Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École doctorale Sciences de la vie et de la santé | ED 414 Université de Strasbourg

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

Institut de médecine translationnelle et maladies hépatiques UMRS 1110

Institut Hospitalo-Universitaire de Strasbourg

THÈSE DE DOCTORAT

présentée par :

Jérémy Dana

soutenue le : **29 août 2024**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Sciences médicales - Recherche clinique et innovation technologique

RISK STRATIFICATION OF HEPATOCARCINOGENESIS USING DEEP LEARNING ANALYSIS OF RADIOLOGICAL IMAGES

THÈSE dirigée par :

Pr Baumert Thomas	PU-PH, Université de Strasbourg
Pr Gallix Benoit	PU-PH, Université de Strasbourg

RAPPORTEURS :

Pr Aubé Christophe	PU-PH, Université d'Angers
Pr Dohan Anthony	PU-PH, Université Paris Cité

AUTRES MEMBRES DU JURY :

Pr Lewin Maïté	PU-PH, Université Paris-Saclay
Pr Lampert Thomas	MdC (Chaire), Université de Strasbourg

Table of Contents

Introduction

Over the last decades, the prevalence of chronic liver diseases and their associated morbidity and mortality markedly increased, especially with the rise of metabolic dysfunction-associated steatotic liver disease (MASLD). A substantial proportion of patients will indeed ultimately develop liver fibrosis and eventually progress towards cirrhosis. Cirrhosis is a clinical dead-end with life-threatening complications (e.g., liver failure, portal hypertension, hepatocellular carcinoma…), which accounts for approximately 1.8% of deaths each year (WHO mortality database). When chronic liver injuries progress, decompensation of the disease (e.g., ascites, jaundice, variceal bleeding or hepatic encephalopathy) may occur, resulting in a dramatic decrease in the overall survival rate. Currently, the clinical predictors of the risk of decompensation have a limited impact on the patient's management and we are unable to accurately characterise the liver parenchyma and monitor its changes or pejorative evolution on imaging alone (e.g., using CT, MRI, ultrasound). The characterisation of chronic liver disease relies on invasive methods such as liver biopsy, to assess fibrosis, steatosis and "activity" (i.e., inflammation) of chronic liver diseases, and trans-jugular catheterization for portal hypertension (measure of the hepatic venous pressure gradient). Such invasive and expensive gold standards are inappropriate for screening and sequential monitoring for obvious reasons. Additionally, liver biopsy is also prone to risks of under-sampling and/or inter-reader variability and does not allow risk stratification of disease progression including hepatocarcinogenesis for instance. All this leads to a necessary and ongoing transition towards non-invasive assessment of chronic liver disease progression and prognostication.

Hepatocellular carcinoma, a life-threatening condition, arises in more than 85% of cases in advanced fibrosis. Over the past decades, the incidence rate of liver cancers has been increasing and the severity of this challenge is amplified by projections that anticipate a 55% increase in new cases of liver cancer by 2040, which would result in 1.3 million deaths worldwide – a 56% increase compared to 2020[1]. The alarming prevalence of hepatocellular carcinoma, combined with a growing understanding of the pathological precursor mechanisms, makes this disease a public health priority for screening programs. Furthermore, there is a significant impact

of early diagnosis on patient prognosis: median survival exceeds five years when hepatocellular carcinoma is diagnosed at an early stage (eligible to curative-intent treatment including liver transplantation, surgical resection, and percutaneous ablation), whereas in advanced hepatocellular carcinoma, survival drops to 3 months. Early-stage hepatocellular carcinoma is defined by the recently updated Barcelona Clinic Liver Cancer (BCLC) classification[2], a guide for management of hepatocellular carcinoma. Incorporating liver function and performance status, the BCLC classification defines 5 stages based on imaging: very early stage (0) with a single ≤ 2 cm nodule, early stage (A) with a single nodule regardless of the size or ≤ 3 nodules each ≤ 3 cm, intermediate stage (B) with multinodular hepatocellular carcinoma, advanced stage (C) with portal invasion and/or extrahepatic spread, and terminal stage (D) corresponding to end-stage liver function or poor performance status regardless of tumour stage. The BCLC 0 and A stages mostly encompass the historical Milan criteria defining the eligibility for liver transplantation by (1) one nodule ≤ 5 cm, (2) ≤ 3 nodules each ≤ 3 cm, (3) no macrovascular invasion, and (4) no extrahepatic spread[3].

In this context, healthcare systems in Europe and North America have included patients with advanced chronic liver disease in screening programs with biannual ultrasound. Screening programs rely on the cost-effectiveness ratio at the collective scale of the population which is determined by the incidence of hepatocellular carcinoma, the cost of the screening program and its benefits including the percentage of patients receiving a curative treatment and the overall survival in at-risk patients. It has been determined that this approach is justified in a given population if the annual incidence of developing hepatocellular carcinoma reaches at least 1.5%[4], with an incremental cost-effective ratio (ICER) within the threshold of willingness-to-pay for a surveillance test, usually accepted as $50,000/quality-adjusted life years[5]. However, the incidence of hepatocellular carcinoma is heterogeneous at the individual level and the epidemiology of chronic liver disease is changing with the eradication of hepatitis C virus, the better control of hepatitis B virus and the rise of MASLD[6]. Although the overall incidence of hepatocellular carcinoma has been currently estimated ranging from 1.5 to 2.5%[7-9], the

proportion of patients with a higher risk of hepatocellular carcinoma remains unknown. Furthermore, ultrasound has significant shortcomings in reproducibility and sensitivity, particularly for detecting early-stage hepatocellular carcinoma with a reported sensitivity of 47% (meta-analysis including 13367 patients[10]) and even more for hepatocellular carcinoma < 2 cm, where sensitivity drops to 22%[11]. Indeed, 80% of patients are diagnosed with advanced hepatocellular carcinoma and only 20% of patients diagnosed with hepatocellular carcinoma are eligible for a first-line curative-intent treatment. This is why the median survival for all stages of hepatocellular carcinoma is only 12 months[12]. Reasons are multiple including heterogeneity of the liver parenchyma in cirrhosis, increased echogenicity with ultrasound attenuation in case of steatosis and obesity, etc.[13-15]. Furthermore, the time interval for screening and its modalities are not personalised. If studies exist on the time interval of screening, they have reported negative results. For instance, Trinchet et al. did not show any survival benefit in performing this screening every 3 months in the overall screening population[16]. In addition, this lack of personalisation results in poor patient compliance.

Thus, surveillance by MRI has been proposed to improve screening as it significantly outperforms ultrasound with a detection rate of 5 times that of ultrasound for very early-stage hepatocellular carcinoma[17]. Considering the higher cost and lower availability of MRI, abbreviated MRI (aMRI) without intravenous contrast (state-of-the-art MRI with fewer acquisition series) was recently introduced as it offers a considerable time-saving advantage, limited to 4.5 to 6 minutes, compared to the conventional MRI protocol, which takes 25 to 40 minutes[18-22]. Various combinations of MRI sequences (without contrast injection, diffusion-weighted imaging (DWI), dynamic sequence with contrast or hepatobiliary phase) have been proposed and all offer higher performance than ultrasound while minimising acquisition time compared to a conventional MRI protocol (30 minutes). The reported sensitivity and specificity of the different protocols (NC-aMRI, DCE-MRI and HB-MRI) ranged from 84.6 to 96% and from 81.6 to 100%[18,21-26]. When results were stratified according to lesion size, the diagnostic sensitivity decreased but remained acceptable for very early-stage hepatocellular carcinoma (< 2 cm) with pooled

sensitivity ranging from 69 to 77.1%, However, although the diagnostic performance of aMRI is superior to that of ultrasound, MRI is an expensive and not easily accessible examination.

Recent analyses of prospective European cohorts including a model-based evaluation of very early-stage hepatocellular carcinoma detection confirmed that MRI surveillance is cost-effective for a baseline yearly incidence of 3% in patients with cirrhosis without active viral replication[17]. Therefore, screening with aMRI can only be considered for a sub-population with a very high risk of hepatocarcinogenesis, which would be selected from the population currently undergoing standard ultrasound screening. Identifying this subset of high-risk patients is crucial as this strategy would detect 5 times more very early-stage hepatocellular carcinoma than ultrasound, with an ICER below 30,000€/life-years gained[27]. Refining and personalising costly hepatocellular carcinoma screening programs based on the individual risk of hepatocellular carcinoma is a timely challenge to provide better care and fairly allocate limited medical resources.

Defining such a population involves developing tools for stratifying the risk of hepatocarcinogenesis[28]. Preliminary models have been developed, either aetiology-specific[29-31] or multi-aetiology[17], incorporating clinical parameters (e.g., age, sex, body mass index, or diabetes) and biological parameters (e.g., GGT, AST/ALT, platelets, or albumin)[30,32,33]. These models demonstrated good discriminative performances and have the advantages of being easy to use and inexpensive. For instance, Nahon et al. developed a multi-aetiology score based on age, sex, platelet count, total bilirubin, GGT, and α-foetoprotein (AFP) that achieved a Harrell's c-index up to 0.76 to identify patients with an annual risk of hepatocellular carcinoma over 3% after 3 years of follow-up[17]. Another example is the aMAP score, developed in chronic hepatitis (age, sex, albumin, bilirubin and platelet count), that also achieved excellent discrimination (Harrell's c-index up to 0.87) and calibration in assessing the 5-year hepatocellular carcinoma risk (up to 19.9% in patients predicted at high-risk)[34]. In addition to these clinical and biochemical parameters, liver stiffness has been shown to be an independent factor associated with hepatocarcinogenesis. Alonso Lopez et al. developed a hepatocellular carcinoma risk model

based on liver stiffness measurements and albumin, achieving a Harrell's c-index of 0.78[35]. Serum protein-based and genetic approaches have also been investigated. An 8-protein serum-based prognostic liver secretome signature (PLSec) recapitulated transcriptome-based hepatic hepatocellular carcinoma risk status with an adjusted hazard ratio of 2.35 and its association with AFP outperformed AFP alone[36-38]. Finally, a 7 single nucleotide polymorphisms genetic risk score, including 6 single nucleotide polymorphisms (PNPLA3, TM6SF2, HSD17B13, APOE, and MBOAT7) affecting lipid turnover and 1 variant involved in the Wnt–b-catenin pathway (WNT3A-WNT9A rs708113), assessed in patients with alcohol-related and/or HCV-cured cirrhosis, achieved good discrimination with a Harrell's c-index up to 0.64 at 5 years but failed to outperform simpler and cheaper clinical risk score such as aMAP score (Harrell's c-index of 0.77) and only showed fair clinical net benefit when associated with the latter score[39].

An alternative approach is to develop prediction models based on the direct analysis of the liver parenchyma. Indeed, these abovementioned models do not consider the analysis of the liver's micro and macrostructure, which reflects the pathophysiological mechanisms responsible for hepatocarcinogenesis. Information regarding the liver morphology is already available by imaging. For example, applying a subjective qualitative ultrasound analysis by the radiologist allows the identification of a population at 20 times greater risk of developing hepatocellular carcinoma when its structure appears heterogeneous and macronodular[40-42]. Despite this data, hepatic morphological and architectural analysis on imaging has never been used to stratify the risk of hepatocarcinogenesis in clinical practice or personalised screening. These models could prove to be complementary to clinical, biochemical, or genetic models.

Quantitative image analysis, which can be achieved with an artificial intelligence (AI) approach, could provide an accurate and reproducible characterisation of liver micro and macrostructure, leading to stratification of the risk of hepatocarcinogenesis, providing aid to the detection of early hepatocellular carcinoma and thus personalising hepatocellular carcinoma screening. Indeed, machine learning can achieve tasks of classification, prediction, segmentation, detection, or image optimization (e.g., faster image acquisition, increased signal-

to-noise ratio, etc.). AI-based imaging models, or the machine learning process, seek to identify and combine new imaging biomarkers, inaccessible to the human eye, in a mathematical model[43]. It aims to provide predictive and/or prognostic information about patients and their pathologies, based on sophisticated statistical analysis[44]. In Dana et al (refer to the section *Additional publications related to the thesis*), we provided a precise overview of quantitative imaging techniques of diffuse liver diseases, together with an explanation of the different concepts of Artificial Intelligence, with short and long-term potential clinical applications for risk stratification and early diagnosis[45].

Objectives

Despite the remarkable rise of quantitative imaging biomarkers for the prediction of pathological features (liver elastography, ultrasound-guided attenuation parameter, or MRI Proton Density Fat Fraction), some decisive clinical needs remain unmet. The assessment of the short- and long-term risk of progression of chronic liver disease towards a pejorative outcome (e.g., liver failure, portal hypertension decompensation or hepatocellular carcinoma) still requires the development of reliable non-invasive tools. This absence can be explained by the difficulty of implementing studies that would need to be exhaustive and prospective over a long period to collect a large number of pejorative events. If fibrosis and steatosis appear as decisive markers for the characterisation of chronic liver disease, they fail to accurately predict the progression of early-stage chronic liver disease to cirrhosis-related complications, such as hepatocellular carcinoma. Furthermore, current hepatocellular carcinoma screening programs based on biannual ultrasound are suboptimal. Refining risk stratification and characterisation of progressive disease would majorly impact screening, monitoring and therapeutic management.

To address the current challenges, this thesis, based on a multidisciplinary approach between hepatology, radiology, and computer science, intends to:

- **Objective 1**: Risk stratify hepatocarcinogenesis in high-risk patients with a deep learning approach using ultrasound and MRI modalities.
- **Objective 2**: Improve detection of early-stage hepatocellular carcinoma in high-risk patients with a deep learning approach using ultrasound and MRI modalities.
- **Objective 3**: Investigate innovative techniques to characterise chronic liver disease.

Risk stratification of hepatocarcinogenesis in high-risk patients

RISK STRATIFICATION OF HEPATOCARCINOGENESIS IN HIGH-RISK PATIENTS WITH A DEEP LEARNING APPROACH USING ULTRASOUND AND MRI IMAGES

The high prevalence of hepatocellular carcinoma, the identification and understanding of risk factors and precursor pathological processes, make hepatocellular carcinoma an optimal disease for screening programs. Another reason is the marked improvement in prognosis when diagnosed early (median survival > 5 years for early hepatocellular carcinoma – BCLC 0 or A - versus 2.5 years, and 3 months for more advanced hepatocellular carcinoma such as in BLCL B and D, respectively), allowing curative treatment (liver resection, percutaneous ablation and liver transplantation)[2,46,47].

As explained in the *Introduction*, biannual ultrasound screening programs are suboptimal with significant shortcomings in sensitivity. To overcome the weaknesses of ultrasound, the use of aMRI has been proposed to improve hepatocellular carcinoma screening as it can detect 5 times more very early-stage hepatocellular carcinomas than ultrasound, with an ICER below 30,000€/life-years gained[27]. Different protocols have been proposed with specific advantages, challenges and limitations: non-contrast (NC; T2-WI with fat suppression and/or DWI and/or T1-WI in/out), dynamic contrast-enhanced (DCE) and hepatobiliary phase (HB)[27]. Abbreviated non-contrast MRI has the advantages of the absence of contrast agent injection, simpler workflow, limited cost and the possibility to repeat poor quality acquisitions[27]. In addition, the inter-reader agreement could be lower with a non-contrast protocol[24]. The challenges and limitations of DCE and HB-MRI are multiple: detection of inconclusive enhancing observations (need for recall examinations), injection of contrast, complex workflow with the need for intravenous access, and higher cost. On the other hand, the reported sensitivity and specificity of the different protocols (NC-aMRI, DCE-MRI and HB-MRI) ranged from 84.6 to 96% and from 81.6 to 100%[18,21-26]. Considering the similar screening performances of the different protocols and their pros and cons, we believe that NC-aMRI is a promising protocol for screening programs. In the coming years, hepatocellular carcinoma screening programs will most likely rely on screening ultrasound

and NC-aMRI for the most at-risk patients (annual incidence > 3%[17]), which motivates the urge to stratify the risk of hepatocarcinogenesis and personalise screening strategies.

Preliminary models have been developed, either aetiology-specific[29-31] or multi-aetiology[17], incorporating clinical parameters (e.g., age, sex, body mass index, or diabetes) and biological parameters (e.g., GGT, AST/ALT, platelets, or albumin)[30,32,33], serum proteins[36-38] or single nucleotide polymorphisms[39]. These models demonstrated good discriminative performances and have the advantages of being easy to use and inexpensive. However, these models do not take into consideration the structural analysis of the liver parenchyma, which reflects the pathophysiological mechanisms responsible for hepatocarcinogenesis. In the 1990s, ultrasound studies examined the incidence of hepatocellular carcinoma according to the liver echostructure[40-42]. Results showed that a nodular heterogeneous echostructure resulted in an adjusted rate ratio estimate of up to 20. However, these findings have not led to a personalisation of the screening strategy. If ultrasound is the modality of choice in clinical practice (availability, less expensive,...) and perfectly suited for a hepatocellular carcinoma-risk stratification Deep Learning model, it may not be contributory in all patients due to different reasons: ultrasound attenuation in the case of steatosis or significant subcutaneous fat in obese patients, etc.[13-15]. Therefore, developing equivalent models on ultrasound and MRI is advisable to address the applicability limitations of ultrasound. In addition, if preliminary risk stratification based on ultrasound echotexture pattern has not led to a personalised screening strategy, it may also be due to significant inter-operator and reader variability. MRI can offer standardised image acquisition, increased reader consistency and superior liver parenchyma characterisation.

Therefore, we hypothesised that non-tumour cirrhotic liver parenchyma is rich in structural information reflecting the severity of liver disease, its carcinogenic risk as well as the process of hepatocarcinogenesis. Its analysis will allow to define a very high-risk population, especially in the context of HCV eradication and HBV control.

The primary objective was to design an imaging-based risk stratification deep learning model for hepatocarcinogenesis on ultrasound to identify a population at very high risk of

developing hepatocellular carcinoma. Developing such a model on non-contrast aMRI is also of importance, for instance to compensate for potential failure of ultrasound in specific patient subgroups, and another study was initiated.

RISK STRATIFICATION OF HEPATOCARCINOGENESIS IN HIGH-RISK PATIENTS WITH A DEEP LEARNING APPROACH USING ULTRASOUND (STARHE CLINICAL TRIAL)

Objective

The primary objective was to design an imaging-based risk stratification deep learning model for hepatocarcinogenesis on ultrasound cine clips of the non-tumoral parenchyma, to identify a population at very high risk of developing hepatocellular carcinoma.

Material and methods

Ethics

This prospective project was approved by the Research Ethics Board (Comité de protection des personnes Sud-Est VI 21.03054.001701-MS03; ClinicalTrial NCT04802954) and followed the ethical principles of the Declaration of Helsinki. All patients provided written informed consent.

Study design

This prospective multicentric study was conducted in 6 centres (Hôpital Beaujon, Hôpital Avicennes, CHU Angers, CHU Montpellier, Hôpital Croix Rousse, Hôpital Edouard Herriot). **Table 1** summarises the study design.

Actions	**Hepatology visit** D-30 to D-7	**Inclusion** D0	**Liver ultrasound** D0	**Follow-up visit** 1 year
Patient information	X			
Collection of informed consent		X		
Verification of inclusion criteria	X	X		
Collection of demographics information, medical history, biochemical workup and liver pathologic report (if available)		X		

Liver ultrasound with elastography			X	
Ultrasound video and elastography acquisition			X	
Collection of the ultrasound report			X	X[(B)]
Adverse Event monitoring		X	X	X[(B)]
End of study			X[(A)]	X[(B)]

Table 1 - Study design. [(A)] for patients in low-risk group (A), without hepatocellular carcinoma on ultrasound on day 0. [(B)] for patients in high-risk group (B), with hepatocellular carcinoma on ultrasound on day 0

Population

The inclusion criteria were as follows:

- Patients over 18 years of age.
- Enrolled in a screening programme for at least 6 months, defined by Child-Pugh A or B histologically proven F3/F4 liver or cirrhosis unequivocally suggested by non-invasive tests of non-viral or controlled/healed B/C viral cause (HBV PCR negative under anti-viral B treatment for more than 12 weeks / HCV PCR negative at least 12 weeks after stopping anti-viral C treatment)
- Patients referred by hepatologist for ultrasound screening.
- No history of treated hepatocellular carcinoma.

Non-inclusion criteria were as follows:

- History of hepatocellular carcinoma.
- Non-cirrhotic viral hepatitis B or uncontrolled HBV cirrhosis (HBV) or uncured HCV cirrhosis (< 3 months).
- Patient under judicial protection, guardianship or curatorship.
- Patient in a situation of social fragility.

