83 %	12 %	87 %	22.8	118.8	51.8	38 %	46.6
7	62 %	95 %	86 %	8 %	87 %	20.1	122.3	50.7	40 %	45.2
8	70 %	94 %	85 %	7 %	86 %	19.8	120.6	52.1	38 %	45.8
9	69 %	93 %	85 %	7 %	86 %	20.0	120.6	52.0	38 %	46.6
10	85 %	98 %	95 %	3 %	96 %	11.4	117.4	25.7	64 %	21.9

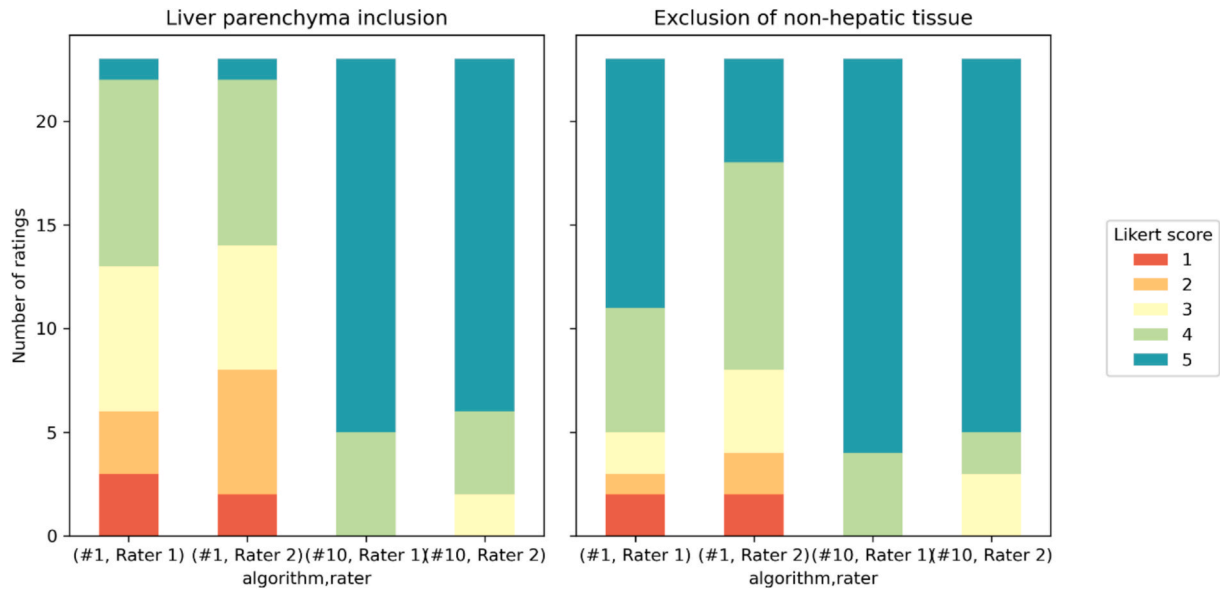


Fig. 1. Qualitative evaluation by rater 1 and 2 of liver segmentations obtained with algorithms #1 and #10 in the internal test cohort.

Table 5

Quantitative comparison of scores given by the two raters to segmentations in the internal test cohort.

	Liver parenchyma inclusion	Exclusion of non-hepatic tissue
Algorithm #1 vs #10 – Rater 1 (Wilcoxon signed-rank test)	4.3*1e-05	0.01
Algorithm #1 vs #10 – Rater 2 (Wilcoxon signed-rank test)	6.6*1e-05	8.8*1e-04
Inter-rater agreement – #1 (Cohen's kappa)	0.85	0.71
Inter-rater agreement – #10 (Cohen's kappa)	0.48	0.73

validating radiomic features that may hold prognostic value. To our knowledge, this is the first study specifically addressing liver segmentation on fat-saturated T2W MRI in patients with PSC. Unlike existing approaches, which are typically based on CT or contrast-enhanced T1W MRI, our work focuses on a sequence and disease that are still less explored. PSC livers are often highly dysmorphic, which makes them particularly challenging to segment accurately. This inherent anatomical variability and structural distortion provide a strong rationale for this study.