- Patients who are subject to a legal protection measure or who are unable to express their consent.

The exclusion criteria were as follows:

- Imaging data (ultrasound videos) not recorded
- Patients in the low-risk group lost on follow-up (refer to reference standard).

Index test

We aimed to develop a deep learning classification model based on ultrasound cine clips of the non-tumoral liver to stratify the risk of hepatocarcinogenesis.

Reference standard

Two groups of patients were defined:

- **High-risk group**: Patients with early-stage hepatocellular carcinoma. To achieve a balanced distribution of patients between both groups, we used a method of reinforcement of pathological cases. All patients from an ultrasound screening programme who had been diagnosed with a BCLC 0 or A hepatocellular carcinoma as per the reference diagnostic standards (radiological – LIRADS v2018 or EASL – or pathologic) were included.
- **Low-risk group**: Patients without hepatocellular carcinoma. These patients were included in the framework of the usual screening. A 1-year interval ultrasound, or dedicated liver CT or MRI if clinically warranted, was performed to confirm the absence of new lesions in the year following the inclusion. The proportion of new hepatocellular carcinoma was expected not to exceed 3-5%. In the case of new hepatocellular carcinoma, these patients were reassigned to the high-risk group.

Outcome

The main outcome of the study was the diagnostic (classification) performances of the AI model for risk stratification of hepatocarcinogenesis.

Collected data

All data were collected at inclusion:

- Clinical: demographics (age and sex), Body Mass Index, liver disease history (aetiology, viral hepatitis status, alcohol consumption), medical history (diabetes, HIV co-infection).
- Biology: liver disease scores (FASTRAK – a multi-aetiology score based on age, sex, platelet count, total bilirubin, GGT, and α-foetoprotein that achieved a Harrell's c-index up to 0.76 to identify patients with an annual hepatocellular carcinoma risk over 3% after 3 years of follow-up – MELD, Child-Pugh, FIB-4), tumour markers (α-foetoprotein), liver function tests (bilirubin, AST, ALT, GGT), haemostasis (platelets, INR, PT), albumin.
- Imaging: ultrasound cine clips (non-tumoral liver parenchyma and hepatocellular carcinoma using B mode ultrasound), liver stiffness using 2D US-guided Shear Wave elastography (kPa).
- Pathology: pathology report of the non-tumoral liver parenchyma: steatosis, activity/inflammation, fibrosis), pathology report of hepatocellular carcinoma if available (focality, prognostic grade, vascular invasion, perineural invasion, capsule).

Ultrasound examination

Ultrasound examinations were performed using two models from two manufacturers: Aplio (Canon Medical Systems, Otawara, Japan) and Aixplorer/MACH 30 (Hologic, Marlborough, Massachusetts, USA; former SuperSonic Imagine, Aix-en-Provence, France). Conventional liver ultrasound was performed using B-mode and colour Doppler. Data acquisition was standardised according to a mandatory protocol implemented in each ultrasound scanner using a low-

frequency abdominal convex transducer (C6-1X for Hologic SuperSonic Image and i8Cx1 for Canon Medical Systems): one B-mode cine clip of 10 seconds in free breathing recorded in an intercostal section of the non-tumoral right liver without passing through the hepatocellular carcinoma (entitled non-tumoral liver); one B-mode cine clip of 10 seconds in free breathing of the hepatocellular carcinoma, any approach allowed (entitled B-mode tumour; refer to Objective 2 of the thesis). Default abdominal preset was used. Depth was initially set at 12 cm with a focal at 7.68 cm on Canon Medical Systems and 7-8 cm on Hologic SuperSonic Image, but ultimately left to the operator's discretion. The cine clips were exported in DICOM format.

2D ShearWave elastography was performed with low-frequency abdominal convex probes (C6-1X for Hologic SuperSonic Image and i8Cx1 for Canon Medical Systems). Liver stiffness measurements were acquired in an intercostal section in the right liver lobe using a fixed-size stiffness colour mapping. 3 reliable measurements were performed in the right liver according to the reference quality standards.

AI methodology

We aimed to develop a deep learning classification model based on ultrasound cine clips of the non-tumoral liver to stratify the risk of hepatocarcinogenesis.

- **Database**: The training/validation and testing sets were stratified according to potential confounders: aetiology of liver disease, FASTRAK score (binary cut-off of 9 points), ultrasound manufacturer, hepatocellular carcinoma size (binary cut-off of 2 cm) and echogenicity (isoechoic or not – refer to Objective 2 of the thesis). To compensate for the imbalance between the two groups, we applied oversampling with data augmentation and weighted loss to penalise model errors for data from the minority group.

- **Labelling of the database**: ultrasound cine clips were labelled by a radiologist subspecialised in liver imaging.
- **Pre-processing of ultrasound images**. ultrasound images were embedded in video layouts influenced by factors such as ultrasound machine brand and display settings. To standardise these images and minimise bias, we developed an automated method to extract the echo region-of-interest from the layout. Our method detected pixels with minimal intensity changes across video timestamps, classifying them as background. This classification allowed to create a binary mask. We then refined this mask using morphological operations to remove artifacts and cropped around the ultrasound region-of-interest.
- **Stratification of the risk of hepatocarcinogenesis based on the non-tumoral liver parenchyma**. We framed the task as a video classification challenge to identify patients at high risk of developing a hepatocellular carcinoma based on a short video clip of non-tumoral liver parenchyma. The video was divided into 10 clips with 16-frame sub-clips sampled from each. We designed a voting system between the predictions of the model for each sub-clip to determine the final classification of the video. Our implementation was based on the MMAction2 library. The training/validation set was split into 5 folds, and we performed cross-validation for model selection and hyperparameters tuning (**Table 2**)[48]. We selected three state-of-the-art algorithms with different sets of parameters for this purpose: MViT, C3D, and I3D. Each model was pretrained on the Kinetics-400 dataset to facilitate transfer learning. First, for each session of CV, all 3 models were trained with different sets of parameters, and their performances were measured on the validation folds. We selected the model or hyperparameter set with the best average performance across all folds as the final model. This final model was retrained using all the training/validation data.

	Model and hyperparameters	Final model
Batch sizes	2, 4, 8, 16, 32	4
Learning rate	0.000005, 0.00001, 0.000015, 0.00002, 0.000025, 0.00003, 0.0000375, 0.00005, 0.0001, 0.0016	45 epochs with learning rate of 0.00002 (divided by 10 at 20 and 40 epochs)
Optimizer	SGD, AdamW	SGD
Input/output size	128x128, 256x256	128x128
Transfer learning	All except prediction head	All except prediction head
Models	MViT, C3D, I3D	C3D

Table 2 - Model and hyperparameters selection using a cross-validation approach

- **Independent testing and sample size calculation**: To ensure the robustness and generalisability of our AI models, we planned to test them in an independent dataset as mentioned above. Approaching this question from a statistical perspective, we planned to include 50 patients in the external independent testing dataset (significance level α of 5%, statistical power of 80%, control-to-case ratio of 1:1, annual incidence of hepatocellular carcinoma of 3% in the low-risk group and a relative risk of 11). This estimate is based on previous studies showing that a macronodular heterogeneous echostructure on ultrasound is associated with an adjusted rate ratio up to 20[40-42]. Therefore, considering the requirements of AI model developments, we intended to include 400 patients to allow an approximate balance of 80%-20% between the training/validation and testing sets and to compensate with excluded patients (expected rate of 10%).

Statistical analysis

Estimates of performance metrics of the classification model were computed for sensitivity, specificity, balanced accuracy, positive and negative predictive values, area under the receiver operating characteristics curve, c-index and odds ratio. Calibration curves were computed.

Results

Population

This study enrolled 403 patients between September 2021 and December 2023 (**Figure 1**). A total of 81 patients were excluded: 44 did not match the inclusion criteria, 19 were lost on follow-up, and 18 had no ultrasound cine clips recorded. In the training/validation dataset, 272 patients were analysed including 145 patients in the low-risk group and 127 in the high-risk group. In the independent testing dataset, 50 patients were analysed including 25 patients in the low-risk group and 25 in the high-risk group.

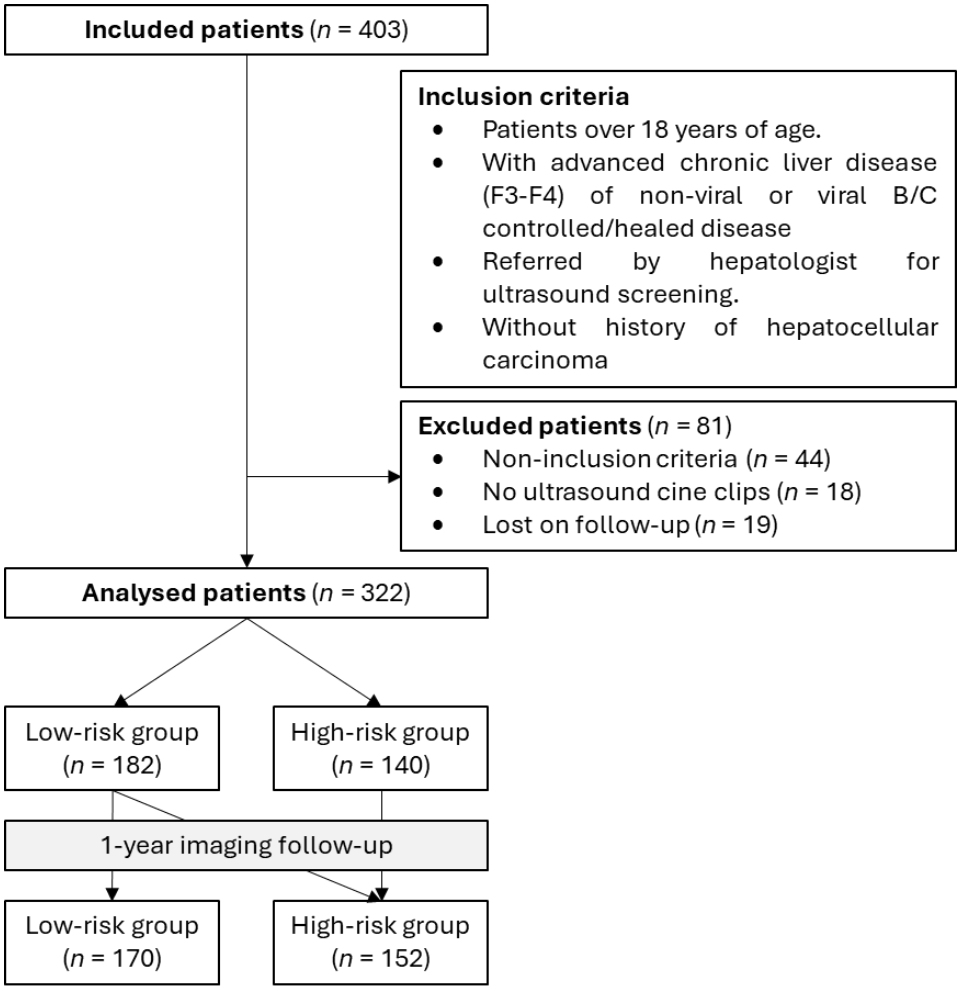


Figure 1 – Flow chart of the study

The demographics description of the dataset is reported in **Table 3**. There was a majority of male patients in the low-risk group (120/170 - 71%) and the high-risk group (137/152 – 81%) with a median age of 63 and 69, respectively. The distribution of aetiologies of chronic liver disease was overall adequately balanced between the low-risk and high-risk groups with a majority of patients with alcohol-related liver disease: 76/170 (45%) in the low-risk group and 63/152 (41%) in the high-risk group. The FASTRAK score was, as expected, higher in the high-risk (median of 11 points versus 7 points). Finally, most patients in the high-risk group were classified as BCLC A (99/140 – 71%).

	Training/validation		Testing		Total	
	Low-risk (n = 145)	High-risk (n = 127)	Low-risk (n = 25)	High-risk (n = 25)	Low-risk (n = 170)	High-risk (n = 152)
Centres						
1	28 (19%)	42 (33%)	3 (12%)	6 (24%)	31 (18%)	48 (32%)
2	31 (21%)	17 (52%)	5 (20%)	6 (24%)	36 (21%)	23 (15%)
3	27 (19%)	21 (17%)	2 (8%)	3 (12%)	29 (17%)	24 (16%)
4	1 (1%)	36 (28%)	0	7 (28%)	1 (1%)	43 (28%)
5	50 (34%)	8 (6%)	10 (40%)	2 (8%)	60 (35%)	10 (7%)
6	8 (6%)	3 (2%)	5 (20%)	1 (4%)	13 (8%)	4 (3%)
Ultrasound manufacturer						
Canon	51 (35%)	93 (73%)	9 (36%)	18 (72%)	60 (35%)	111 (73%)
Supersonic (Hologic)	94 (65%)	34 (27%)	16 (64%)	7 (28%)	110 (65%)	41 (27%)
Age	63 [57-68]	69 [63-75]	61 [56-70]	68 [61-74]	63 [56-69]	69 [62-75]
Sex						
Male	102 (70%)	113 (89%)	18 (72%)	24 (96%)	120 (71%)	137 (81%)
Female	43 (30%)	14 (11%)	7 (28%)	1 (4%)	50 (29%)	15 (19%)
Chronic Liver Disease						
Aetiology of liver disease						
ALD	63 (44%)	47 (37%)	11 (44%)	8 (32%)	74 (44%)	55 (36%)
MASLD	27 (19%)	15 (16%)	5 (20%)	6 (24%)	32 (19%)	21 (14%)
MetALD	23 (16%)	31 (12%)	3 (12%)	3 (12%)	26 (15%)	34 (22%)
HBV	6 (4%)	3 (2%)	1 (4%)	1 (4%)	7 (4%)	4 (3%)
HCV	15 (10%)	14 (11%)	3 (12%)	2 (8%)	18 (11%)	16 (11%)
ALD + HBV	0 (0%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)
ALD + HCV	3 (2%)	4 (3%)	1 (4%)	5 (20%)	4 (2%)	9 (6%)
MASLD + HBV	1 (1%)	1 (1%)	0 (0%)	0 (0%)	1 (1%)	1 (1%)
MASLD + HCV	1 (1%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)	0 (1%)
HBV + HCV	0 (0%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)
Other	6 (4%)	10 (8%)	1 (4%)	0 (0%)	7 (4%)	10 (6%)
FASTRAK score	7 [4-11]	10 [9-13]	8 [5-11]	11 [9-13]	7 [5-11]	10 [9-13]
Child-Pugh						
A5	88 (59%)	70 (55%)	15 (60%)	13 (52%)	103 (61%)	83 (55%)
A6	24 (19%)	27 (21%)	3 (12%)	3 (12%)	27 (16%)	30 (20%)
B7	12 (8%)	10 (8%)	1 (4%)	2 (8%)	13 (8%)	12 (8%)
B8	3 (21%)	1 (1%)	0 (0%)	2 (8%)	3 (2%)	3 (2%)
B9	0 (0%)	2 (2%)	0 (0%)	0 (0%)	0 (0%)	2 (1%)
Missing data	18 (12%)	17 (13%)	6 (24%)	5 (20%)	24 (14%)	22 (14%)
Type 2 diabetes	61 (42%)	52 (41%)	7 (28%)	10 (40%)	68 (40%)	62 (41%)
BMI ≥ 25	104 (72%)	87 (69%)	18 (72%)	19 (76%)	122 (72%)	106 (70%)
Biology						
Alpha-foetoprotein (ng/mL)	4.1 [2.6-5.2]	5.7 [3.0-10.5]	3.2 [2.4-5.5]	5.7 [3.6-7.0]	4.1 [2.6-5.2]	5.7 [3.0-10.5]
GGT (UI/L)	87 [39-161]	131 [64-270]	83 [36-124]	118 [58-267]	87 [39-161]	131 [64-270]
Total Bilirubin (µmol/L)	14 [9-19]	15 [9-23]	14 [11-24]	14 [10-28]	14 [9-19]	15 [9-23]
Platelet (G/L)	158 [107-200]	126 [95-172]	158 [99-190]	115 [82-201]	158 [107-201]	126 [95-172]
INR	1.2 [1.1-1.3]	1.1 [1.1-1.3]	1.1 [1.0-1.2]	1.1 [1.0-1.2]	1.2 [1.1-1.3]	1.1 [1.1-1.3]
Albumin (g/L)	41 [36-44]	40 [36-43]	42 [40-44]	39 [32-43]	41 [36-44]	40 [36-42]
ShearWave Elastography (kPa)						
	14.4 [10.1-21.9]	12.8 [9.9-18.9]	13.1 [10.3-20.9]	11.3 [9.0-16.0]	14.1 [10.1-21.9]	12.5 [9.8-19.0]
Hepatocellular carcinoma (at inclusion)						
Number of nodules						
1	NA	87 (69%)	NA	18 (72%)	NA	105 (75%)
2		22 (17%)		7 (28%)		29 (21%)
3		6 (5%)		0		6 (4%)
Largest nodule size (mm)	NA	25 [20-31]	NA	22 [20-30]	NA	25 [20-31]
Nodule echogenicity	NA		NA		NA	

Homogeneous						
Hypoechoic		33 (37%)		8 (28%)		41 (36%)
Isoechoic		24 (27%)		8 (28%)		32 (28%)
Hyperechoic		19 (21%)		5 (20%)		24 (21%)
Heterogeneous						
Iso and hypoechoic		6 (7%)		2 (8%)		8 (7%)
Iso and hyperechoic		4 (4%)		1 (4%)		5 (4%)
Hypo and hyperechoic		4 (4%)		1 (4%)		5 (4%)
BCLC stage						
0	NA	31 (24%)	NA	10 (40%)	NA	41 (29%)
A		84 (76%)		15 (60%)		99 (71%)
Ultrasound cine clips (B mode)						
Data						
Non-tumour liver parenchyma	145 (100%)	122 (96%)	25 (100%)	25 (100%)	170 (100%)	147 (97%)
Hepatocellular carcinoma	NA	122 (96%)	NA	25 (100%)	NA	147 (97%)

Table 3 - Demographics description of the population. *Notes: ALD = Alcohol-related liver disease; MASLD; Metabolic Dysfunction-Associated Steatotic Liver Disease; MetALD = MASLD and ALD; HBV = Hepatitis B virus; HCV = Hepatitis C virus; AI = auto-immune; Other (auto-immune, primary biliary cholangitis, granulomatosis, hemochromatosis, Wilson disease, iatrogenic,*

Risk stratification of hepatocarcinogenesis

The trained C3D classification model achieved good diagnostic performances in the testing set with an accuracy of 0.72 (95% CI 0.57-0.84), a sensitivity of 0.72 (95% CI 0.51-0.88), a specificity of 0.72 (95% CI 0.51-0.88), a positive predictive value of 0.72 (95% CI 0.57-0.83), a negative predictive value of 0.72 (95% CI 0.57-0.83), and an AUC of 0.75 (95% CI 0.61-0.86). A patient predicted at high-risk by the model had an odds ratio of 6.6 (95% CI 1.9-22.7; p=0.003). The c-index was 0.75. On the other hand, the model achieved moderate calibration (**Figure 2**). **Figure 3** illustrates two examples of a patient predicted at low risk and a patient predicted at high risk. In comparison, the classification model achieved an accuracy of 0.81 and a c-index of 0.92 in the training set.

Considering the study design, we cannot exclude that these results may be underestimated. Indeed, although we intended to mitigate this potential bias with a follow-up at 1 year, a few patients at high risk of hepatocarcinogenesis might have been included in the low-risk group and may have developed an HCC in the next months following the 1-year follow-up.

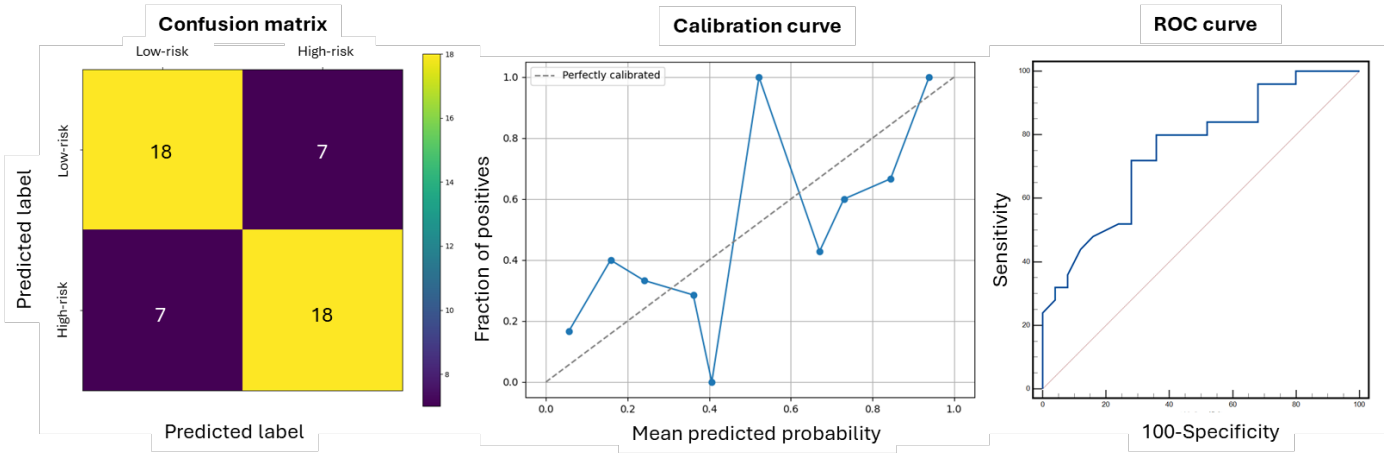


Figure 2 – Confusion matrix, calibration curve and ROC curve of the C3D classification model to stratify the risk of hepatocarcinogenesis in the testing set

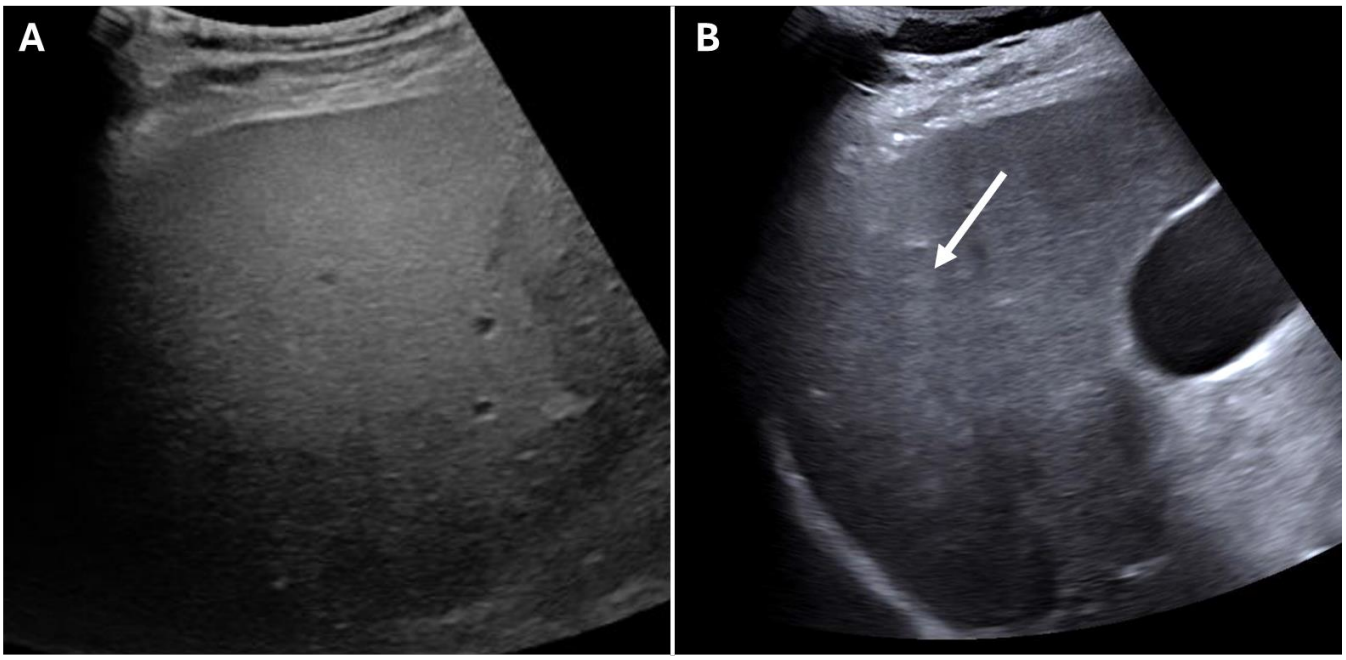


Figure 3 – Increased homogeneous echostructure in a 75 years-old man with alcohol-related cirrhosis correctly predicted at low-risk (A) compared to a macronodular echostructure in a 58 years-old man with alcohol-related cirrhosis and BCLC A hepatocellular carcinoma correctly predicted at high-risk on this view of the non-tumoral liver parenchyma (B).

Conclusion

The developed classification model achieved good diagnostic performances with an odds-ratio of 6.6 and an accuracy of 0.72 to predict patients at high risk of hepatocarcinogenesis based on the analysis of the non-tumoral parenchyma. This new imaging biomarker could help stratify the risk of hepatocarcinogenesis alongside clinical or biochemical biomarkers and allow risk-based personalised screening strategies.

RISK STRATIFICATION OF HEPATOCARCINOGENESIS IN HIGH-RISK PATIENTS WITH A DEEP LEARNING APPROACH USING MRI

Ultrasound will remain the cornerstone of chronic liver disease characterisation, including liver elastography and steatosis quantification, and hepatocellular carcinoma surveillance. It is therefore logical to maintain ultrasound-based tools as a first-line screening strategy alongside blood biomarkers. However, MRI biomarkers can be used to refine the screening strategy. In addition, quality control of ultrasound (LI-RADS visualisation score) should also be taken into consideration. In a retrospective cohort study, about 20% of patients with cirrhosis had moderately to severely limited ultrasound visualisation for hepatocellular carcinoma nodules, particularly those with obesity, ALD or MASLD cirrhosis[49]. This the reason why developing a risk stratification deep learning model on NC-aMRI clinically matters.

Material and methods

Ethics

This retrospective project was approved by the Research Ethics Board (McGill University Health Centre REB F20-113490) and followed the ethical principles of the Declaration of Helsinki. Written consent was waived.

Study design

This retrospective multicentre IRB-approved study (2023-9568) was coordinated at the Research Institute of McGill University Health Centre (MUHC). It included patients from November 2011 to September 2023 at the Royal Victoria Hospital, Montreal General Hospital, and Lachine Hospital.

Population

The inclusion criteria were as follows:

- Patients over 18 years of age.
- Adult patients enrolled in a screening programme for at least 6 months, defined by Child-Pugh A or B histologically proven F3/F4 liver or cirrhosis unequivocally suggested by non-invasive tests, of non-viral or controlled/healed B/C viral cause (HBV PCR negative under anti-viral B treatment for more than 12 weeks / HCV PCR negative at least 12 weeks after stopping anti-viral C treatment).
- Liver MRI with complete characterisation protocol, including post-contrast sequences, to determine reference standard, performed in the screening setting (liver disease or nodule characterisation).
- No history of treated hepatocellular carcinoma
- Composite reference gold standard
    - No or benign (LR-1 or 2) liver observations on baseline MRI (low-risk group) with a 1-year follow-up by dedicated ultrasound, CT, or MRI.
    - LR-3 or LR-4 liver observations on baseline MRI with pathological proof (low or high-risk group) or stability/regression over a two-year follow-up (low-risk group)
    - LR-5 liver observations, i.e. hepatocellular carcinoma (high-risk group)

The exclusion criteria were as follows:

- Poor quality of the imaging data
- No acute hepatic event at the time of the MRI

Index test

We aimed to develop a classification deep learning model based on non-contrast aMRI (T1-weighted in/out-of-phase images, T2-weighted with fat suppression images, diffusion-weighted imaging, using the same acquisition parameters as a state-of-the-art MRI to maintain

the same contrast, spatial resolution, and signal-to-noise ratio) to risk stratify hepatocarcinogenesis.

Reference standard

Two groups of patients were defined:

- *High-risk group: Patients with BCLC 0 or A hepatocellular carcinoma* as per the reference diagnostic standards (radiological – LIRADS v2018[50] or EASL – or pathologic) were included.
- *Low-risk group: Patients without hepatocellular carcinoma* on baseline MRI. A 1-year interval follow-up by ultrasound, dedicated liver CT or MRI, was required to confirm the absence of new lesions in the year following the baseline MRI.

Outcome

The main outcome of the study was the diagnostic performances of the AI models for risk stratification of hepatocarcinogenesis.

Collected data

All data were collected at inclusion:

- Clinical: demographics (age and sex), Body Mass Index, liver disease history (aetiology, viral hepatitis status, alcohol consumption), medical history (diabetes, HIV co-infection).
- Biology: liver disease scores (FASTRAK – a multi-aetiology score based on age, sex, platelet count, total bilirubin, GGT, and α-foetoprotein that achieved a Harrell's c-index up to 0.76 to identify patients with an annual hepatocellular carcinoma risk over 3% after 3 years of follow-up – MELD, Child-Pugh, FIB-4), tumour markers (α-foetoprotein), liver function tests (bilirubin, AST, ALT, GGT), haemostasis (platelets, INR, PT), albumin (g/L).

- Imaging: MRI dicoms.
- Pathology: pathology report of the non-tumoral liver parenchyma: steatosis, activity/inflammation, fibrosis), pathology report of hepatocellular carcinoma if available (focality, prognostic grade, vascular invasion, perineural invasion, capsule).

AI methodology

The database was randomly divided into training, validation and test sets stratified according to potential confounders including aetiology of liver disease, FASTRAK score, MRI manufacturer and model, hepatocellular carcinoma number and size of the largest nodule. To compensate for the imbalance between the two groups, we applied oversampling with data augmentation and weighted loss to penalise model errors for data from the minority group. To build reproducible, high-performance models, we followed best practices in developing and evaluating machine learning models, including independent testing

The AI system was developed in the following stages:

- **Pseudonymization** of data per Canadian and Quebec legislations.
- **Annotation and labelling of databases**. MRIs were annotated and labelled by a radiologist subspecialized in liver imaging: (1) presence of hepatocellular carcinoma; (2) localisation and segmentation of hepatocellular carcinoma (3D Slicer).
- **Liver segmentation**. The liver was automatically segmented on T2-weighted images using a deep learning model provided by Guerbet® from which the tumour volume was cropped by a fellowship-trained radiologist. Based on the annotation and the automated segmentation of the liver, the nontumoral liver parenchyma was segmented and extracted to develop the risk stratification model.
- **Pre-processing of MRI images**. The non-uniformity of the intensities of the 3D MRI imaging volumes will be corrected using the *N4 Bias Field Correction* algorithm[51], then

normalised according to the *z-score*. Any misalignment of images from the different MRI sequences used, most often due to breathing inconsistency, was corrected.

- **Stratification of the risk of hepatocarcinogenesis based on the non-tumoral liver parenchyma**. The development of the classification is under progress and will rely on an advanced approach centred around training a 3D Res-Net on a multiple-input including T2-weighted images with fat suppression, Diffusion-weighted images with ADC map, and T1-weighted images in/out. This neural network will be specifically designed to analyse the non-tumoral liver parenchyma, with the aim of detecting patients at high risk of developing hepatocellular carcinoma. This approach leverages the detailed spatial and structural information provided by 3D imaging, enhancing the model's ability to discern complex patterns that might not be evident in two-dimensional analyses or less sophisticated models. The training/validation set will be split into 5 folds, and we will perform cross-validation for model selection and hyperparameters tuning[48]. Considering the presence of bounding boxes to exclude hepatocellular carcinoma areas from the liver images, different approaches will be investigated to prevent any bias in the training: (1) cropping of random bounding boxes during data augmentation in patients in the low-risk group; (2) patch-based approach; (3) inpainting approach; (4) if none of the first 3 approaches is successful, we will train a 2D CNN model only on liver images without bounding boxes and design a voting system between the predictions of the model for each image to determine the final classification of the MRI.
- **Performance analysis by profile**. Machine learning models can sometimes predict poor performance in specific subgroups of patients with different profiles despite high overall performance in the test dataset. To anticipate these failures, we aim to develop a measure of prediction confidence for the model by forming a second-layer model called conditional accuracy that aims to identify prediction errors specific to certain patients or groups of patients. This approach identifies patient profiles with higher prediction errors than the initial classification model.

- **External independent testing**: To ensure the robustness and generalisability of our AI model, we will first test it in an independent internal dataset. Then, we will test it on two external independent databases, each with different prevalence of hepatocellular carcinoma, as the AI model performance may be affected by the prevalence of the disease:
    - CHUM [Centre Hospitalier de l'Université de Montréal]: retrospective cohort with identical inclusion criteria and similar high and low-risk patients' distribution than the training dataset. Approaching this question from a statistical perspective[52], we plan to include at least 58 patients balanced in the two groups in this external independent testing dataset (significance level α of 5%, statistical power of 80%, control-to-case ratio of 1:1, annual incidence of hepatocellular carcinoma of 3% in the low-risk group and relative risk of 10). This estimate is based on previous studies showing that a macronodular heterogeneous echostructure on ultrasound is associated with an adjusted rate ratio up to 20[40-42].
    - FASTRAK [*FAST-IRM for hepatocellular carcinoma suRveillance in pAtients with high risK of liver cancer* (NCT05095714)[53]]: prospective multicentre cohort including 950 patients with advanced chronic liver disease of non-viral or viral B/C controlled/healed disease (HBV PCR negative on anti-viral B therapy for more than 12 weeks / HCV PCR negative at least 12 weeks after stopping anti-viral C therapy), absence of hepatocellular carcinoma on imaging less than 3 months old, or history of treated hepatocellular carcinoma, and estimated annual risk of hepatocellular carcinoma > 3% based on a clinical-biochemical score. This study will be completed in December 2027.

Statistical analysis

Estimates of performance metrics will be computed for sensitivity, specificity, balanced accuracy, positive and negative predictive values, area under the receiver operating characteristics curve, odds ratio and c-index. Calibration curves will be computed.

Results

Population

This study enrolled 368 patients between November 2011 and September 2023 (**Figure 4**). A total of 35 patients were excluded: 25 had extensive artifacts on the MRI and 10 had the MRI at the time of an acute liver event (e.g., acute portal vein thrombosis). An additional total of 41 patients were excluded from the analysis of hepatocellular carcinoma risk stratification: 2 had non-controlled HBV, 38 had non-cured HCV and 1 had no information regarding the aetiology of the liver disease.

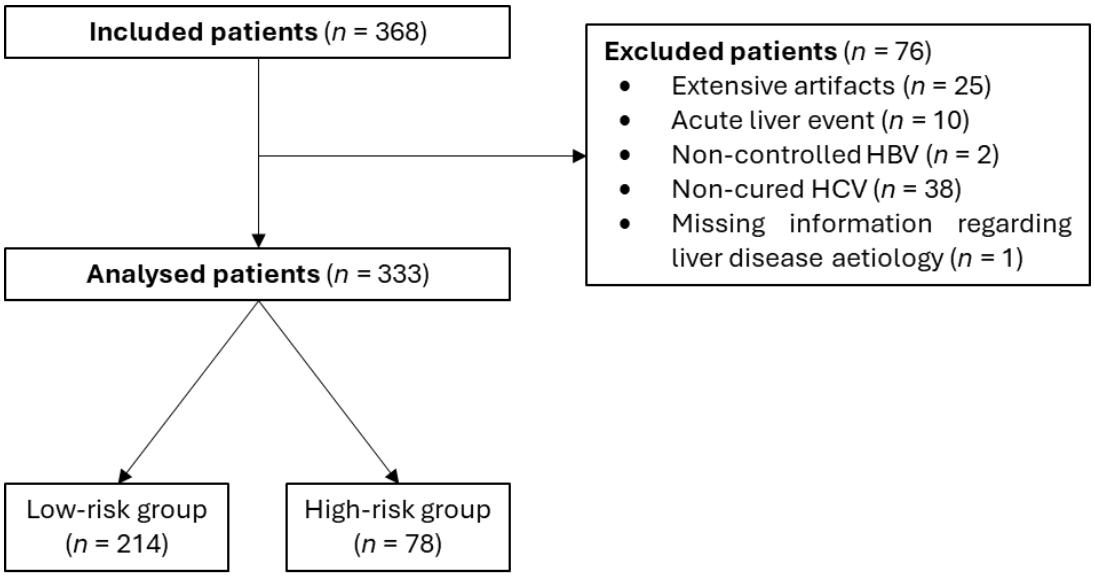


Figure 4 – Flow chart of the study

The demographics description of the dataset is reported in **Table 4**. There was a majority of male patients in the training/validation (209/263 – 79%) and testing (48/59 – 81%) datasets with a median age comprised between 62 and 69. The distribution of aetiologies of chronic liver disease was overall adequately balanced between the low-risk and high-risk groups in both training/validation and testing datasets. Although there were more patients with viral cured/controlled HBV or HCV in the testing dataset than in the training dataset, patients with ALD (12/59 – 20%), MASLD (12/59 – 20%), and MetALD (6/59 – 10%) were adequately represented in the testing datasets.

	Low-risk (*n* =214)	High-risk (*n* = 78)
Centres		
1 (Royal Victoria Hospital)	122 (57%)	49 (63%)
2 (Montreal General Hospital)	76 (36%)	17 (22%)
3 (Lachine Hospital)	16 (7%)	12 (15%)
MRI manufacturer		
General Electrics Optima MR450w (1.5T)	59 (28%)	24 (31%)
General Electrics SIGNA Excite (1.5T)	104 (49%)	18 (23%)
General Electrics SIGNA Artist (1.5T)	31 (14%)	18 (23%)
Siemens AERA (1.5T)	16 (7%)	12 (15%)
Siemens SKYRA (3.0T)	4 (2%)	6 (8%)
Age	62 [52-70]	65 [59-74]
Sex		
Male	127 (59%)	60 (77%)
Female	87 (41%)	18 (23%)
Chronic Liver Disease		
Aetiology of liver disease		
ALD	30 (14%)	12 (15%)
MASLD	53 (25%)	24 (31%)
MetALD	9 (4%)	6 (8%)
HBV	26 (12%)	9 (12%)
HCV	15 (7%)	21 (27%)
ALD + HBV	4 (2%)	1 (1%)
ALD + HCV	2 (1%)	1 (1%)
MASLD + HBV	2 (1%)	1 (1%)
HBV + HCV	2 (1%)	0
PSC	18 (8%)	0
Auto-immune and overlap syndrome	16 (7%)	3 (4%)
Indeterminate	10 (5%)	0
Other	27 (13%)	0
FASTRAK score	7 [6-10]	10 [8-12]
Child-Pugh		
A5	131 (61%)	36 (46%)
A6	32 (15%)	18 (23%)
B7	25 (12%)	13 (17%)
B8	8 (4%)	8 (10%)
B9	2 (1%)	1 (1%)
C10	0	1 (1%)
Missing data	16 (7%)	1 (1%)
Type 2 diabetes	70 (33%)	41 (53%)
Biology		
Alpha-foetoprotein (ng/mL)	4.0 [2.7-6.0]	5.6 [3.1-8.6]
GGT (UI/L)	65.5 [33.3-140.8]	84.5 [46.8-149.3]
Total Bilirubin (µmol/L)	16.8 [12.2-24.3]	18.8 [11.4-31.0]
Platelet (G/L)	134 [91-177]	108 [77-150]
INR	1.1 [1.0-1.2]	1.1 [1.0-1.3]
Albumin (g/L)	39 [36-42]	38 [33-43]
Hepatocellular carcinoma (at inclusion)		
Number of nodules		
1	NA	58 (74%)
2		14 (18%)
3		6 (8%)
Size (mm)	NA	23.0 [15.0-28.0]
BCLC stage		
0	NA	31 (40%)
A		47 (60%)

Table 4 - Demographics

Risk stratification of hepatocarcinogenesis

The development is in progress and the results are not available yet.

The retrospective database of 333 patients is already available and fully annotated. We have started to work on the development of the deep learning model but are facing methodological challenges. Indeed, we intend to develop a fully automated model and this is why we planned to use the liver mask automatically segmented using a deep learning model provided by Guerbet®. Because, the model requires to be trained on the non-tumoral parenchyma, we then manually cropped the tumour volume, which resulted in holes in the mask.