Currently, the most successful segmentation strategies for medical images are based on deep learning models trained on large datasets with manually annotated ground truth. However, most liver segmentation tools available to the scientific community, both open-source and commercial, are specific to CT images. A notable example is the Liver Tumor Segmentation Challenge (LiTS) [9], which significantly advanced the field of liver CT segmentation. In the MRI domain, segmentation algorithms are fewer and predominantly designed for T1W images, which provide clearer anatomical and perfusion information when contrast-enhanced [10,11]. In PSC, a disease leading to chronic inflammation and progressive fibrosis of the peribiliary liver

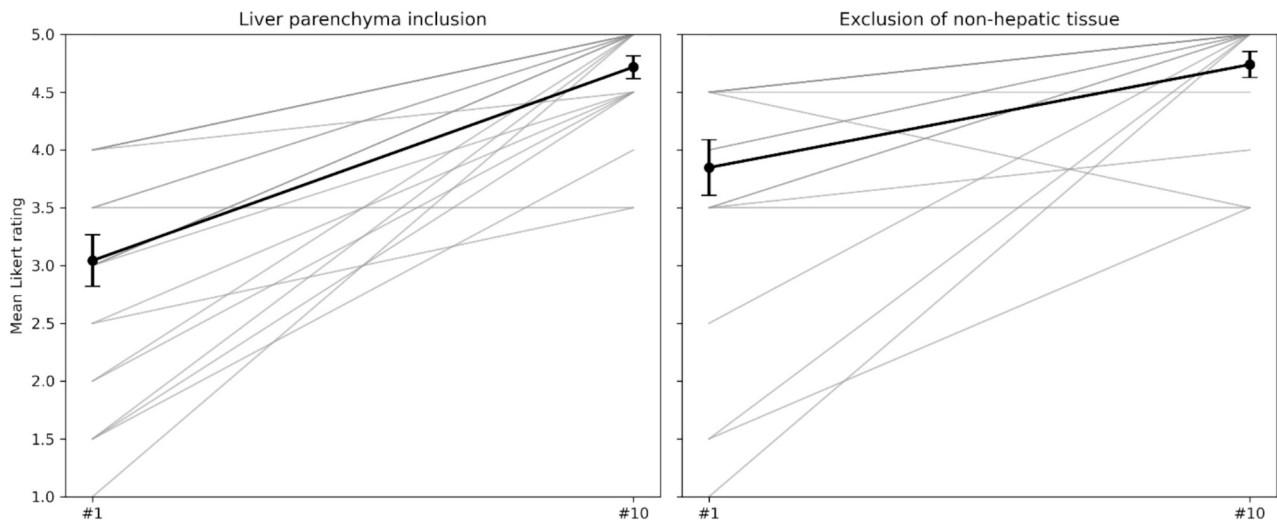


Fig. 2. Paired slope plot of the average ratings across the two raters on the internal test cohort, with the mean Likert score highlighted in bold.