As a first step, we performed random cropping of bounding boxes (1-3 boxes with a random size between 20 and 30 mm) during data augmentation in patients in the low-risk group to mimic the non-tumoral liver mask in patients in the high-risk group. Using a validation set of 8 patients, we trained a 3D ResNet10 implemented in Pytorch (Gaussian Error Linear Unit activation function, learning rate of 0.01, ADAM optimizer, batch size of 8, binary cross-entropy loss and loss scaled to class imbalance x3.4 for patients in the high-risk group) on T2-weighted images with fat suppression. In addition to the random cropping of bounding boxes, data augmentation included random rotation, flips, and intensity shifts. Using an early stopping approach, the model achieved an accuracy of 1.0 after only 11 epochs without data augmentation and 0.93 after 15 epochs with data augmentation. These perfect results made us suspect a bias in the training related to the random boxes. This is why we are exploring other alternatives with patch-based and inpainting approaches. Random standardised cropping in the non-tumoral liver could also be used as input. In parallel, we are working on the co-registration of the MRI sequences (T2-weigthed imaged with fat suppression, diffusion weighted-images and ADC map, and T1-weighted images in/out-of-phase) to use the whole as a multiple input.

Detection of early-stage hepatocellular carcinoma in high-risk patients

DETECTION OF EARLY-STAGE HEPATOCELLULAR CARCINOMA IN HIGH-RISK PATIENTS WITH A DEEP LEARNING APPROACH USING ULTRASOUND AND MRI IMAGES

Currently in France, screening for hepatocellular carcinoma remains uniform for all patients and is based solely on abdominal ultrasound every 6 months. Although abdominal ultrasound is an inexpensive and radiation-free examination, it has significant shortcomings in sensitivity and inter-observer reproducibility in the detection of early-stage hepatocellular carcinoma. Tzartzeva reported a sensitivity of 47% in the detection of early-stage hepatocellular carcinoma (1 nodule < 5 cm or ≤ 3 nodules, each < 3 cm in diameter, without gross vascular invasion or extrahepatic metastases) in a meta-analysis including 13367 patients[10] while Park reported a sensitivity of 22.5% in the detection of very early-stage (< 2 cm) hepatocellular carcinoma[11]. Reasons are multiple: heterogeneity of the liver parenchyma in advanced cirrhosis, increased echogenicity with ultrasound attenuation in case of steatosis and obesity, etc.[13-15].

Thus, surveillance by MRI has been proposed to improve screening as it significantly outperforms ultrasound with a detection rate of 5 times that of ultrasound for very early-stage hepatocellular carcinoma[17]. Considering the higher cost and lower availability of MRI, abbreviated MRI (aMRI) (state-of-the-art MRI with fewer acquisition series) was recently introduced as it offers a considerable time-saving advantage, limited to 10 minutes at most, compared to the conventional MRI protocol, which takes 25 to 40 minutes[18-22]. Different protocols have been proposed with specific advantages, challenges and limitations: non-contrast (NC; T2-WI with fat suppression and/or DWI and/or T1-WI in/out), dynamic contrast-enhanced (DCE) and hepatobiliary phase (HB)[27]. Abbreviated non-contrast MRI has the advantages of the absence of contrast agent injection, simpler workflow, limited cost and the possibility to repeat poor quality acquisitions[27]. In addition, the inter-reader agreement could be lower with a non-contrast protocol[24]. The challenges and limitations of dynamic contrast-enhanced and hepatobiliary MRI are multiple: detection of inconclusive enhancing observations (need for recall examinations), injection of contrast, complex workflow with the need for intravenous access, and higher cost.

In NC-aMRI, the pooled per-patient sensitivity for hepatocellular carcinoma detection has been reported to range from 85 to 86.8% and specificity from 90.3 to 96% in three meta-analyses[18,23,24]. When results were stratified according to lesion size, the diagnostic sensitivity remained acceptable for very early-stage hepatocellular carcinoma (< 2 cm) with pooled sensitivity ranging from 69 to 77.1%, although it was lower than for larger hepatocellular carcinoma. This range of sensitivity for very early-stage hepatocellular carcinoma compared well with the pooled sensitivity of 70% for detecting 1-2 cm hepatocellular carcinomas on CE-MRI reported in another meta-analysis[54]. In DCE-aMRI, the reported per-patient sensitivity and specificity were also high, 84.6-92.1% and 81.6-100%, respectively[21,22,25,26]. In HB-aMRI, two meta-analyses reported a pooled sensitivity of 86-88.7% and 93-94%[18,24]. Considering the similar detection performances of the different protocols and their pros and cons, we believe that NC-aMRI is a promising protocol for screening programs. In the coming years, hepatocellular carcinoma screening programs will most likely rely on screening ultrasound and NC-aMRI, which motivates the urge to improve their detection performances.

The primary objective of the research was to develop an object detection model to improve detection of early-stage hepatocellular carcinoma on screening ultrasound and NC-aMRI.

The secondary objective was to develop an on-the-fly method for live ultrasound video annotations.

DETECTION OF EARLY-STAGE HEPATOCELLULAR CARCINOMA IN HIGH-RISK PATIENTS WITH A DEEP LEARNING APPROACH USING ULTRASOUND (STARHE CLINICAL TRIAL)

Material and methods

The methodology of the clinical study STARHE has been detailed in the previous section and only relevant specific methodological information is provided in this section.

Index test

We aimed to develop an object detection deep learning model based on ultrasound cine clips passing through the hepatocellular carcinoma to improve early-stage hepatocellular carcinoma detection.

Outcome

The main outcome of the study was the diagnostic performances of the AI model for the detection of early-stage hepatocellular carcinoma (1 nodule of any size or ≤ 3 nodules, each < 3 cm in diameter).

AI methodology

- **Database**: as mentioned above, the training, validation and testing sets were stratified according to potential confounders: aetiology of liver disease, FASTRAK score (binary cut-off of 9 points), ultrasound manufacturer, hepatocellular carcinoma size (binary cut-off of 2 cm) and echogenicity (isoechoic or not).

- **Annotation and labelling of databases**: ultrasound videos were annotated and labelled by a radiologist subspecialised in liver imaging using the MOSaiC Annotation Platform co-developed by the IHU Strasbourg and Université de Strasbourg[55]. Hepatocellular carcinomas were annotated using a standardised annotation pipeline with keyframe interpolation, as follows (**Figure 5**): (i) clicking on the two corners of a tight box around the object ① ②, (ii) navigating the video with play/forward/backward ③, (iii) pause on the next keyframe, (iv) go back to step (i).

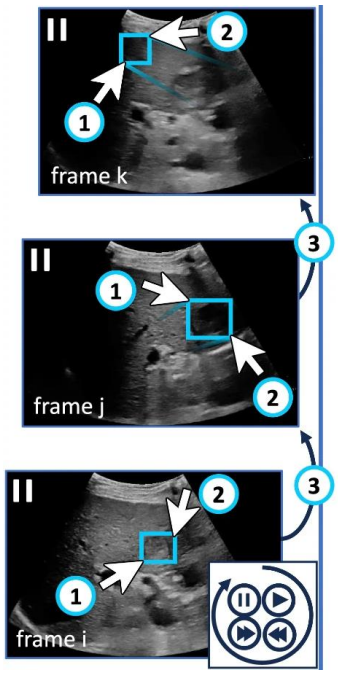


Figure 5 – Annotation process

- **Pre-processing of ultrasound images** (identical).
- **Early hepatocellular carcinoma detection**. The goal was to detect and localize hepatocellular carcinoma as the video is played, highlighting the hepatocellular carcinoma with a bounding box. We selected three state-of-the-art models for this purpose: Faster-RCNN, DINO-DETR, and RTMDet. Each model was pretrained on the COCO dataset to facilitate transfer learning. Our implementation was based on the MMDetection library. We used a stratified validation set of 15 patients to select the model architecture and tune the hyperparameters (Table 5).
- **Independent testing**: To ensure the robustness and generalisability of our AI models, we planned to test them in an independent dataset (same 25 patients in the high-risk group included in the classification testing set). First, we tested the detection model on the ultrasound cine clips of the hepatocellular carcinoma of the same 25 patients included in the classification testing set from the high-risk group and computed the detection metrics (refer to statistical analysis section). Then, we intended to simulate a testing dataset as representative as possible to real life practice with homogeneous liver, heterogeneous liver, and focal liver lesions. Therefore, we ran the detection model on the ultrasound cine clips of the non-tumoral parenchyma of the same 50

patients included in the testing dataset designed for the classification task (risk stratification of hepatocarcinogenesis), resulting in 50 additional ultrasound cine clips: 25 from the low-risk group, i.e. hypothesised to represent more homogeneous liver, and 25 from the high-risk group, i.e. more heterogeneous liver. The objective was to evaluate the rate of false positives.

	Model and hyperparameters	Final model
Batch sizes	2, 4, 8, 16	8
Learning rate	0.0008, 0.0006, 0.0004, 0.0002, 0.0001, 0.001, 0.002, 0.005	45 epochs with learning rate of 0.0002 (divided by 10 at 20 and 40 epochs)
Frozen stages	-1, 0, 1	-1 (none)
Batch requires grads	True, False	True
Transfer learning	All except prediction head	All except prediction head
Models	Faster-RCNN, DINO-DETR, RTMDet	RTMDet

Table 5 - Model and hyperparameters selection using a stratified validation set (15 patients)

Statistical analysis

The estimate of performance metrics of the detection model was computed for mean average precision (mAP; area under the precision-recall curve) with a predefined intersection over union (IoU) of 10, 50, and 75% (these arbitrary thresholds were chosen because they are commonly reported in the literature). Confusion matrices were computed for each patient at different confidence levels for an intersection over the union of 10% to assess the rate of true positive and false positive. A 10% threshold was chosen because of the screening strategy where detecting a lesion is more important than correctly delineating it.

Results

Detection of early-stage (BCLC 0 and A) hepatocellular carcinoma

RTMDet achieved good to excellent performances in the testing dataset with a mAP10 of 0.67 and a mAP50 of 0.50 (**Figure 6**).

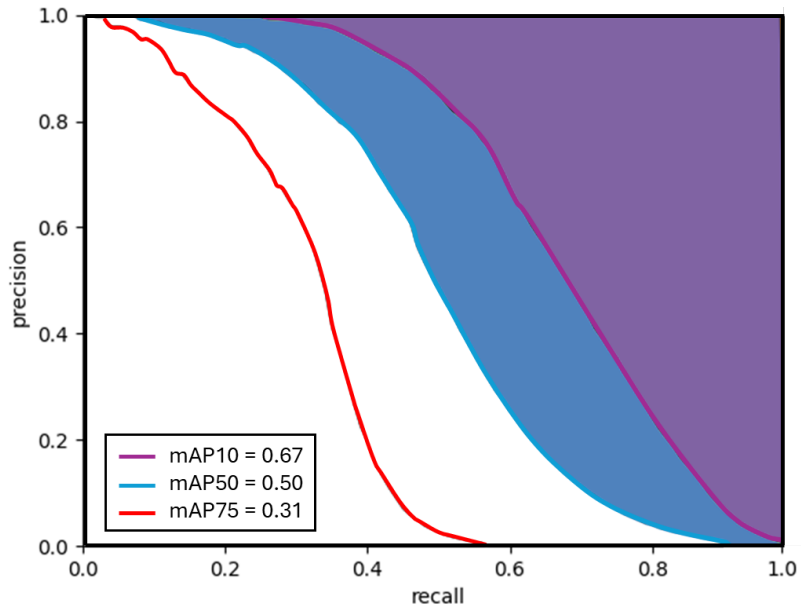


Figure 6 - Mean Average Precision (Intersection over Union of 10%, 50%, 75%) with precision-recall curve obtained by plotting the model's precision and recall values as a function of the model's confidence score threshold.

Table 6 demonstrates the rate of detected lesions and false positives with its median (predicted boxes in normal liver) when the intersection over the union between the predicted box and the annotated box was set to 10%. The confidence level of 70% appeared clinically interesting as 68% of lesions were correctly detected with false positives in only 20% of patients (median of false positives per video of 1 [IQR = 1-2]). Figure 6 illustrates prediction examples in patients with small hepatocellular carcinomas.

	Rate of detected lesions	Rate of false positives	Median number of false positives
Confidence of 30%	96%	96%	62 [19-113]
Confidence of 40%	96%	88%	24 [5-51]
Confidence of 50%	84%	72%	8 [3-22]
Confidence of 60%	76%	52%	4 [1-8]
Confidence of 70%	68%	20%	1 [1-2]
Confidence of 80%	40%	0%	NA

Table 6 – Rate of detected lesions and false positives on ultrasound cine clips of 10 seconds (total number of frames between 200-250) in patients with early-stage hepatocellular carcinoma (intersection over union between the predicted box and the annotated box set to 10%)

At a confidence level of 70% and an intersection over the union of 10% (**Table 7**), the rate of detected lesions was excellent for small hepatocellular carcinomas (67%) and larger hepatocellular carcinomas (80%). On the other hand, although the rate of detected hypoechoic and hyperechoic lesions was excellent (75% and 100%, respectively), it was only moderate for isoechoic and heterogeneous lesions (50% for both).

	Rate of detected lesions	Rate of false positives	Median number of false positives
Nodule size			
≤ 20 mm (*n* = 12)	67%	33%	2 [1-4]
20-30 mm (*n* = 8)	63%	13%	1 [1-1]
> 30 mm (*n* = 5)	80%	0%	NA
Nodule echogenicity			
Hypoechoic (*n* = 8)	75%	13%	1 [1-1]
Isoechoic (*n* = 8)	50%	25%	5 [3-6]
Hyperechoic (*n* = 5)	100%	20%	1 [1-1]
Heterogeneous (*n* = 4)	50%	25%	2 [2-2]

Table 7 – Subgroup analysis with a confidence level of 70% and an intersection over the union of 10%.

Table 8 shows the rate of false positives on ultrasound cine clips of 10 seconds (total number of frames between 200-250) of early-stage hepatocellular carcinoma and non-tumoral parenchyma with a confidence level of 70% and an intersection over the union of 10% were similar.

	Rate of detected lesions	Rate of false positives	Median number of false positives
Ultrasound cine clips of early-stage HCC (*n* = 25)	68%	20%	1 [1-2]
Ultrasound cine clips of high-risk non-tumoral parenchyma (*n* = 25)	NA	8%	2 [2-2]
Ultrasound cine clips of low-risk non-tumoral parenchyma (*n* = 25)	NA	25%	1 [1-22]
Non-tumoral parenchyma (*n* = 50)	NA	18%	2 [1-14]
All	68%	19%	2 [1-8]

Table 8 – Rate of detected lesions and false positives on ultrasound cine clips of 10 seconds (total number of frames between 200-250) of early-stage hepatocellular carcinoma (HCC) and non-tumoral parenchyma with a confidence level of 70% and an intersection over the union of 10%.

Figures 6, **7** and **8** shows illustrative examples of true positive and false positive of hepatocellular carcinomas of different sizes and echogenicities.

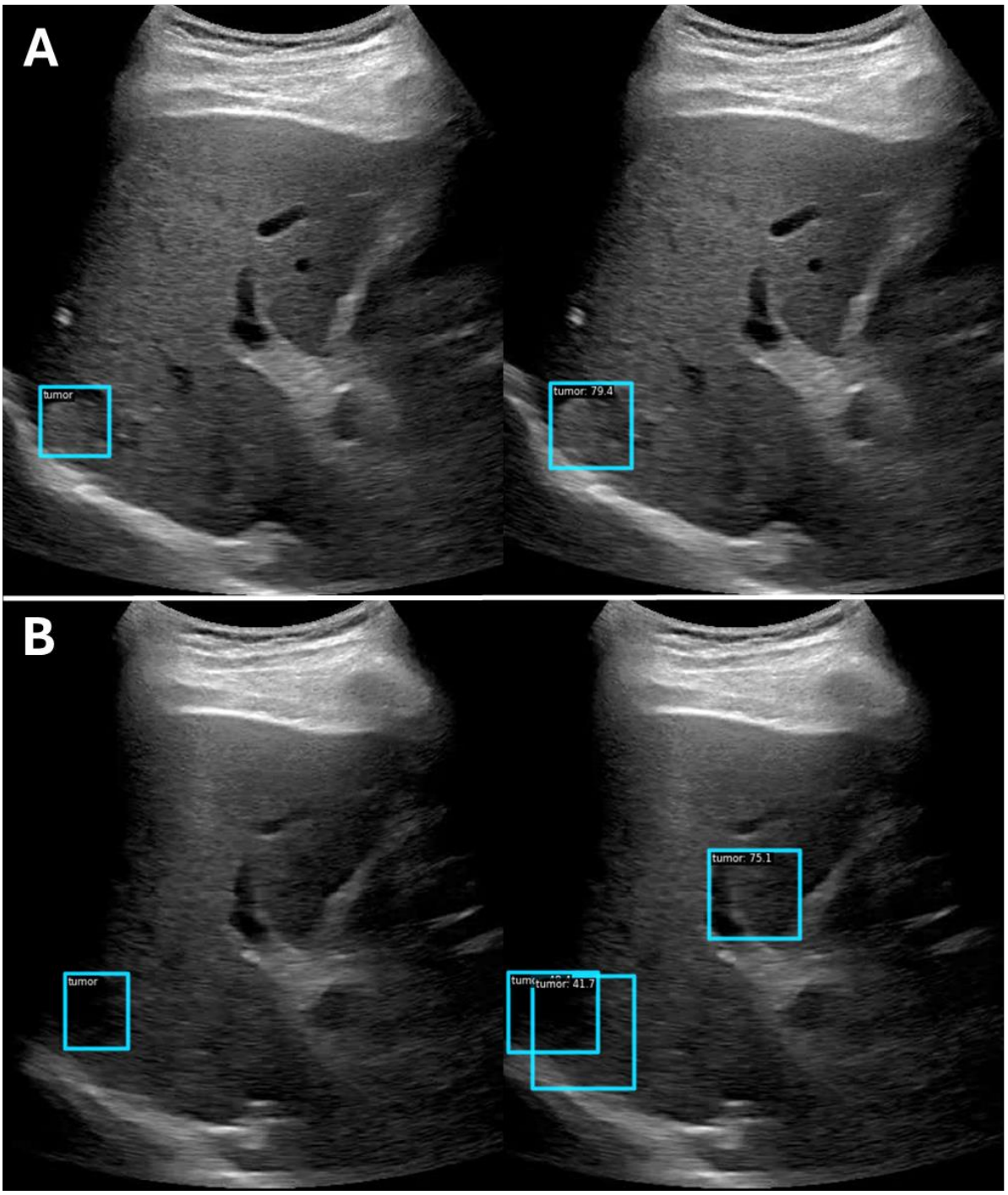


Figure 6 – True positive (A) and false positive (B) in a 74-year-old patient with metabolic-dysfunction associated steatotic liver disease and 17 mm isoechoic hepatocellular carcinoma. Figure 6A shows the deep hepatocellular carcinoma correctly detected with a confidence level of 79%. Figure 6B shows a false positive prediction in the non-tumoral parenchyma with a high confidence level (75%). On Figure 6B, the partially obscured lesion was detected with a lower confidence level (42%).

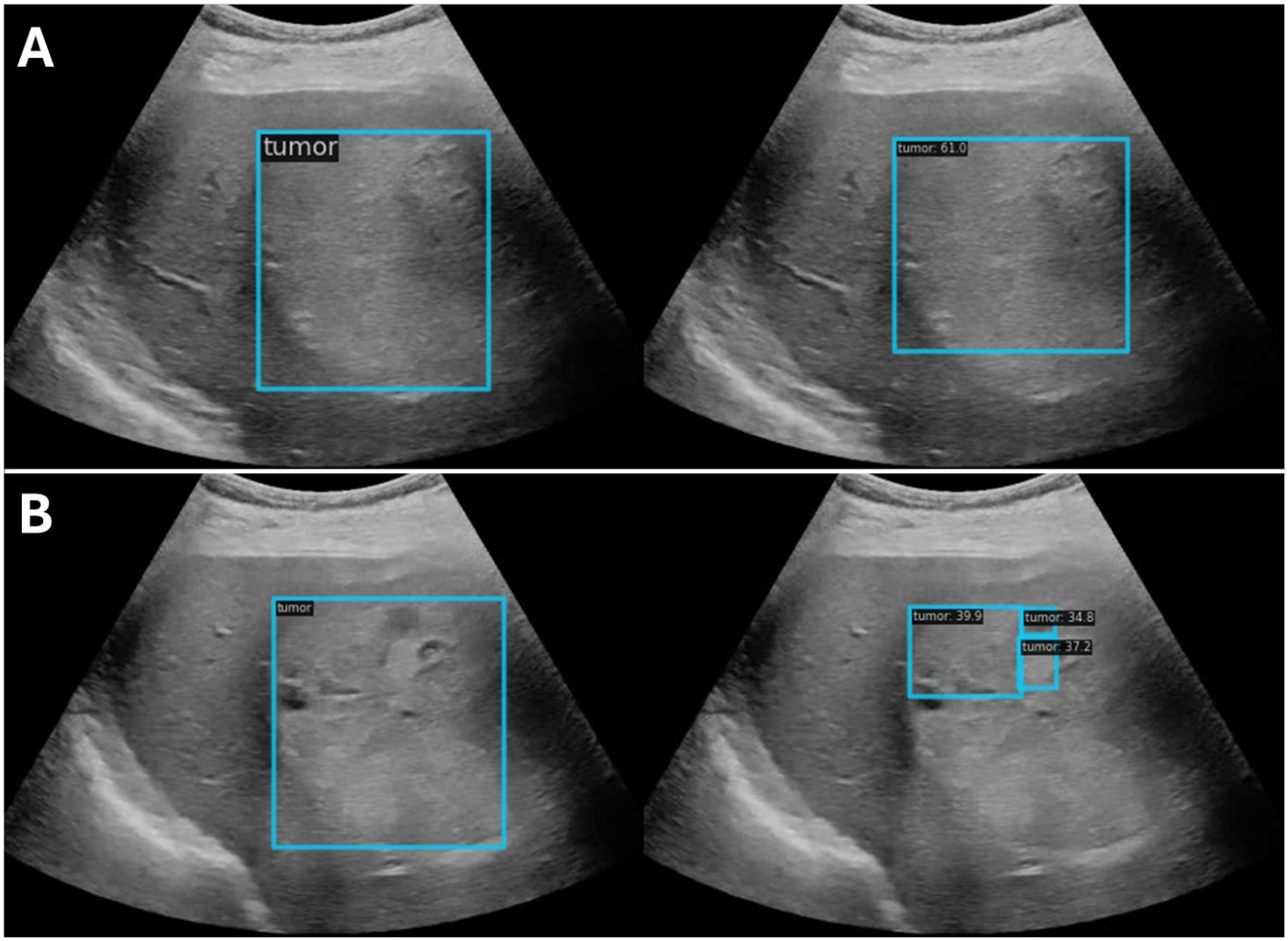


Figure 7 – Large mildly hyperechoic hepatocellular carcinoma in a 74-year-old patient with alcohol-related cirrhosis. Multiple boxes were predicted in the tumour area, mostly of smaller size. However, only one reached the confidence level of 60% and none reached the confidence level of 70%. This might be explained by the underrepresentation of such large lesions in the training dataset.

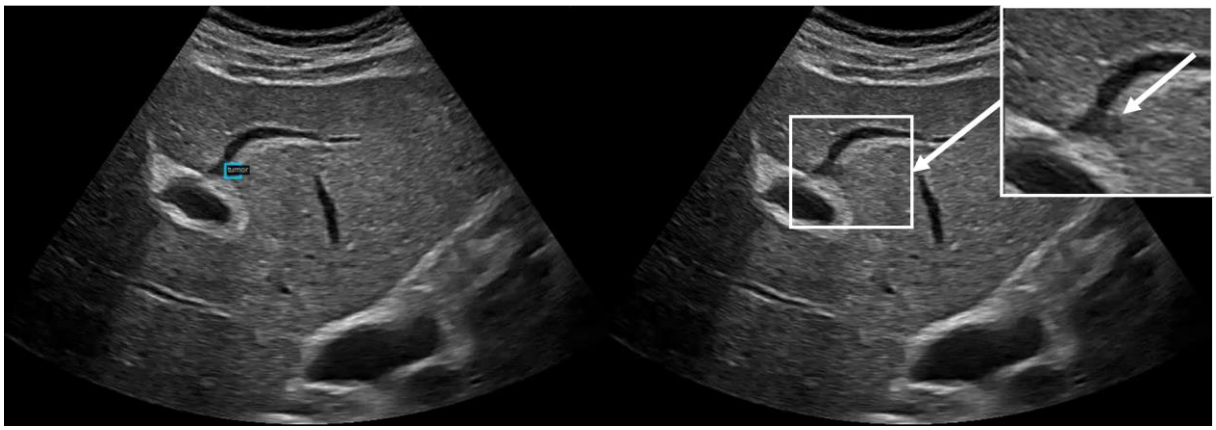


Figure 8 – False negative in a 61-year-old patient with alcohol related cirrhosis and 10 mm hypoechoic hepatocellular carcinoma.

Conclusion

The developed object detection model achieved excellent performances in detecting very-early stage (< 2 cm) and early-stage hepatocellular carcinomas (overall rate of detected lesions = 68% and mAP10 = 0.67) on ultrasound cine clips. A threshold of 10% was chosen for the intersection over the union between the ground truth annotation and the predicted boxes. Indeed, in a screening strategy, detecting a lesion is more important than correctly delineating it. Furthermore, the confidence level in the predicted box of 70%, to be considered a positive prediction, was chosen based on a clinically relevant balance between the rate of detected lesions and that of false positives, a potential time-consuming hurdle. This threshold should be tested in prospective longitudinal studies alongside radiologists' reading.

ON-THE-FLY POINT ANNOTATION FOR FAST MEDICAL VIDEO LABELLING

As outlined in the prior sections, we aim to improve the detection of early-stage hepatocellular carcinoma in high-risk patients with a deep-learning approach using ultrasound. This medical challenge is considered an object detection task in the field of computer science. Object detection is increasingly recognized as a critical tool in medical video analysis, with applications ranging from identifying landmarks, organs and lesions to tracking surgical instruments in real-time procedures to assess safety. Yet, while modern deep learning-based object detectors have shown notable success, their effectiveness is largely dependent on the availability of extensive annotated data. This becomes particularly challenging in the medical domain, given the labour-intensive nature of annotation and the constraints on experts' time, especially given their primary clinical duties. Currently, the process of video object annotation is frame-based, which is suboptimal and highly time-consuming, even with the use of interpolation tools. As a result, most studies rarely annotate more than one or two hundred videos. This methodological constraint hampers the democratization and scalability of deep learning in medical procedures, as there is a compelling requirement for the collection and annotation of diverse, multicentre, and multi-operator data.

To address these challenges in domains where the opportunity cost of expert time is high, we proposed a new annotation paradigm focused on live video annotation, which corresponds to the challenges of ultrasound. Our proposal is a shift from the conventional frame-based approach to a more dynamic video-based point annotation strategy. We introduced an on-the-fly point annotation pipeline, developed and tested on the STARHE dataset (NCT04802954), enabling live video annotation to mitigate the tedious efforts specifically associated with video annotation. Every frame in the video was weakly-annotated, ensuring expert guidance throughout the process. Our method proved to allow precise tracking of structures, which can enable useful pseudo-labels generation compatible with weakly semi-supervised object detection pipelines, outperforming conventional annotation method at equivalent annotation budgets. A notable

reduction in annotation costs was observed through the utilization of this strategy. The findings of this study underscored the need for optimization of video annotation processes, enabling the development of high-quality datasets. This approach fostered a more efficient utilization of expert resources, optimizing the balance between annotation accuracy and cost-effectiveness in medical imaging studies.

This work was published in the International Journal of Computer Assisted Radiology and Surgery and was awarded the CASCINATION & Zeiss Machine Learning in CAI Award: Runner-up at the International Conference on Information Processing in Computer-Assisted Interventions (IPCAI) 2024.

International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-024-03098-y

ORIGINAL ARTICLE

On-the-fly point annotation for fast medical video labeling

Adrien Meyer[1] · **Jean-Paul Mazellier**[1,2] · **Jérémy Dana**[2,3] · **Nicolas Padoy**[1,2]

Received: 1 March 2024 / Accepted: 4 March 2024

Abstract

Purpose: In medical research, deep learning models rely on high-quality annotated data, a process often laborious and time-consuming. This is particularly true for detection tasks where bounding box annotations are required. The need to adjust two corners makes the process inherently frame-by-frame. Given the scarcity of experts' time, efficient annotation methods suitable for clinicians are needed.

Methods: We propose an *on-the-fly* method for *live video annotation* to enhance the annotation efficiency. In this approach, a continuous single-point annotation is maintained by keeping the cursor on the object in a live video, mitigating the need for tedious pausing and repetitive navigation inherent in traditional annotation methods. This novel annotation paradigm inherits the point annotation's ability to generate pseudo-labels using a point-to-box teacher model. We empirically evaluate this approach by developing a dataset and comparing on-the-fly annotation time against traditional annotation method.

Results: Using our method, annotation speed was 3.2× faster than the traditional annotation technique. We achieved a mean improvement of 6.51 ± 0.98 AP@50 over conventional method at equivalent annotation budgets on the developed dataset.

Conclusion: Without bells and whistles, our approach offers a significant speed-up in annotation tasks. It can be easily implemented on any annotation platform to accelerate the integration of deep learning in video-based medical research.

Keywords Live video annotation · Deep learning · Object detection · WSSOD

Introduction

Object detection is increasingly recognized as a critical tool in medical video analysis, with applications ranging from identifying landmarks, organs and lesions to tracking surgical instruments in real-time procedures to assess safety [1, 2]. Yet, while modern deep learning-based object detectors have shown notable success, their effectiveness is largely dependent on the availability of extensive annotated data [3–5]. This becomes particularly challenging in the medical domain, given the labor-intensive nature of annotation and the constraints on experts' time, especially given their primary clinical duties. Currently, the process of video object annotation is frame-based, which is suboptimal and highly time-consuming, even with the use of interpolation tools. As a result, most studies rarely annotate more than one or two hundred videos [6, 7]. This methodological constraint hampers the democratization and scalability of deep learning in medical procedures, as there is a compelling requirement for the collection and annotation of diverse, multicentre, and multi-operator data.

To address these challenges in domains where the opportunity cost of expert time is high, methods to accelerate the annotation process generally fall into one of two non-exclusive categories:

1. Methods to scale the pool of available annotators by lowering the expertise barrier. For instance, experts could annotate keyframes that non-experts, possibly crowd-sourced contributors, use as references to expand the dataset [8]. However, it requires sufficient resources for

✉ Adrien Meyer
adrien.meyer@ihu-strasbourg.eu

Jean-Paul Mazellier
jean-paul.mazellier@ihu-strasbourg.eu

Jérémy Dana
jeremy.dana@ext.ihu-strasbourg.eu

Nicolas Padoy
npadoy@unistra.fr

[1] ICube, CNRS, University of Strasbourg, Strasbourg, France

[2] IHU Strasbourg, Strasbourg, France

[3] Department of Diagnostic Radiology, McGill University, Montréal, Canada

Published online: 04 April 2024

managing human capital at scale. The gains in annotation production must outweigh the overheads of annotators training, follow-up of annotation quality and project management.
2. Methods to lessen the annotation burden per annotator through the use of weaker localization labels, such as absence/presence, gaze, scribbles, or points [9, 10]. The weak labels are used in a (Weakly) Semi-Supervised Object Detection ((W)SSOD) pipeline, which uses a small set of box-level labeled images as well as a larger set of weakly labeled images to train detectors. In Point-DETR [10] and Group R-CNN [11], the authors propose to learn a point-to-box teacher model to generate pseudo-box labels.

We position our work on the later categories. We propose a new annotation paradigm focused on *live video annotation*. Our proposal is a shift from the conventional *frame-based* approach to a more dynamic *video-based point annotation* strategy. Unlike bounding boxes, point annotations have only two degrees of freedom (x and y coordinates) and can be adjusted with a single drag of a pointer (i.e., mousepad, pencil on tactile screen or eye gaze). Leveraging this inherent property, we introduce *on-the-fly* annotation, where a 2D cursor enables continuous tracking on *live video*, producing a point annotation similar to a temporal scribble (see Fig. 1b). This eliminates frame-by-frame annotation, speeding up the process while maintaining the advantages of point-based WSSOD. We leverage point-to-box teacher models [10, 11] for the generation of pseudo-box labels derived from the point annotations to train detectors.

To validate our annotation method, we constructed a liver ultrasound video dataset and compare the efficiency of standard bounding box and on-the-fly point annotation (OTF) methods. Given the high level of expertise required in the ultrasound domain, enhancing the annotation efficiency of experts is crucial. With varying annotation budgets, we train two type of teacher models: one using point-to-box models in a WSSOD paradigm with OTF and the other as a traditional object detector. These teachers generate pseudo-labels for training student models. Unlike methods utilizing interpolation or crowdsourcing, our approach ensures that every frame in the video is weakly-annotated by the experts.

Our contributions are twofold: (1) We introduce the novel task of live video annotation, and (2) we present an on-the-fly point annotation method optimized for this task within a WSSOD framework.

Related work

Crowdsourcing annotations

Crowdsourcing aims to match expert annotation performance in tasks like image annotation. In healthcare domains, the study by [12] on hepatic steatosis reveals that crowdworker annotation reliability is not guaranteed by annotator certainty or agreement, yet a larger crowd slightly outperformed a few experts. The study by [13] suggests that crowds can refine automatic 3D segmentation of liver CT scans to a level comparable to experts, although at a slower pace. Crowdsourcing can offer scalability in annotations but requires substantial setup that becomes cost-effective only at large project scales. Its suitability varies by project and modality and may underperform in tasks demanding high expertise.

Semi-supervised/weakly supervised object detection

Semi-Supervised Object Detection (SSOD) and Weakly-Supervised Object Detection (WSOD) aim to mitigate the high cost of data annotation. SSOD methods leverage a mix of a few box-level labeled images and many unlabeled ones with two main approaches; consistency regularization techniques, to stabilize the detector's predictions across variably augmented images [14], and pseudo-labeling, where a teacher model trains on labeled data to generate pseudo-labels for unlabeled data. A student model then trains on both datasets for improved performance [15, 16]. WSOD methods use abundant but weakly annotated data, such as image labels [17, 18]. The studies by [19] and [20] utilize class activation maps in WSOD methods to enable both detection and localization of surgical tools in endoscopic videos and breast cancer in ultrasound images respectively, without spatial annotations.

Combining these approaches, Weakly Semi-Supervised Object Detection (WSSOD) methods use both box-level and weakly labeled images to train detectors, aiming to propose a favorable trade-off between annotations cost and performances. [21] detect lung consolidations in ultrasound videos, using video-level labels (presence in at least one frame) and a teacher-student training strategy.

Methodology

In this section, we first introduce the task of live video annotation and discuss why bounding box annotation is suboptimal from a video annotation point of view. Next, in order to address it, we illustrate our novel on-the-fly point annotation as an efficient alternative.

Live video annotation

We introduce Live Video Annotation as a new approach to streamline the process of object annotation in videos. The key goal of live video annotation is to reduce or completely eliminate the need to frequently pause the video, thereby making the annotation process more fluid and efficient. This raises a critical question: How can a dataset be densely annotated with spatial instance information (Fig. 1a) without substantial pausing?

Considering the standard annotation pipeline as performed on a dedicated video annotation software with keyframes interpolation (Fig. 1a), the annotation process is as follows: (i) clicking on the two corners of a tight box around the object ① ②, (ii) navigating the video with play/forward/backward ③, (iii) pause on the next keyframe, (iv) go back to step (i). These convoluted steps are the result of the multi-click nature of boundary annotation, which is not compatible with a continuous annotation on a streamed video. In this paper, the aforementioned method will be referred to as the 'BBox method' for ease of discussion and simplicity. From those observations, we propose to use a weaker localisation label such as the point, which only requires one click/drag to adjust its position over a continuous video playback.

On-the-fly point annotation

We propose a novel *on-the-fly* point annotation (OTF) strategy for video streaming. In this scenario (Fig. 1b), the user is asked to continuously point at the targeted structure during the video playback ①, reducing the tedious pausing and back-and-forth video navigation associated with the standard annotation method (Fig. 1a). In practice, the annotator still needed to pause occasionally for video understanding, taking breaks, or stopping the live annotation when the object disappeared from view. In videos where objects frequently appear and disappear, continuously stopping OTF annotations can be inefficient as it interrupts the live annotation process to precisely find the frames where the object is not visible. A smoother, less conservative approach could involve performing annotations purely on-the-fly, without pausing the video, thereby enhancing workflow fluidity and reducing annotation time. However, to maintain the quality of annotations and mitigate the risk of introducing false positives, it might be prudent to exclude annotations at the temporal edges corresponding to the annotation stoppages. In our experiment, we adopt the more conservative approach of stopping the video and precisely stopping the annotation when the object disappears. In the annotation process for our study, we utilized a reduced playback speed of $0.2\times$. This slower speed was essential to accurately track rapid changes in the videos, which are difficult to observe at the normal speed ($1\times$). This adjustment helped minimize potential errors in annotation. However, we recognize that this method may not be universally applicable, particularly in scenarios involving faster movements, and the choice of playback speed might need to be tailored to the specific dataset being annotated. The resulting annotation maintains the advantages of point-based WSSOD, i.e. Point-DETR [10] and Group R-CNN [11].

We adopt the self-training pipeline of [10]. Given a small number of supervised images and a large number of weakly supervised images: (i) Train a teacher model on available labeled images, (ii) Generate pseudo-labels of weakly OTF annotated images using the trained teacher model and (iii) Train a student model with fully labeled images and pseudo-labeled images. It's important to note that the box-level pseudo-labels produced by these teacher models are neither verified nor corrected during the training of the student models, in order to keep the annotation budget manageable.

To verify the benefit of the proposed OTF, we used a dataset named STARHE of liver ultrasound videos, as described below, with annotated lesions with both bounding box annotation and OTF. We timed both annotation methods on subsets of the annotated videos to compare the annotation speed. Next, we studied whether the points annotated using OTF accurately lie within the corresponding bounding boxes. This assessment allowed us to determine the consistency of the OTF points and whether they accurately tracked and moved in alignment with the objects being annotated. Finally, we conducted comparative studies between two self-training scenario S_{OTF} and S_{BBox}. The first leveraged OTF while the latter did not. This comparison aimed to assess the compatibility and effectiveness of OTF-based pseudo-labels within a WSSOD pipeline.

We employ DETR [22] and Faster R-CNN [23] as our student models. DETR uses the transformer architecture to simplify object detection, removing hand-crafted elements like non-maximum suppression and anchor generation, while maintaining performance on par with Faster RCNN [23]. We employ Point-DETR and Group R-CNN as a teacher models in S_{OTF}. Point-DETR extends DETR by incorporating both images and point annotations as inputs. It employs a point encoder to map these point annotations to object queries, enhancing detection performance through strong prior localization and class. Group R-CNN [11], building on

classic R-CNN architecture, introduces instance-level proposal grouping and assignment, coupled with instance-aware representation learning, to effectively translate point annotations into precise box proposals. Those models are trained on a box-level annotated video set $Box(S_{OTF})$, and subsequently used to create pseudo- labels on the weakly labeled videos $OTF(S_{OTF})$. Therefore, the corresponding annotation budget B_{OTF} can be expressed as

$$B_{OTF} = T_{Box} \times |Box(S_{OTF})| + T_{OTF} \times |OTF(S_{OTF})| \tag{1}$$

where T_{BBox} and T_{OTF} represent the average annotation times required to annotate a video using the bounding box or OTF method, respectively. $|Box(S_{OTF})|$ and $|OTF(S_{OTF})|$ are the number of annotated videos with BBox or OTF method for the scenario S_{OTF}, respectively. For S_{BBox}, which utilizes the classic DETR as a teacher model, the annotation budget B_{BBox} is calculated as

$$B_{BBox} = T_{BBox} \times |Box(S_{BBox})| \tag{2}$$

Since no annotation is required during inference, this model budget solely depends on the BBox method. To ensure a fair comparison between the two models, we increase the number of annotated videos $|Box(S_{BBox})|$ such that $B_{BBox} = B_{OTF}$. In this way, the time spent on weak annotations is effectively converted into additional box-level annotated videos.

Experimental setup

STARHE dataset

We developed and tested our OTF method using a newly created dataset, STARHE (Risk Stratification of Hepatocarcinogenesis), registered at ClinicalTrials.gov (Identifier: NCT04802954). This dataset gathered liver ultrasound videos acquired using a standardized protocol. The current hepatocellular carcinoma (HCC) screening program in France relies on biannual liver ultrasound. However, the performance of this screening program is poor which can be explained by poor liver visualization using ultrasound in some patients (e.g., obesity, steatosis,...), operator dependency, or limited patient compliance. Given the anticipated surge in HCC-related mortalities by 2030, a more effective screening strategy is needed. Our aim is to develop an automated method for detecting HCC lesions during ultrasound screenings, thereby minimizing missed lesions and delays in diagnosis.