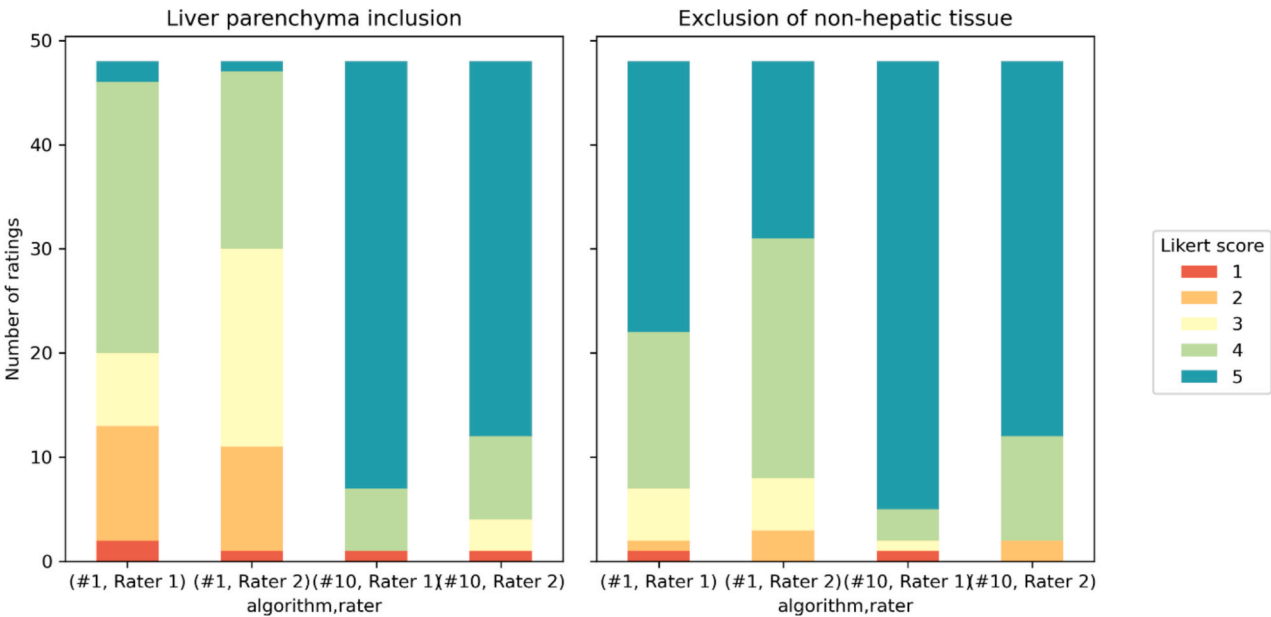


Fig. 3. Qualitative evaluation by rater 1 and 2 of liver segmentations obtained with algorithms #1 and #10 in the external test cohort.

Table 6
Quantitative comparison of scores given by the two raters to segmentations in the external test cohort.

	Liver parenchyma inclusion	Exclusion of non-hepatic tissue
Algorithm #1 vs #10 – Rater 1 (Wilcoxon signed-rank test)	4.7*1e-09	1.7*1e-03
Algorithm #1 vs #10 – Rater 2 (Wilcoxon signed-rank test)	8.7*1e-09	2.9*1e-04
Inter-rater agreement – #1 (Cohen's kappa)	0.76	0.62
Inter-rater agreement – #10 (Cohen's kappa)	0.68	0.8

parenchyma, fat-sat T2W sequences, such as T2W-SPIR, are specifically designed to enhance thickened, inflamed, or fibrotic bile duct walls. Fat-sat T2W sequences do not always clearly depict overall liver anatomy. Consequently, segmentation strategies for the liver on this sequence have not traditionally been developed. Furthermore, given the rarity of PSC and its distinct imaging appearance, segmentation algorithms trained on more common liver diseases (e.g., hepatocellular carcinoma, fatty liver) may perform sub-optimally in these patients, especially due to the dysmorphic appearance of hepatic parenchyma.

With these considerations in mind, we initially tested TotalSegmentator, an open-source, pretrained CNN ensemble not specifically trained on T2W-SPIR images or PSC patients. Despite these limitations, TotalSegmentator is readily accessible and easy to use, making it an attractive initial candidate. To compensate for the lower quality and contrast of T2W-SPIR images, we considered applying TotalSegmentator to higher-quality Out Phase T1W images followed by spatial registration to T2W-SPIR. However, this approach yielded inferior results compared to directly applying TotalSegmentator to the T2W-SPIR images. While direct application generally produced acceptable segmentation in cases with typical liver morphology, it failed in more complex scenarios (such