In our study, an experienced clinician (radiologist) annotated a set of 125 ultrasound videos with dedicated video annotation software with interpolation tools, employing both OTF and BBox annotation methods for each video, as shown in Fig. 1. The annotations were performed using a mouse cursor. A minimum one-month interval was maintained between each annotation type to ensure no recall bias, with previous annotations being hidden during the subsequent session. For a subset of 27 videos, the annotation process was timed to compare the efficiency of both methods. Specifically, in

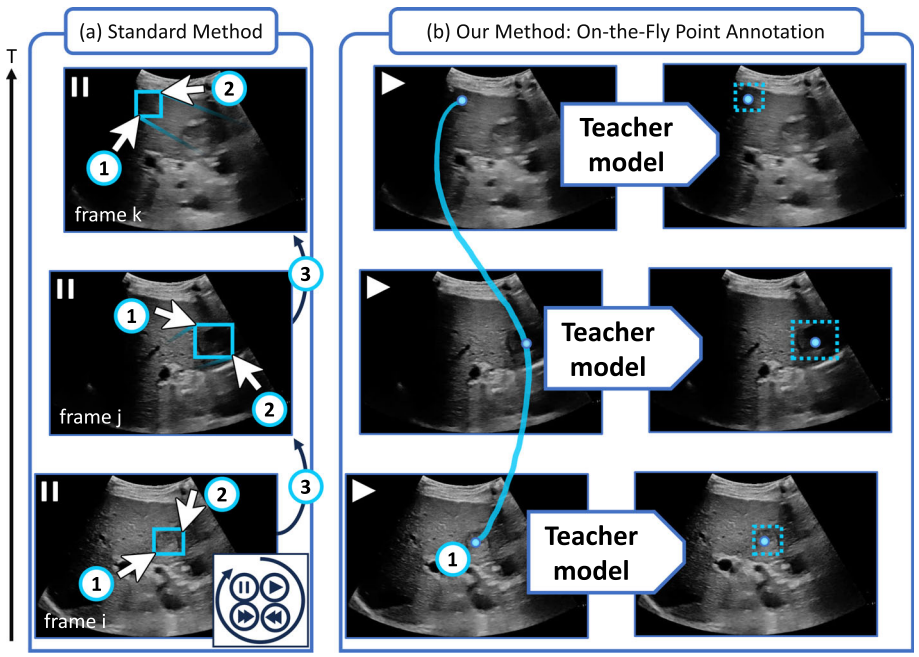


Fig. 1 **a** Conventional bounding box annotation approach on static frames. ① ② adjust the two corners, ③ video navigation; **b** Our proposed *on-the-fly* point annotation method on live video. ① pointing of the targeted structure. Box cyan—lesion; solid lines—ground truth; dashed lines—predictive pseudo-labels

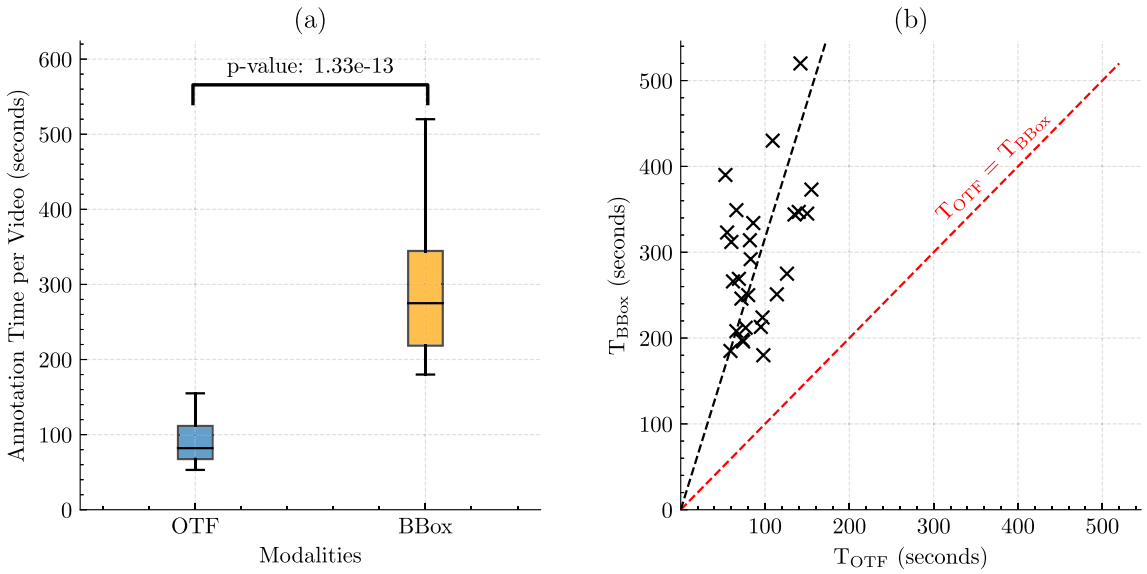


Fig. 2 **a** Box plot comparing annotation times between OTF and BBox method. **b** Pairwise comparison of annotation times for each timed video, with a fitted line illustrating the relationship between T_{BBox} and T_{OTF}

S_{OTF}, annotations were made on live videos played at 0.2× speed, ensuring a comprehensive understanding of the video content throughout the annotation process. The videos in our dataset had a duration of 10 s. We partitioned our dataset as follows: 20% for testing (25 videos), 10% for validation (13 videos), and 70% for training (87 videos). Within the training set, we further divided the data into a box-level annotated set and a weakly annotated set to conduct experiments with varying annotation budgets. Initially, the division was set at 20% box-level annotated and 80% weakly annotated. We then incrementally transferred 5% from the weakly annotated set to the box-level annotated set, until the distribution reached 60% box-level annotated and 40% weakly annotated. We report the average precision at an intersection over union threshold of 0.5 (AP@50), averaged over 3 runs with random data splitting.

Teacher models training

In S_{OTF}, we use Point-DETR and Group R-CNN as our teacher models. We pretrain the models on 20% of the COCO dataset [24], limited by computational budget constraints, preventing the use of the full dataset. 70k iterations of AdamW training with mini-batch 4 on 2× Nvidia V100 are performed for fine-tuning on our STARHE dataset, using an initial learning rate of $1e^{-4}$. We divide the learning rate by 10 at 50k iterations. We use random flipping, resizing and random cropping as data augmentation. In our experiments, we train the teacher models with noise on the point up to 25% of its respective box dimension. Note that point noising

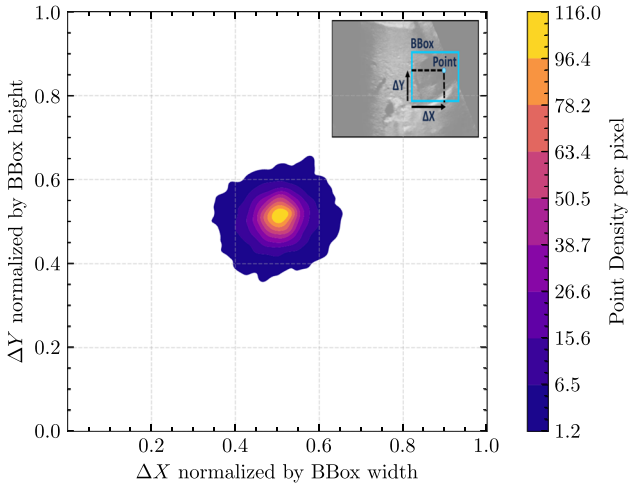


Fig. 3 Spatial density of OTF annotation locations across all videos with respect to corresponding boxes. The x-axis represents the horizontal position, and the y-axis represents the vertical position of the OTF within the box. Yellow areas indicate regions with a higher density of annotations, while dark blues indicate a lower density

serve as a data augmentation and is not used during inference. To reduce frame redundancy, we trained the models using every eighth frame from the annotated videos, approximately equating to 2.5 frames per second. Our implementation is based on the MMDetection library [25]. Similarly, in S_{BBox}, we employ DETR as our baseline teacher model. In both S_{OTF} and S_{BBox}, we use DETR and Faster R-CNN as our student models. The student models utilizes pretrained weights from the entirety of the COCO dataset.

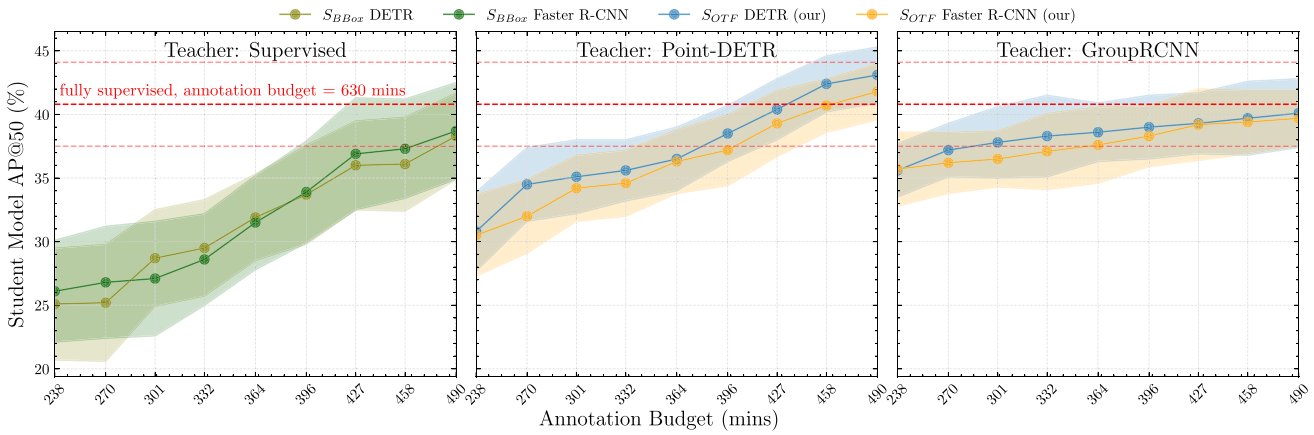


Fig. 4 AP@50 of Student models under similar annotation budgets, utilizing a blend of box-level and pseudo-labels. Results, along with the standard deviation, are computed based on three individual runs. S_{OTF} pseudo-labels are from point-to-box models using OTF annotation, whereas S_{BBox} uses Faster R-CNN or DETR-derived pseudo-labels without prior

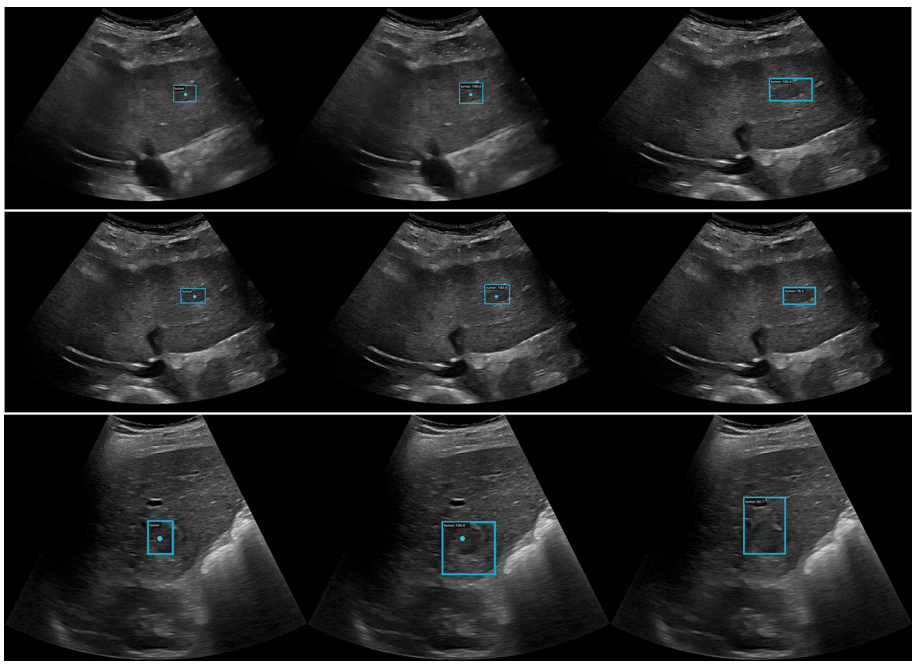


Fig. 5 Qualitative results of S_{OTF} are presented. The left column displays the ground truth, featuring both point and corresponding bounding box annotations for lesions. The middle column depicts the predicted pseudo-labels with S_{OTF}, and the right column the prediction from the fully supervised model

Results

In our initial experiment, we examined the annotation speed between the OTF and BBox methods. As illustrated in Fig. 2a, we observed a statistically significant acceleration, with the OTF method being on average 3.2 times faster than the BBox method ($p = 1.33e^{-13}$). We present a pairwise comparison of annotation times for each video in Fig. 2b, illustrating a distinct relationship between T_{BBox} and T_{OTF}. This visualization highlights a consistent pattern: the longer a video takes to annotate using one method, the longer it tends to take using the other method as well (see the fitted line). This suggests a consistent difficulty level in video annotations, such as lesion visualization due to poor conspicuity, irrespective of the method used. Still, the OTF method consistently proves to be significantly faster than the BBox method in our annotation scenarios.

To better understand the distribution and localization of annotated points with respect to corresponding BBox, we employ a kernel density estimation plot as illustrated in Fig. 3. This allows us to present the data as a heatmap. The axes, ranging from 0 to 1, represent the relative dimensions of the boxes, with values indicating the normalized position of corresponding OTF annotation within them. Areas with a higher color intensity signify regions with a denser concentration of annotations. Two key observations are made regarding the

OTF annotations. First, all OTF annotations systematically fall within their respective boxes, confirming the precision of this annotation method. Secondly, a significant concentration of OTF annotations is observed around the near-center regions of the boxes, despite the annotator not being explicitly instructed to target the center of the structures. This denotes that the OTF method effectively facilitates accurate tracking of the anatomical structures in question. This tendency toward centre-annotation could potentially be an instinctive approach adopted when tracking oval-like structures, such as lesions, enabling a more intuitive annotation process.

Finally we investigated the integration of OTF labels into a WSSOD pipeline. A comparative study was conducted between the two self-training scenarios, S_{OTF} and S_{BBox}, to evaluate the effectiveness of our method in generating pseudo-labels for downstream applications. We report the AP@50 and standard deviation, calculated over three runs, with equivalent annotation budget in Fig. 4. S_{OTF} achieves a mean improvement of 6.51 ± 0.98 AP@50 over S_{BBox}. Using Group R-CNN, we achieved better results than Point-DETR, especially in scenarios with smaller annotation budgets. With a 238-minute annotation budget, Group R-CNN combined with DETR achieved an AP@50 of 35.7%, and 35.6% when paired with Faster-RCNN. This success is due to Group R-CNN's multi-scale approach and CNNs' efficiency with limited data. Interestingly, S_{OTF} even surpasses the performance of the fully supervised scenario, which we infer to be a consequence of a label smoothing effect induced during the pseudo-label generation process which aligns more closely with the expectations of the student model, facilitating more effective learning and acting as a regularization mechanism. While S_{OTF} exceeds the fully supervised model's performance for annotations budget over 427 min, it remains within the error margin of the supervised model. Overall, we achieve the same performance as the fully supervised baseline with 68% of its annotation budget. Examining the performance of teacher models, Group R-CNN achieves an AP@50 of 65.2% with 20% of the data strongly labeled and 73.4% with 60%, while Point-DETR reaches 58.2% and 73.9%, respectively, under the same conditions. The teacher models achieve a notably high AP@50, and the student models, trained using the pseudo-labels derived from these teachers, attain results comparable to their fully supervised counterparts. This underscores the scalability and effectiveness of the proposed method. As a comparison, we employ WSOD methods[17, 18] to leverage image-level labels (indicating the presence or absence of lesions, without spatial localisation) which are faster to annotate than point annotations. Both PCL and WS-DETR demonstrated significantly lower performance compared to our proposed WSSOD pipeline. Specifically, the average precision at 50% (AP@50) was less than 1% for PCL and 6% for WS-DETR with an annotation budget corresponding to the entire training set labeled at the image level (195 min). We find that, due to the unclear boundaries of lesion areas, the region proposals are inaccurate, which results in low AP.

In Fig. 5, we showcase qualitative results of our study. The left column displays the ground truth, featuring both point and corresponding bounding box annotations for lesions. The middle column depicts the predicted pseudo-labels with S_{OTF}. The right column displays predictions from the fully-supervised model. The first two rows display instances where accurate pseudo-labels were generated. However, the last row reveals a case where the model failed, incorrectly interpreting the ultrasound artifact, known as acoustic shadowing, as the periphery of the lesion. As mentioned, the OTF point is consistently localized on the near-center of the lesions.

Our method primarily focuses on annotation, making it inherently adaptable and compatible with various optimization strategies, such as self-supervised learning and active learning. Active learning streamlines the training of models by strategically selecting a subset of unlabeled data. This method focuses on choosing samples that, once annotated, contribute most effectively to the model's performance. This iterative process of model improvement aims to achieve high accuracy with fewer labeled instances, which is valuable when data annotation is expensive or time-consuming. [26, 27] focus on identifying and annotating the most uncertain or challenging samples, thereby optimizing the learning process and efficiently utilizing the annotation budget. Those approaches, which emphasizes learning from complex cases, can be integrated with on-the-fly annotation strategies to allows for effective decision-making about which videos to annotate in-depth (box-annotation) and which to annotate weakly, ensuring a targeted and resource-efficient learning process.

Another interesting application of our approach is to leverage a strong pretrained foundation model such as Segment Anything (SAM)[9], which acts as a class-agnostic segmentation model and exploits prompts such as points for improved accuracy. Our on-the-fly annotation can be used to prompt these models without the need to train a separate teacher model. However, at this time, the capacity of SAM is limited on ultrasound data due to the domain gap with its training database, yielding noncompetitive results on the STARHE dataset.

Conclusion

In this paper, we introduce an on-the-fly point annotation pipeline, enabling live video annotation to mitigate the tedious efforts specifically associated with video annotation. Every frame in the video is weakly-annotated, ensuring expert guidance throughout the process. Our method proves

to allow precise tracking of structures, which can enable useful pseudo-labels generation compatible with weakly semi-supervised object detection pipelines, outperforming conventional annotation method at equivalent annotation budgets. A notable reduction in annotation costs is observed through the utilization of this strategy. The findings of this study underscore the need for optimization of video annotation processes, enabling the development of high-quality datasets. This approach fosters a more efficient utilization of expert resources, optimizing the balance between annotation accuracy and cost-effectiveness in medical imaging studies.

Acknowledgements This research was conducted within the framework of the APEUS and TheraHCC 2.0 projects, which are supported by the ARC Foundation (www.fondation-arc.org). This work was also partially supported by French state funds managed within the 'Plan Investissements d'Avenir', funded by the ANR (reference ANR-10-IAHU-02 and ANR-21-RHUS-0001 DELIVER). This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011013698R1).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards (ID-RCB 2020-A02949-30; NCT04802954).