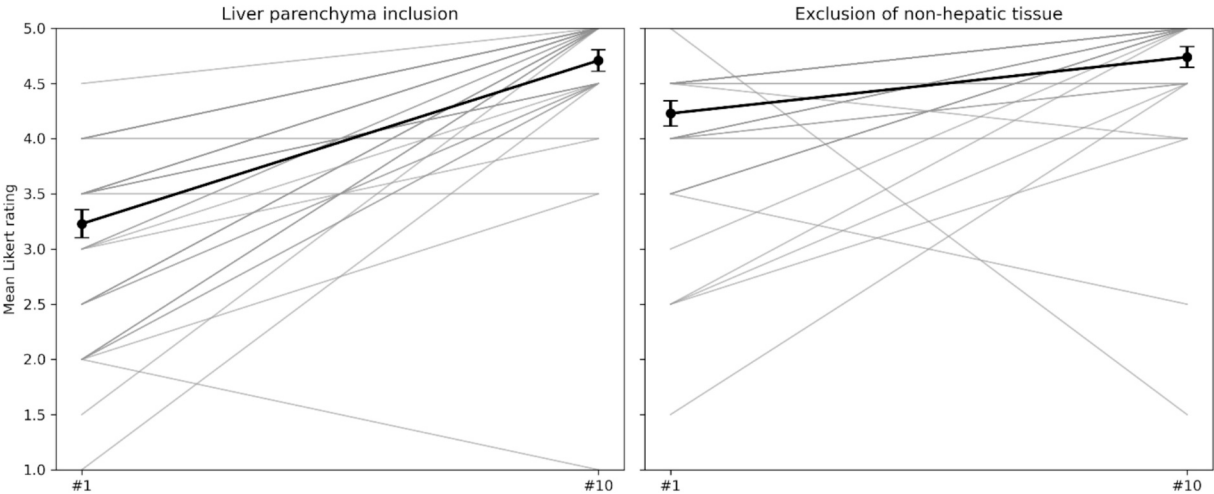


Fig. 4. Paired slope plot of the average ratings across the two raters on the external test cohort, with the mean Likert score highlighted in bold.

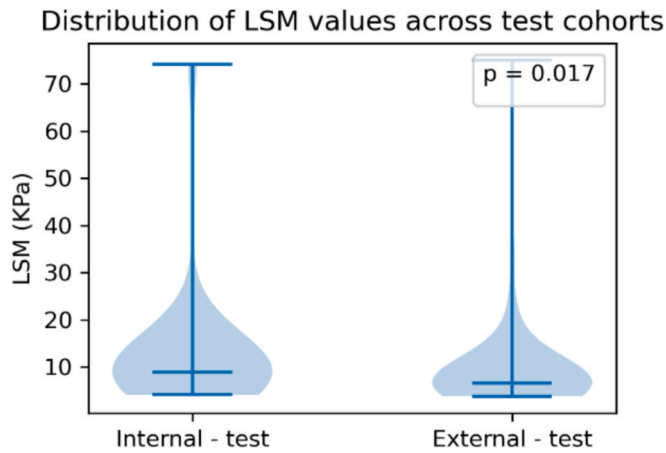


Fig. 5. Violin plot representing the distribution of LSM values in the internal and in the external test cohort. Internal test cohort: median = 8.0 KPa, IQR = 5.0 KPa; external test cohort: median = 6.6 KPa, IQR = 4.4 KPa. The P-value of the Wilcoxon Rank-Sum Test is shown in the legend.

Table 7

Results of the mixed-effects model fitting for liver parenchyma inclusion (left) and exclusion of non-hepatic tissue (right). For each coefficient, estimated value, 95 % CI and P-value are reported.

	Liver parenchyma inclusion			Exclusion of non-hepatic tissue		
	Estimated value	95 % CI	P-value	Estimated value	95 % CI	P-value
β_1	-0.03	[-0.04–0.01]	0.000	-0.01	[-0.02–0.00]	0.286
β_2	0.11	[-0.18–0.40]	0.450	0.36	[0.06–0.66]	0.018
β_3	1.64	[1.34–1.93]	0.000	0.87	[0.55–1.19]	0.000
β_4	0.02	[0.00–0.03]	0.007	0.01	[-0.00–0.03]	0.205
β_5	-0.14	[-0.50–0.22]	0.446	-0.35	[-0.74–0.04]	0.075

as non-standard liver shapes or extensive fibrosis), likely due to the absence of such cases in its training dataset. This limitation is understandable, as TotalSegmentator is a general-purpose tool designed to segment 50 anatomical structures rather than being optimized for liver

pathology. We also evaluated a vendor-specific tool developed for contrast-enhanced T1W liver imaging, a practical choice for institutions using Philips MRI scanners. This tool performed better in managing abnormal liver morphologies, as indicated by lower Hausdorff distance values. However, the necessary image registration step to align segmentations with T2W-SPIR images again proved unsatisfactory, limiting its utility for our specific purpose. Consequently, we trained a dedicated model using nnU-Net, the same framework on which TotalSegmentator is based, to create a customized liver segmentation tool for T2W-SPIR images in PSC patients. Initially trained on a cohort of 50 patients with manually delineated ground truth contours, and subsequently on an expanded cohort of 112 patients, the model demonstrated superior performance relative to previous methodologies. This finding underscores the principle that AI models exhibit optimal performance when trained on data that accurately reflect the target domain, specifically the imaging sequence and pathological characteristics of the pathology under study.