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Buch VH, Ahmed I, Maruthappu M (2018) Artificial intelligence in medicine: current trends and future possibilities. Br J Gen Pract 68(668):143–144
2. Mascagni P, Alapatt D, Sestini L, Altieri MS, Madani A, Watanabe Y, Alseidi A, Redan JA, Alfieri S, Costamagna G et al (2022) Computer vision in surgery: from potential to clinical value. npj Digital Med 5(1):163
3. Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, Ni L, Shum H (2022) Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arxiv 2022. arXiv preprint arXiv:2203.03605
4. Lyu C, Zhang W, Huang H, Zhou Y, Wang Y, Liu Y, Zhang S, Chen K (2022) RTMDet: an empirical study of designing real-time object detectors
5. Barua I, Vinsard DG, Jodal HC, Løberg M, Kalager M, Holme Ø, Misawa M, Bretthauer M, Mori Y (2020) Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis. Endoscopy 53(03):277–284
6. Twinanda AP, Shehata S, Mutter D, Marescaux J, De Mathelin M, Padoy N (2016) Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging 36(1):86–97
7. Srivastav V, Issenhuth T, Kadkhodamohammadi A, Mathelin M, Gangi A, Padoy N (2018) Mvor: a multi-view rgb-d operating room dataset for 2d and 3d human pose estimation. arXiv preprint arXiv:1808.08180
8. Krenzer A, Makowski K, Hekalo A, Fitting D, Troya J, Zoller WG, Hann A, Puppe F (2022) Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists. Biomed Eng Online 21(1):1–23
9. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y, et al (2023) Segment anything. arXiv preprint arXiv:2304.02643
10. Chen L, Yang T, Zhang X, Zhang W, Sun J (2021) Points as queries: weakly semi-supervised object detection by points. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8823–8832
11. Zhang S, Yu Z, Liu L, Wang X, Zhou A, Chen K (2022) Group r-cnn for weakly semi-supervised object detection with points. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9417–9426
12. Rother A, Niemann U, Hielscher T, Völzke H, Ittermann T, Spiliopoulou M (2021) Assessing the difficulty of annotating medical data in crowd working with help of experiments. PLoS ONE 16(7):0254764
13. Heim E, Roß T, Seitel A, März K, Stieltjes B, Eisenmann M, Lebert J, Metzger J, Sommer G, Sauter AW et al (2018) Large-scale medical image annotation with crowd-powered algorithms. J Med Imaging 5(3):034002–034002
14. Jeong J, Lee S, Kim J, Kwak N (2019) Consistency-based semi-supervised learning for object detection. Adv Neural Inf Process Syst, 32
15. Liu Y-C, Ma C-Y, He Z, Kuo C-W, Chen K, Zhang P, Wu B, Kira Z, Vajda P (2021) Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480
16. Wang Z, Li Y, Guo Y, Fang L, Wang S (2021) Data-uncertainty guided multi-phase learning for semi-supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4568–4577
17. LaBonte T, Song Y, Wang X, Vineet V, Joshi N (2023) Scaling novel object detection with weakly supervised detection transformers. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 85–96
18. Tang P, Wang X, Bai S, Shen W, Bai X, Liu W, Yuille A (2018) PCL: proposal cluster learning for weakly supervised object detection. IEEE Trans Pattern Anal Mach Intell 42(1):176–191
19. Vardazaryan A, Mutter D, Marescaux J, Padoy N (2018) Weakly-supervised learning for tool localization in laparoscopic videos. In: Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis: 7th joint international workshop, CVII-STENT 2018 and third international workshop, LABELS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, pp 169–179. Springer
20. Kim J, Kim HJ, Kim C, Lee JH, Kim KW, Park YM, Kim HW, Ki SY, Kim YM, Kim WH (2021) Weakly-supervised deep learning for ultrasound diagnosis of breast cancer. Sci Rep 11(1):24382
21. Ouyang J, Chen L, Li GY, Balaraju N, Patil S, Mehanian C, Kulhare S, Millin R, Gregory KW, Gregory CR et al (2023) Weakly semi-supervised detection in lung ultrasound videos. In: International conference on information processing in medical imaging, pp 195–207. Springer
22. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision, pp 213–229. Springer
23. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28
24. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: Computer vision–ECCV 2014: 13th European Con-

ference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, pp 740–755. Springer
25. Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, Sun S, Feng W, Liu Z, Xu J et al (2019) Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155
26. Kim DD, Chandra RS, Peng J, Wu J, Feng X, Atalay M, Bettegowda C, Jones C, Sair H, Liao W-h et al (2023) Active learning in brain tumor segmentation with uncertainty sampling, annotation redundancy restriction, and data initialization. arXiv preprint arXiv:2302.10185
27. Wang K, Zhang D, Li Y, Zhang R, Lin L (2016) Cost-effective active learning for deep image classification. IEEE Trans Circuits Syst Video Technol 27(12):2591–2600

DETECTION OF EARLY-STAGE HEPATOCELLULAR CARCINOMA IN HIGH-RISK PATIENTS WITH A DEEP LEARNING APPROACH USING MRI

Material and methods

The methodology of the MRI study has been detailed in the previous section and only relevant specific methodological information are provided in this section.

Population

To maximise the number of cases to train the detection model, patients with early-stage hepatocellular carcinoma developed on non-controlled or non-healed B/C viral F3/F4 hepatitis were included for the detection task.

Index test

We aimed to develop an object detection deep learning model based on non-contrast aMRI (T1-weighted in/out-of-phase images, T2-weighted with fat suppression images, diffusion-weighted imaging, using the same acquisition parameters as a state-of-the-art MRI to maintain the same contrast, spatial resolution, and signal-to-noise ratio) to improve early-stage hepatocellular carcinoma detection.

Outcome

The main outcome of the study was the diagnostic performances of the AI model for the detection of early-stage hepatocellular carcinoma (1 nodule of any size or ≤ 3 nodules, each < 3 cm in diameter, without gross vascular invasion or extrahepatic metastases).

AI methodology

The AI system will follow the same stages detailed in the Objective 1 section (risk stratification):

- **Annotation and labelling of databases** (identical).
- **Liver segmentation** (identical).
- **Pre-processing of MRI images** (identical).
- **Early hepatocellular carcinoma detection**. A 3D U-Net model, a convolutional neural network architecture renowned for its proficiency in medical image segmentation tasks, will be trained for the detection task in the clinical perspective that in a screening setting, it values more to detect a lesion than accurately delineate its contour. We will use the same liver volumes, liver segmentation mask and hepatocellular carcinoma bounding boxes. This task requires the network to effectively capture both the global context and fine-grained details of the liver's anatomical structure.
- **External independent testing**: To ensure the robustness and generalisability of our AI models, we will test the model on the same independent internal dataset used for the classification model (risk stratification of hepatocarcinogenesis) and then, on the same two external datasets (CHUM and FASTRAK).

Statistical analysis

The estimate of performance metrics of the detection model was computed for mean average precision (area under the precision-recall curve) with a predefined intersection over union (IoU) of 10, 50 and 75%. Confusion matrices were computed for each patient at different confidence levels for an intersection over the union of 10% to assess the rate of true positive and false positive. A 10% threshold was chosen because of the screening strategy where detecting a lesion is more important than correctly delineating it.

Results

Population

This study enrolled 368 patients between November 2011 and September 2023 (**Figure 9**). A total of 35 patients were excluded: 25 had extensive artifacts on the MRI and 10 had the MRI at the time of an acute liver event (e.g., acute portal vein thrombosis). An additional total of 41 patients were excluded from the analysis of hepatocellular carcinoma risk stratification: 2 had non-controlled HBV, 38 had non-cured HCV and 1 had no information regarding the aetiology of the liver disease. The demographics description of the dataset is reported in **Table 9**.

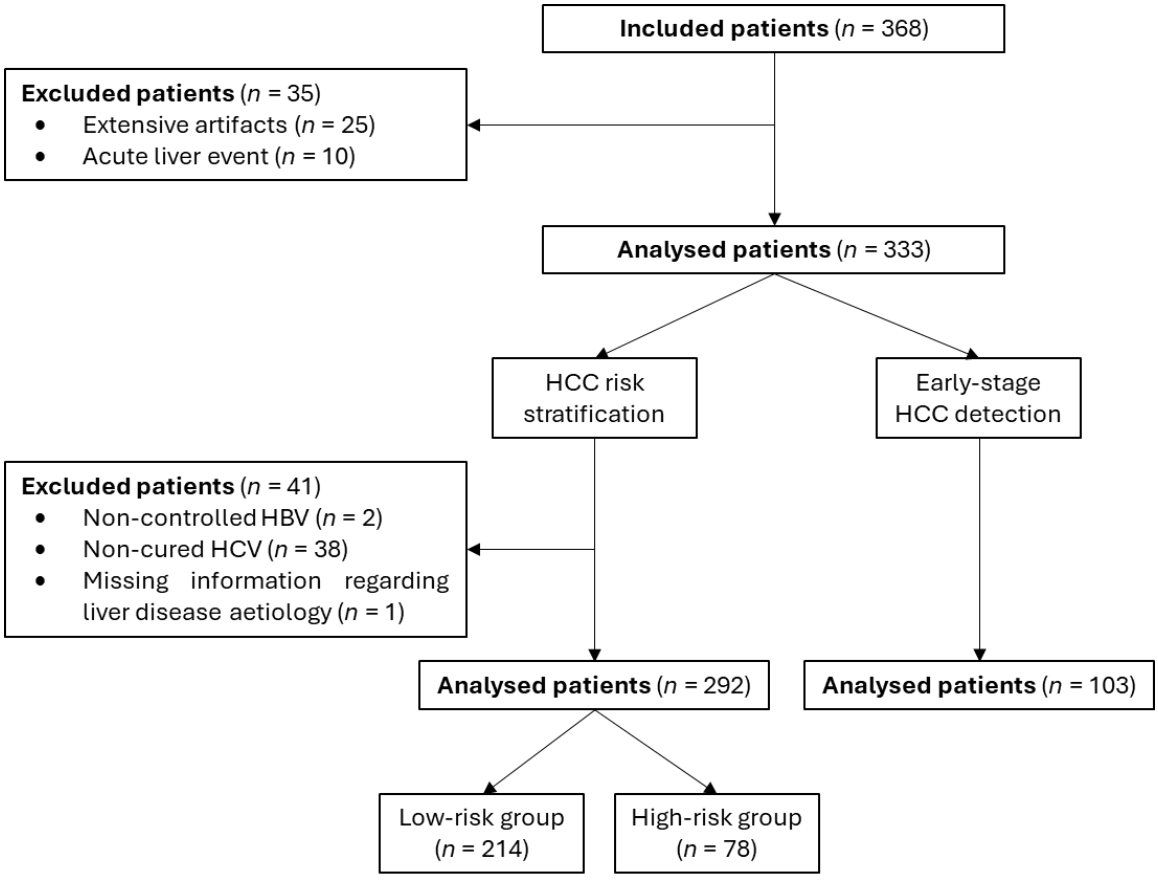


Figure 9 – Flow chart of the study

	No HCC (*n* =214)	Early-stage HCC (*n* = 103)
Centres		
1 (Royal Victoria Hospital)	122 (57%)	68 (66%)
2 (Montreal General Hospital)	76 (36%)	22 (21%)
3 (Lachine Hospital)	16 (7%)	13 (13%)
MRI manufacturer		
General Electrics Optima MR450w (1.5T)	59 (28%)	28 (27%)
General Electrics SIGNA Excite (1.5T)	104 (49%)	36 (35%)
General Electrics SIGNA Artist (1.5T)	31 (14%)	18 (17%)
Siemens AERA (1.5T)	16 (7%)	13 (13%)
Siemens SKYRA (3.0T)	4 (2%)	8 (8%)
Age	62 [52-70]	65 [58-74]
Sex		
Male	127 (59%)	77 (75%)
Female	87 (41%)	36 (25%)
Chronic Liver Disease		
Aetiology of liver disease		
ALD	30 (14%)	12 (12%)
MASLD	53 (25%)	24 (23%)
MetALD	9 (4%)	9 (9%)
HBV	26 (12%)	10 (10%)
HCV	15 (7%)	40 (39%)
ALD + HBV	4 (2%)	1 (1%)
ALD + HCV	2 (1%)	5 (5%)
MASLD + HBV	2 (1%)	0
HBV + HCV	2 (1%)	1 (1%)
PSC	18 (8%)	0
Auto-immune and overlap syndrome	16 (7%)	0
Indeterminate	10 (5%)	1 (1%)
Other	27 (13%)	0
FASTRAK score	7 [6-10]	10 [8-12]
Child-Pugh		
A5	131 (61%)	42 (41%)
A6	32 (15%)	28 (27%)
B7	25 (12%)	18 (17%)
B8	8 (4%)	8 (8%)
B9	2 (1%)	2 (2%)
C10	0	1 (1%)
Missing data	16 (7%)	4 (4%)
Type 2 diabetes	70 (33%)	50 (49%)
Biology		
Alpha-foetoprotein (ng/mL)	4.0 [2.7-6.0]	5.9 [3.3-15.1]
GGT (UI/L)	65.5 [33.3-140.8]	78 [44-133]
Total Bilirubin (µmol/L)	16.8 [12.2-24.3]	20.8 [12.6-31.4]
Platelet (G/L)	134 [91-177]	105 [72-139]
INR	1.1 [1.0-1.2]	1.1 [1.0-1.3]
Albumin (g/L)	39 [36-42]	37 [33-42]
Hepatocellular carcinoma (at inclusion)		
Number of nodules		
1	NA	74 (72%)
2		21 (20%)
3		8 (8%)
Size (mm)	NA	18.0 [12.8-25.3]
BCLC stage		
0	NA	40 (39%)
A		63 (61%)

Table 9 – Demographics

Detection of early-stage (BCLC 0 and A) hepatocellular carcinoma

The development is in progress and the results are not available yet. We are currently working on the risk stratification model and then we'll work on the detection model

Investigate innovative techniques to characterise chronic liver disease

HIGH-RESOLUTION 7T MRI FOR A PATHOLOGY-LIKE EXAMINATION OF LIVER FIBROSIS

Accurate prediction of the progression of early-stage chronic liver disease to cirrhosis-related complications is a critical unmet need. Although several imaging-based quantitative biomarkers have emerged, characterisation of chronic liver disease is still limited and histopathology remains the gold standard. Fibrosis can be non-invasively assessed by ultrasound- or MRI-based elastography techniques. Approaches have been developed to estimate steatosis, exploiting the attenuation of ultrasonic waves or employing advanced MRI acquisition techniques (e.g., multi-echo DIXON, spectroscopy). Several bio-clinical scoring systems based on routine parameters and liver elastography have proven valuable in predicting the first liver-related event and overall survival in patients with cirrhosis. Yet, major improvements in the field of image acquisition and management/analysis of imaging data are required to be able to accurately characterise liver parenchyma, monitor its changes and predict any pejorative evolution. As previously mentioned, the structural analysis of the liver parenchyma has been shown to reflect the pathophysiological mechanisms responsible for hepatocarcinogenesis. In the 1990s, ultrasound studies examined the incidence of hepatocellular carcinoma according to the liver echostructure[40-42]. Results showed that a nodular heterogeneous echostructure resulted in a relative risk estimate of up to 20. Therefore, if the liver parenchyma can be more accurately characterised with higher spatial resolution at a microscopic level, the liver architecture could be better appreciated which could lead to a shift in paradigm in the monitoring and therapeutic management of patients with chronic liver disease.

With recent improvements in micro-imaging techniques, high-resolution Magnetic Resonance Imaging (MRI) (high-field MR imaging) recently emerged as a promising tool to image fresh ex vivo tissue and provide histopathology-like examination, with a spatial resolution approximating that of histology (i.e., < 100 µm, compared to ~1 mm for clinical MRI). High-resolution MRI appears as a promising “missing link” between conventional imaging and histology. Additionally, unlike histology, which is limited to 2D images, MRI provides images of

the entire tissue sample as a volume. This volumetric analysis is also an advantage over traditional histology, particularly relevant for the study of fibrosis-related structural distortions in tissues. Liver fibrosis appeared as an appropriate model, because of the existence of widely validated histological classifications of liver fibrosis, allowing both qualitative and quantitative analysis, to demonstrate that MRI could be disruptively used to provide a histopathology-like examination. Furthermore, in chronic liver disease, liver fibrosis is easily recognized on high-resolution MR images as T2 hyperintense tracts and is associated with micro- and macro-architectural changes in the liver parenchyma. Characterized by the deposition of extracellular matrix proteins, including collagen, liver fibrosis progresses from fibrous portal expansion to bridging fibrosis, and finally to cirrhosis.

We aimed to investigate the capabilities of high-resolution MRI to provide a histopathology-like examination of ex vivo liver tissues. This would be the first step to pave the way for future research developments to bridge the gap with in-vivo MRI to impact clinical practice and provide true non-invasive microscopic histologic examination.

Material and methods

This prospective project was approved by the Research Ethics Board (Protocol RIPH2, LivMod N°IDRCB 2019-A00738-49 ClinicalTrial NCT04690972) and followed ethical principles of the Declaration of Helsinki. All patients provided written informed consent.

Population

Twenty patients aged > 18 years old, who underwent surgical liver resection between November 2021 and April 2023, were prospectively included. We used ex-vivo fresh liver tissue (~ 1 cm3) from surgically resected livers. Each fragment was sectioned in half, and the sections were identified, so that the MRI acquisition plane was as close as possible to the histological one. We imaged the first half, placed in Fluorinert™ Electronic Liquid FC40 (Sigma-Aldrich), using a 7T

MRI with a cryoprobe (Bruker BioSpin; Fat-Suppressed Turbo Spin Echo (Rapid Imaging with Refocused Echoes, RARE) T2-weighted sequence; echo time = 30 ms; repetition time = 3000 ms; averages = 25; Rare Factor = 2; matrix = 266x200x42, yielding a slice thickness of 200 µm; field-of-view = 20x15x6mm), allowing a spatial resolution of 75 x 75 x 200 µm, in a 2-hour acquisition time. The second half was fixed in formalin, embedded in paraffin, cut at a 4 µm thickness in the same plane as the MRI acquisition, and then stained using Masson's Trichrome and Perls. The minimum processing time for pathology samples was around ~ 48h.

Assessment of MRI and pathology images

Three subspecialty-trained abdominal radiologists (Benoit Gallix with 27 years of experience, Valérie Vilgrain with 35 years of experience, and Aïna Venkatasamy with 7 years of experience) and three subspecialty-trained abdominal pathologists (Antonin Fattori with 5 years of experience, Valérie Paradis with 30 years of experience, and Aurélie Beaufrère with 10 years of experience) reviewed, independently and blinded to medical records, MR, and pathology images. The readings were carried out in 2 separate sessions, over a month apart.

Pathology images

In the first session, the 3 pathologists independently staged fibrosis on histological slides in accordance with the METAVIR score. The consensus fibrosis stage for pathology was based on the agreement of at least 2 of the 3 pathologists.

MR images

The second session consisted of the reading of MR images, the cases being presented in a different random order, by all 3 radiologists and 3 pathologists, given that this new field of application is halfway between the two specialities. First, all readers (radiologists and pathologists) received training on the METAVIR classification of fibrosis and the semiology of the fibrosis on high-resolution MRI (MR-derived METAVIR score as described below), using two

previously published cases [2] and schematic illustrations of high-resolution MRI for each stage (F0 to F4, **Figure 10**). Then, all reviewers were asked to review the MR images (6 images per patient) and stage fibrosis, according to the MR-derived METAVIR score. The consensus fibrosis stage for MRI was based on the agreement of at least 2 of the 3 radiologists. To create the MR-derived METAVIR score, one subspecialty-trained abdominal radiologist (JD, who did not participate in the reading), reviewed all MR images and analysed the MRI features of fibrosis in correlation with their corresponding histological section. The MR-derived METAVIR score has been developed in analogy with the histological staging criteria, based on the presence and distribution of fibrosis [2].