These findings also provide insight into the limitations of broadly trained or foundation-style segmentation models when applied to highly specific imaging scenarios. Although TotalSegmentator and the vendor-specific Philips tool achieved overall good performance despite not being trained on PSC T2W-SPIR images, both approaches showed reduced robustness in cases with marked liver dysmorphism, advanced fibrosis, or substantial anatomical distortion. In contrast, the PSC-specific nnU-Net model maintained more stable performance across disease severity levels and across internal and external cohorts. This observation suggests that, for chronic liver diseases with characteristic structural alterations and parenchymal inhomogeneities due to fibro-inflammatory areas, disease- and sequence-specific training may offer advantages over more general-purpose segmentation frameworks, particularly when accurate quantitative downstream analysis is required.

The nnU-Net model was trained on data acquired using a single MR scanner with a standardized T2W-SPIR MultiVane sequence, which generates characteristic image texture and liver edge profile that may differ from those generated by other vendor-specific implementations of the fat-suppressed T2W sequence. Therefore, an important aspect to evaluate was the generalizability of the trained nnU-Net model to external datasets. The external dataset used for validation was highly heterogeneous, comprising studies collected from 45 different imaging centres. Despite this variability, the model demonstrated robust performance as to both the inclusion of liver parenchyma and the exclusion of non-hepatic structures, with a cohort effect lower than that observed

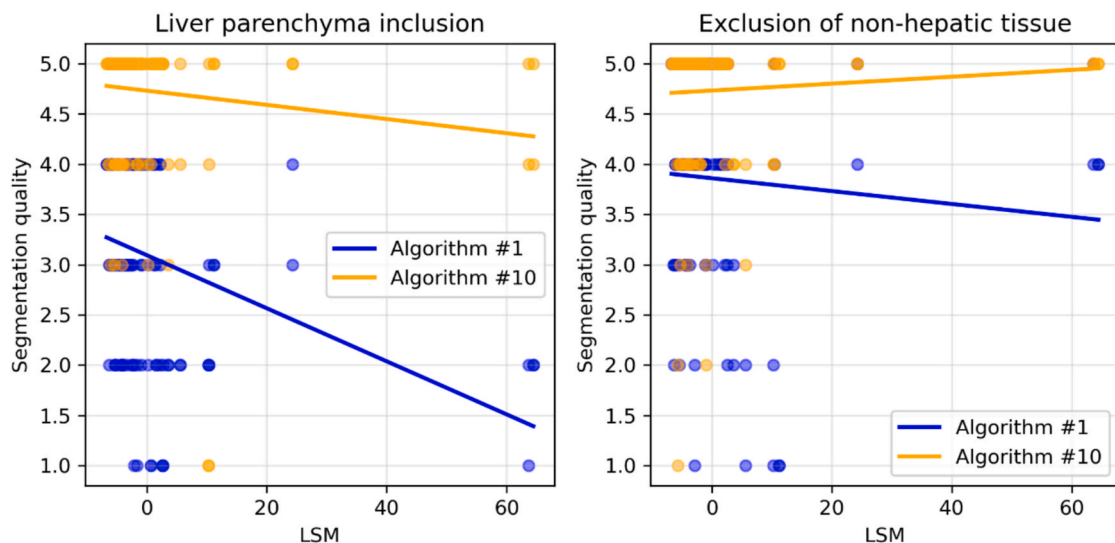


Fig. 6. Dependence of segmentation quality scores on liver stiffness for algorithms #1 and #10.

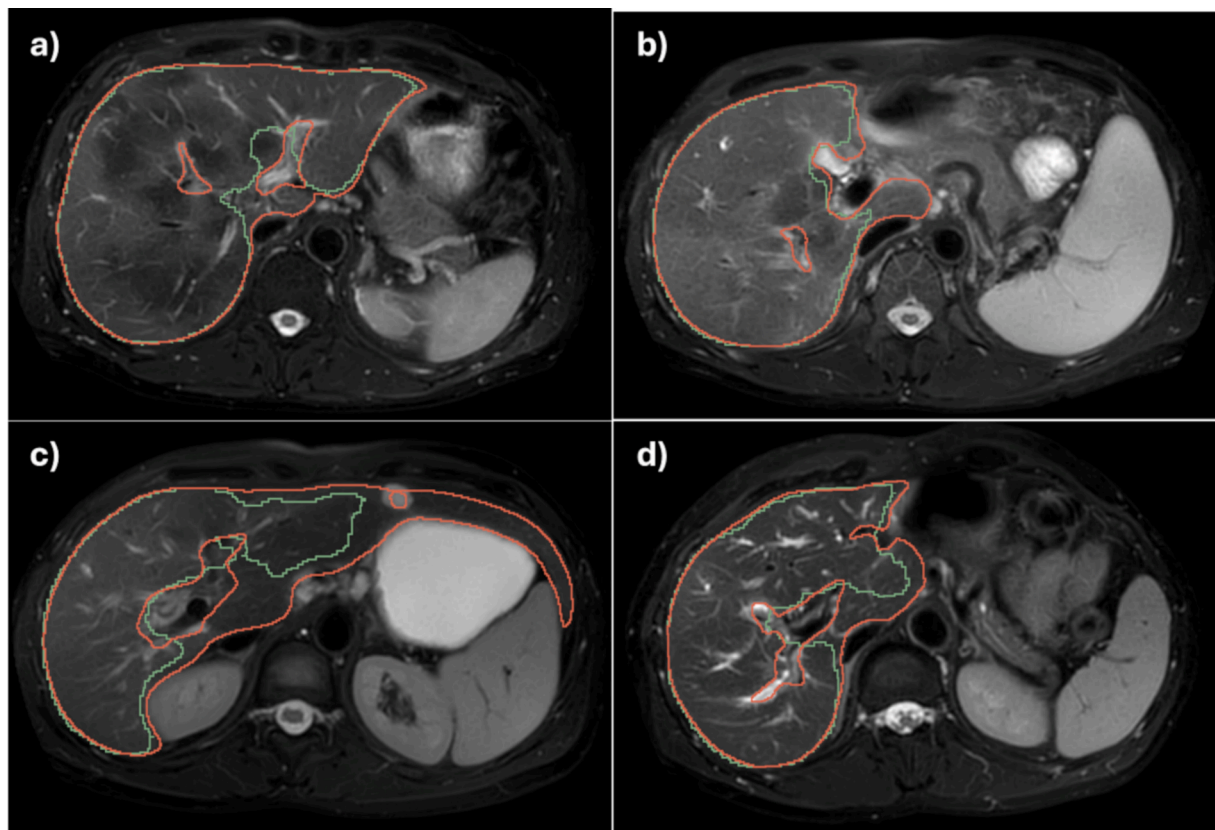


Fig. 7. Axial slices of four patients of the internal test cohort with superimposed liver contours obtained with algorithm #1 (green) and algorithm #10 (red). a) LSM = 11.4 Kpa; mean score liver parenchyma inclusion = 4 algorithm #1, 5 algorithm #10; mean score exclusion non-hepatic tissues = 5 algorithm #1, 5 algorithm #10. b) LSM = 11.7 Kpa; mean score liver parenchyma inclusion = 3.5 algorithm #1, 5 algorithm #10; mean score exclusion non-hepatic tissues = 4.5 algorithm #1, 5 algorithm #10. c) LSM = 13.2 Kpa; mean score liver parenchyma inclusion = 1.5 algorithm #1, 5 algorithm #10; mean score exclusion non-hepatic tissues = 4.5 algorithm #1, 5 algorithm #10. d) LSM = 7.9 Kpa; mean score liver parenchyma inclusion = 3 algorithm #1, 5 algorithm #10; mean score exclusion non-hepatic tissues = 4.5 algorithm #1, 5 algorithm #10.

for TotalSegmentator, which is inherently designed for broad generalizability, having been trained on a large multicentre dataset encompassing diverse scanners, imaging protocols, and pathological conditions. The good performance observed in the external cohort is particularly encouraging given the substantial variability in scanner vendors and acquisition protocols encountered in routine clinical practice. A high degree of robustness to scanner and protocol variability is a desirable property for potential future multicentre applications.

From a radiomics perspective, the precision of liver segmentation is a pivotal factor influencing feature reliability. Reducing segmentation variability may also improve the reproducibility of longitudinal radiomics measurements, an important prerequisite for future clinical implementation of imaging biomarkers. Adding large vessels or bile ducts may make heterogeneity seem higher than it really is, while leaving out fibrotic or morphologically changed areas may lead to underestimate disease burden seem. In this context, the improved ability of the nnU-Net model to accurately capture the full liver parenchyma while excluding non-hepatic tissues is expected to increase the robustness and biological relevance of the extracted radiomic features. Therefore, the proposed segmentation strategy can represent a relevant step for the large-scale validation of GLRM-RunEntropy as a prognostic biomarker in patients with PSC [2].

The proposed model is based on the widely adopted nnU-Net framework and performs fully automated liver segmentation, without the need for manual refinement in most instances. The inference time for a single study is on the order of a few minutes, making the approach compatible with routine research workflows and large-scale multicentre radiomics studies. The utilization of the standardized open-source nnU-

Net framework facilitates reproducibility and enables the potential sharing and deployment of the trained model across research centres. Nonetheless, practical deployment considerations, including automated processing pipelines implementation, definition of hardware requirements and integration into analysis frameworks necessitate further work.

In the present study, liver segmentation was primarily developed to enable robust assessment and multicenter validation of GLRM-RunEntropy. An important downstream clinical implication of improved segmentation accuracy is the potential for reliable longitudinal monitoring of GLRM-RunEntropy over time, once its prognostic value has been fully validated. However, a segmentation tool of this kind may also support several additional quantitative imaging applications requiring whole-liver delineation. These include liver volumetry, quantitative assessment of liver morphology and surface irregularity, fibrosis-related imaging analysis, and the extraction and validation of additional radiomics biomarkers. The proposed model is currently optimized for fat-saturated T2-weighted images only; therefore, its applicability to other MRI sequences (e.g., T1-weighted acquisitions) should be specifically evaluated. It is likely that additional model fine-tuning or retraining on sequence-specific datasets would be required to ensure comparable performance across different imaging contrasts.

This study has some limitations. Liver contours were initially delineated on T1W-HBP images and subsequently propagated to T2W-SPIR images, due to the superior contrast of T1W-HBP images for more accurate boundary definition. This approach may introduce small spatial inconsistencies due to registration errors. However, all segmentations were refined and reviewed by expert radiologists, which likely

minimized this effect. In addition, the reference contours were generated through a multi-step workflow consisting of semi-automatic segmentation, manual refinement by a first experienced abdominal radiologist, and subsequent review by a second one. A formal quantitative assessment of interobserver variability for the reference segmentations was not performed and should therefore be considered a further limitation of this study. Ideally, the second radiologist would have had to independently refine all segmentations, followed by a consensus process to establish the final reference standard. However, the substantial time commitment required for such an approach was not compatible with the clinical workload of the participating experts. To mitigate this limitation, the second expert review step was specifically incorporated to account for potential reader variability, reduce individual observer bias, and improve the consistency of the reference standard. It should be noted that interobserver variability was instead explicitly addressed in the qualitative evaluation of the test cohort segmentations, where independent assessments from two readers were included into the analysis.

Additional limitations should be acknowledged. Although this study included both internal and external multicentre cohorts, the overall dataset size remains relatively limited compared with large-scale deep learning studies, reflecting the rarity of PSC and the challenges associated with collecting manually annotated MRI datasets in this population. Furthermore, demographic diversity was not specifically controlled or stratified, and the cohorts reflected the patient population followed in an Italian referral Center. Therefore, additional validation on larger and more demographically heterogeneous populations will be important to further assess model generalizability across more heterogeneous groups.

5. Conclusions

This study identifies and validates a disease-specific deep learning model for accurate liver segmentation on fat-saturated T2-weighted MR images in patients with PSC, a population characterised by marked hepatic dysmorphism that represents a recognised challenge for general-purpose segmentation tools. By enabling reliable and reproducible whole-liver delineation across multicentre cohorts, this framework lays the groundwork for a broader radiomics research programme in PSC. As an immediate application, the model will support the multicentre validation of GLRM-RunEntropy as a prognostic imaging biomarker; more broadly, it provides the infrastructure to systematically explore additional radiomics features of potential clinical relevance in this rare disease with limited therapeutic options. The trained model will be made available to collaborators interested in contributing to these validation efforts.

6. Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Open Paperpal to improve English and readability. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the content of the published article.

7. Ethical section

The study was conducted in accordance with the guidelines of the Declaration of Helsinki and the principles of good clinical practice. All participants provided written informed consent. The study was approved by the University of Milan-Bicocca Research Ethics Committee (study name: PSC [Database-Clinicaltrial.gov](https://clinicaltrials.gov/ct2/show/study/NCT05618145) ID NC705618145).

CRedit authorship contribution statement

Elisabetta De Bernardi: Writing – original draft, Software, Project

administration, Methodology, Investigation, Formal analysis, Conceptualization. **Laura Cristofori:** Writing – review & editing, Investigation, Data curation, Conceptualization. **Paolo Niccolò Franco:** Writing – review & editing, Investigation. **Daphne D'Amato:** Writing – review & editing, Data curation. **Davide Ippolito:** Writing – review & editing, Supervision. **Pietro Invernizzi:** Writing – review & editing, Supervision. **Marco Carbone:** Writing – review & editing, Supervision, Conceptualization. **Cesare Maino:** Writing – original draft, Supervision, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by 1) Dipartimenti di Eccellenza 2023-2027 IMPACT MEDICINE project of the Department of Medicine and Surgery, University of Milano-Bicocca; 2) PRIN 2022 project 2022AEFRKC, “Biomarkers of Inflammaging in Liver Diseases (BILD): new directions for basic research and therapeutic approaches in chronic cholangiopathies and cholangiocarcinoma”. Both projects are funded by the Italian Ministry of University and Research (MUR).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ejrad.2026.113010>.

References

- [1] R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: images are more than pictures, they are data, *Radiology* 278 (2) (2016) 563–577.
- [2] L. Cristofori, C. Maino, D.P. Bernasconi, I. Ripamonti, M. Scaravaglio, A. Savino, E. Pesatori, A. Gerussi, E. Nofit, O. Falco, D. D'Amato, F. Gallivanone, C. Alberzoni, R. Viganò, C. Mazzei, L.S. Belli, M.E. Dinelli, M. Mutignani, M.G. Valsecchi, R. Corso, P. Invernizzi, M. Carbone, D. Ippolito, E. De Bernardi, Radiomics-based prognostication in primary sclerosing cholangitis: a proof-of-concept study, *Liver Int.* 45 (11) (2025) e70348.
- [3] J.J.M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G.H. Beets-Tan, J.C. Fillon-Robin, S. Pieper, H.J.W.L. Aerts, Computational radiomics system to decode the radiographic phenotype, *Cancer Res.* 77 (21) (2017) e104–e107.
- [4] D.L. Pham, Xu. Chenyang, J.L. Prince, Current methods in medical image segmentation, *Annu. Rev. Biomed. Eng.* 2 (1) (2000) 315–337.
- [5] Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D., Cyriac, J., Yang, S., Bach, M., Segeroth, M., 2023. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence*. <https://doi.org/10.1148/ryai.230024>.
- [6] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, *Nat. Methods* 18 (2) (2021) 203–211.
- [7] European Association for the Study of the Liver. EASL Clinical Practice Guidelines on sclerosing cholangitis. *J Hepatol.* 2022 Sep;77(3):761–806. doi: 10.1016/j.jhep.2022.05.011. Epub 2022 Jun 21. Erratum in: *J Hepatol.* 2023 Nov;79(5): 1339. doi: 10.1016/j.jhep.2023.09.005. PMID: 35738507.
- [8] Ippolito D, Maino C, Arrivè L, Ba-Salamah A, Cannella R, Furlan A, Grigoriadis A, Pezzullo M, Pöetler Lang S, Schmidt Kobbe S, Vernuccio F, Bali MA. ESGAR consensus statement on MR imaging in primary sclerosing cholangitis. *Eur Radiol.* 2025 Oct;35(10):6495–6506. doi: 10.1007/s00330-025-11583-4. Epub 2025 Apr 26. PMID: 40285815; PMCID: PMC12417234.
- [9] P. Bilic, P. Christ, H.B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, B. Menze, The liver tumor segmentation benchmark (lits), *Med. Image Anal.* 84 (2023) 102680.
- [10] K. Bousabarah, B. Letzen, J. Tefera, L. Savic, I. Schobert, T. Schlachter, L.H. Staib, M. Kocher, J. Chapiro, M. Lin, Automated detection and delineation of hepatocellular carcinoma on multiphasic contrast-enhanced MRI using deep learning, *Abdom. Radiol.* 46 (2021) 216–225, <https://doi.org/10.1007/s00261-020-02604-5>.
- [11] A. Hansch, G. Chlebus, H. Meine, F. Thielke, F. Kock, T. Paulus, N. Abolmaali, A. Schenk, Improving automatic liver tumor segmentation in late-phase MRI using multi-model training and 3D convolutional neural networks, *Sci. Rep.* 12 (2022) 12262, <https://doi.org/10.1038/s41598-022-16388-9>.