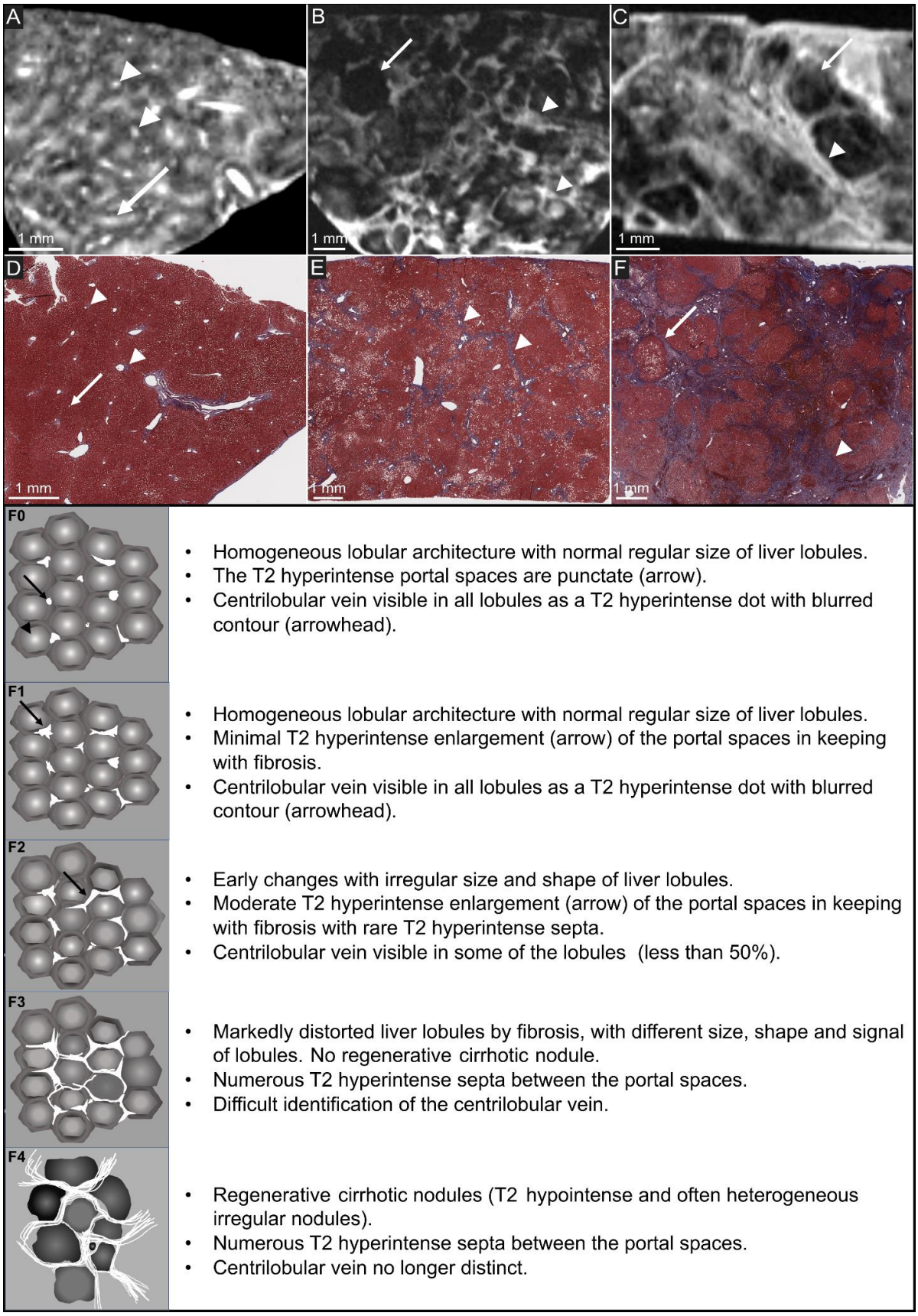


Figure 10 – Schematic illustrations of the MR-derived METAVIR score with High-resolution liver MRI and histopathology correlation (Masson's Trichrome) in a normal liver (A and D), F3 liver (B and E), and F4 (cirrhosis) liver (C and F).

Statistical analysis

The inter-reader agreement was calculated using Fleiss Kappa[56]. The agreement between the pathologic gold standard and the MRI consensus was calculated using weighted Kappa (linear weights). The agreement was interpreted according to Kappa value as follows: < 0 (poor); 0–0.2 (slight); 0.21–0.40 (fair); 0.41–0.60 (moderate); 0.61–0.80 (substantial); 0.81–1.00 (almost perfect)[57]. All statistical analysis was performed using IBM SPSS Statistics (Version 29).

Results

Population

A total of 20 patients scheduled for liver surgery were included in the study. Three patients were excluded due to poor specimen quality, not allowing MRI or pathology examination (n = 1) and cancellation of the surgery (n = 2). In addition, 2 of the 20 included patients, previously published [2] were used for reader training, and not included in the final dataset. The final dataset consisted of 15 patients (median age 68 [60-73], 12 men and 3 women) who underwent liver surgery for primary liver lesions (n=10) or colorectal cancer liver metastases (n=5). Chronic liver disease of several aetiologies was present in 10/15 patients of the final cohort (67%) (**Table 10**).

The pathologists' consensus results for fibrosis staging on Masson's Trichrome stained slides, which served as the gold standard, were as follows: 2/15 patients (13%) had a METAVIR score of F0, 5 (33%) had a score of F1, 5/15 (33%) had a score of F3, and 3/15 (20%) had a score of F4 (Table S1). In all cases, no consensus review was necessary and the agreement between the pathologists regarding the review of histopathology images was very good (κ = 0.77 [95 CI 0.61-0.93).

Age	Sex	Chronic Liver Disease	Surgical indication	Pathological Fibrosis stage	MRI Fibrosis stage
59	M	Hepatitis B and C viruses	Hepatocellular carcinoma	F2	Training
49	F	None	Metastases	F0	Training
73	M	Metabolic Dysfunction Associated Steatotic Liver Disease (MASLD)	Hepatocellular carcinoma	F1	F0
57	M	Indeterminate aetiology	Metastases	F3	F2
73	M	None	Metastases	F0	F0
64	M	MetALD (MASLD and increased alcohol intake)	Hepatocellular carcinoma	F3	F3
58	M	None	Biliary cystadenoma	F1	F1
65	M	Alcohol Associated Liver Disease	Hepatocellular carcinoma	F4	F4
73	M	MetALD (MASLD and increased alcohol intake)	Hepatocellular carcinoma	F1	F1
72	F	Alcohol Associated Liver Disease	Hepatocellular carcinoma	F3	F4
71	M	None	Metastases	F1	F0
62	M	Alcohol Associated Liver Disease	Hepatocellular carcinoma	F4	F4
56	F	None	Metastases	F0	F0
68	M	Indeterminate aetiology	Hepatocellular carcinoma	F3	F3
60	F	None	Metastases	F1	F1
68	M	Alcohol Associated Liver Disease	Hepatocellular carcinoma	F3	F4
80	M	Metabolic Dysfunction Associated Steatotic Liver Disease (MASLD)	Hepatocellular carcinoma	F4	F4

Table 10 – Demographics, clinical characteristics, surgical indications, and results of consensus review of pathology slides and 7T MRI fibrosis staging.

Accuracy of 7T MRI

The accuracy of MRI for a histopathology-like diagnosis of the absence (F0) or very early (F1) fibrosis, and the presence of advanced fibrosis (F3-F4), was excellent (0.93 95CI [68-100] –

Table 11). MR imaging correctly classified almost all patients with advanced fibrosis on pathology (*n* = 7/8, one patient was F3 on histologic examination and staged F2 on MRI). Conversely, MRI correctly excluded fibrosis (i.e., F0 or F1 stages) in all cases (*n* = 7) compared to histology. The concordance of MR image analysis compared to histopathology was excellent (κ = 0.81 95CI [0.67-0.95]) for all fibrosis stages, with an accuracy of 93%.

	Inter-reader agreement	Agreement with pathologic gold standard		
		n	Accuracy	Weighted Kappa
Radiologists				
No or very-early fibrosis (F0-F1)	κ = 0.46 [0.17-0.76]	*n* = 7/7	93% [68-100]	κ = 0.81 [0.67-0.95]
Advanced fibrosis (F3-F4)	κ = 0.42 [0.29-0.55]	*n* = 7/8		
Pathologists				
No or very-early fibrosis (F0-F1)	κ = 0.64 [0.35-0.93]	*n* = 6/7	87% [60-98]	κ = 0.50 [0.26-0.73]
Advanced fibrosis (F3-F4)	κ =0.82 [0.53-1.0]	*n* = 7/8		

Table 11 – Accuracy of MRI to identify and stage hepatic fibrosis and inter-reader agreement (Inter-reader agreement was calculated using Fleiss Kappa; [95% CI are in squared brackets])

Conclusion

High-resolution 7T MRI provides an assessment similar to low-magnification histology and its accuracy was excellent to grade liver fibrosis (93%). Beyond fibrosis staging of liver tissue, MRI enables a cross-sectional volumetric exploration of the entire specimen, without cutting or destroying the sample. With short-time image acquisition and immediate image reading, high-resolution MRI could become a new modality for extemporaneous tissue analysis, especially in oncologic surgery. Future research developments should focus on bridging the gap with in-vivo MRI to impact clinical practice and provide true non-invasive microscopic histologic examination.

The preliminary aspect of this work was published in Radiology[58]. The larger study is under submission.

Radiology

High-Resolution (7-T) Liver MRI for Pathologic Examination

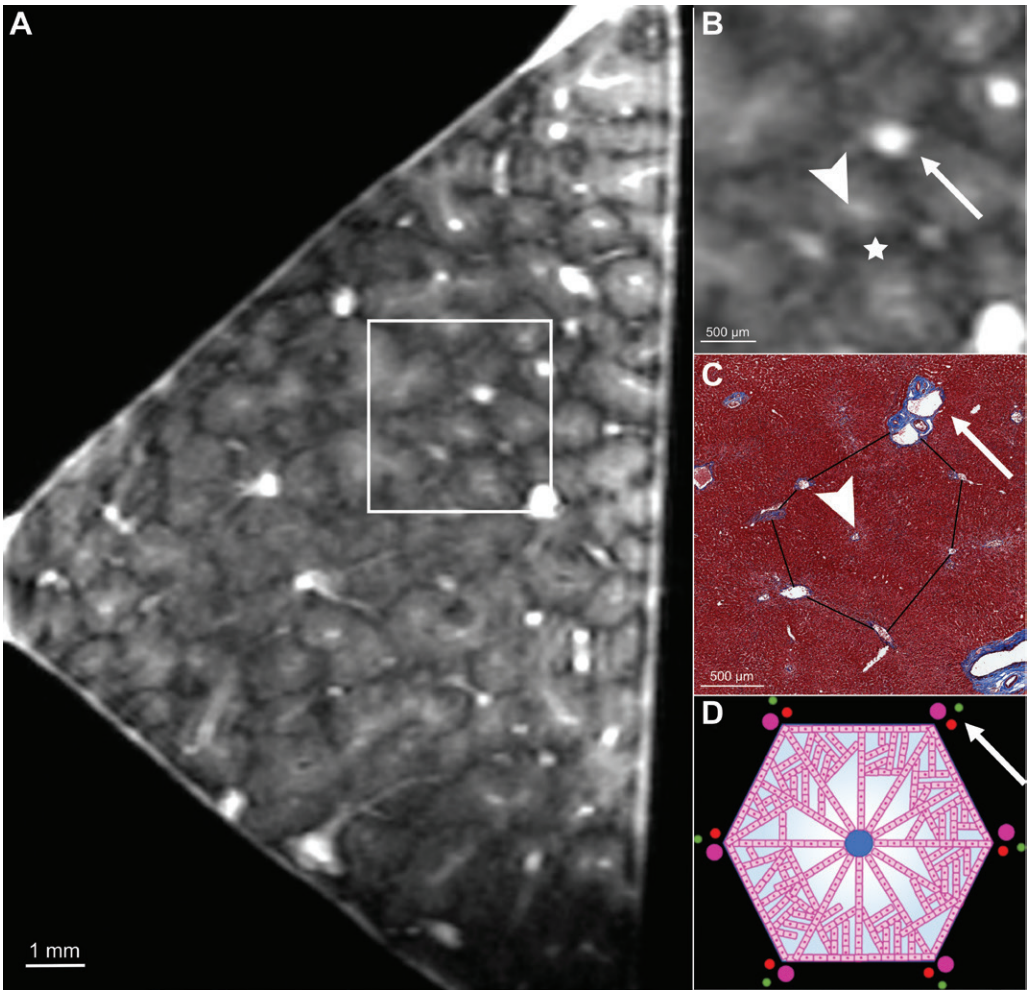
Jérémy Dana, MD • *Aïna Venkatasamy, MD, PhD*

From the Institut de Recherche sur les Maladies Virales et Hépatiques, Université de Strasbourg, Inserm, U1110, 3 Rue Koeberlé, 67000 Strasbourg, France (J.D.); Institut Hospitalo-Universitaire, Strasbourg, France (J.D., A.V.); Department of Diagnostic Radiology, McGill University Health Centre, Montreal, Canada (J.D.); Streinth Laboratory (Stress Response and Innovative Therapies), Inserm UMR_S 1113 IRFAC, Interface Recherche Fondamentale et Appliquée à la Cancérologie, Strasbourg, France (A.V.); and Department of Radiology–Medical Physics, University Hospital Freiburg, Freiburg, Germany (A.V.). Received February 23, 2022; revision requested March 22; revision received April 9; accepted April 25. **Address correspondence to** J.D. (email: *jeremy.dana@etu.unistra.fr*).

Conflicts of interest are listed at the end of this article.

Online supplemental material is available for this article.

Radiology 2023; 306:74–75 • **https://doi.org/10.1148/radiol.220410** •



Images show radiologic-pathologic correlation of normal liver. **(A)** Fat-suppressed T2-weighted MRI scan, with a spatial resolution of 75 × 75 × 200 μm, shows the homogeneous architecture of normal liver. **(B)** MRI scan (magnified image) depicts the anatomy of the primary liver lobule with its centrilobular vein (arrowhead) and peripheral portal venous tracts (arrow) within the interlobular delineations (star). **(C)** Mirrored pathologic slide, with Masson trichrome staining at four times magnification, identifies the liver lobule with the centrilobular vein (arrowhead) and peripheral portal tracts (arrow) within the interlobular delineations (black lines). **(D)** Schematic representation of the primary liver lobule with the centrilobular vein (central blue dot) and peripheral portal venous tracts (arrow).

Two different liver samples from patients who underwent surgical resection were imaged with a 7-T MRI scanner by using fat-suppressed fast spin-echo T2-weighted sequence, reaching spatial resolution of 75 × 75 × 200 μm in 2 hours.

An MRI scan clearly demonstrated the homogeneous architecture of the normal liver (Figure) and identified the primary liver lobule, with its centrilobular vein and peripheral portal tracts (Figure, B and C) within the interlobular delineations (Figure, B). In stage 2 fibrotic liver (Fig E1 [online]), an MRI scan of the liver sample demonstrated correlation with pathologic findings, easily identifying the same fibrous portal bridges (Fig E1A and E1B [online]) and enlarged fibrous portal tracts (Fig E1A and E1B [online]), thus enabling a similar grading of fibrosis. With a spatial resolution close to that of the pathologic examination, MRI enabled an easier depiction of the liver lobule

boundaries, allowing early detection of fibrosis-related architectural distortion. Additionally, the specimen remained intact, as no specific preparation was required for the MRI scan, and could undergo standard pathologic processing after image acquisition.

Acknowledgments: We thank Chrystelle Po, PhD (ICube UMR 7357, Université de Strasbourg/CNRS, Fédération de Médecine Translationnelle de Strasbourg, France), Antonin Fattori, MD (Département de Pathologie, Hôpitaux Universitaires de Strasbourg, France), Catherine Schuster, PhD (Université de Strasbourg, Inserm, Institut de Recherche sur les Maladies Virales et Hépatiques UMR_S1110, Strasbourg, France), Thomas F. Baumert, MD, PhD (Université de Strasbourg, Inserm, Institut de Recherche sur les Maladies Virales et Hépatiques UMR_S1110, Strasbourg, France), Benoit Gallix, MD, PhD (Institut Hospitalo-Universitaire, Strasbourg, France), and Pr Patrick Pessaux, MD, PhD (Département de Chirurgie Viscérale et Digestive, Pôle Hépato-digestif, Nouvel Hôpital Civil, Hôpitaux Universitaires de Strasbourg, Strasbourg, France) for their crucial contribution. Supported by French state funds managed within the "Plan Investissements d'Avenir" and by the ANR (ANR-10-IAHU-02 to B.G.), the ARC, Paris and Institut Hospitalo-Universitaire, Strasbourg (TheraHCC 2.0 and IHUARC2019 to T.F.B.), the European Union (ERC-AdG-2014-671231-HEPCIR and ERC-AdG-2020-667273-FIBCAN to T.F.B.), ANRS, Paris (ECTZ160436 and ECTZ103701 to T.F.B), the Foundation of the University of Strasbourg (HEPKIN to T.F.B.), and the Institut Universitaire de France (T.F.B.). This work has been published under the framework of the LABEX ANR-10-LABX-0028_HEPSYS and Inserm Plan Cancer.

Disclosures of conflicts of interest: **J.D.** Research grant from the French Society of Radiology. **A.V.** Support from Bracco to attend RSNA 2021; faculty member for ESHNR.

Discussion and Perspectives

Chronic liver diseases, resulting from chronic injuries of various causes, lead to cirrhosis with life-threatening complications including liver failure, portal hypertension, and hepatocellular carcinoma. A key unmet medical need is robust non-invasive biomarkers to predict patient outcomes, stratify patients for risk of disease progression and monitor response to emerging therapies[45]. Furthermore, the identification of patients more prone to a pejorative outcome (e.g., liver failure, portal hypertension or hepatocellular carcinoma) is crucial to allow the early onset of regular monitoring and chemoprevention. If no approved therapy exists to treat liver fibrosis and prevent the progression of early-stage chronic liver disease, the discovery of Claudin-1, a tight junction protein expressed in hepatocytes, as a therapeutic target for liver fibrosis and hepatocellular carcinoma may revolutionise therapeutic management. Novel specific drugs including the claudin-1-specific monoclonal antibody will require a stratification of patients, i.e. companion biomarker, to achieve the best clinical outcome/cost ratio.

The transition towards non-invasive characterisation and longitudinal follow-up of chronic liver diseases is ongoing. To date, reproducible quantitative imaging biomarkers are available to assess liver fibrosis with liver stiffness measured by elastography or steatosis with Proton Density Fat Fraction on Magnetic Resonance Imaging. Nevertheless, if fibrosis and steatosis appear as decisive markers for the histopathologic characterization of chronic liver disease, they fail to accurately predict the progression of early-stage chronic liver disease to cirrhosis-related complications. Major improvements, in the field of image acquisition and analysis, are still required to be able to accurately characterise the liver parenchyma, monitor its changes and predict any pejorative evolution across disease progression. Artificial Intelligence has the potential to augment the exploitation of massive multi-parametric data to extract valuable information and achieve precision medicine[44]. Machine learning algorithms have been developed to assess non-invasively certain histological characteristics of chronic liver diseases, including fibrosis and steatosis[59-64]. Although still at an early stage of development, AI-based imaging biomarkers provide novel opportunities to predict the risk of progression from early-stage chronic liver disease towards cirrhosis-related complications, with the ultimate perspective of

precision medicine. AI could also help maximise diagnostic performances of ultrasound and automatise time-consuming tasks such as measurement of the liver volume using deep CNN[65,66], a simple prognostic biomarker of the pejorative outcome of acute liver failure[67]. These quantitative imaging techniques, either based on conventional imaging or using artificial intelligence approaches, can provide reproducible and reliable quantitative information. They are not exclusive from each other and can be complementary to biochemical biomarkers. But to become clinical tools, AI models should be developed following a high standard process to achieve generalizability and transferability including training on large datasets representing the wide spectrum of the disease expression to avoid selection biases, and independent and prospective testing to avoid overfitting[68]. Such databases also come with multiple challenges such as the important resources that they require in terms of annotation/labelling time by experts. In this thesis, we have investigated a new innovative approach to shorten the annotation time of imaging videos[69].

The absence of personalisation of hepatocellular carcinoma screening programs is an undeniable example of the lack of prediction biomarkers in clinical practice. Currently, in France, screening programs still rely on biannual ultrasound with poor performances for early-stage hepatocellular carcinoma detection. To overcome the weaknesses of ultrasound, the use of aMRI has been proposed to improve hepatocellular carcinoma screening because it offers higher performance than ultrasound while minimising acquisition time compared to a conventional MRI protocol. The reported sensitivity and specificity of the different protocols (NC-aMRI, DCE-MRI and HB-MRI) ranged from 84.6 to 96% and from 81.6 to 100%[18,21-26]. Although the diagnostic performance of MRI is superior to that of ultrasound, MRI is an expensive and not easily accessible examination. Recent analyses of prospective European cohorts including a model-based evaluation of very early-stage hepatocellular carcinoma detection confirmed that MRI surveillance is cost-effective for a baseline yearly incidence of 3% in patients with cirrhosis without active viral replication[17]. Therefore, screening with aMRI can only be considered for a sub-population with a very high risk of hepatocarcinogenesis, which would be selected from the

population currently undergoing standard ultrasound screening. Identifying this subset of high-risk patients is crucial as this strategy would detect 5 times more very early-stage hepatocellular carcinoma than ultrasound, with an ICER below 30,000€/life-years gained[27].

Risk stratification of hepatocarcinogenesis is therefore needed, for patients to benefit from a personalised screening strategy. Defining such a population involves developing tools for stratifying the risk of hepatocarcinogenesis[28]. Preliminary models have been developed, either aetiology-specific[29-31] or multi-aetiology[17], incorporating clinical parameters (e.g., age, sex, body mass index, or diabetes) and biological parameters (e.g., GGT, AST/ALT, platelets, or albumin)[30,32,33]. These models demonstrated good discriminative performances and have the advantages of being easy to use and inexpensive. Serum protein-based[36-38] and genetic approaches[39] have also been investigated. However, these models do not take into consideration the structural analysis of the liver parenchyma, which reflects the pathophysiological mechanisms responsible for hepatocarcinogenesis. In the 1990s, ultrasound studies examined the incidence of hepatocellular carcinoma according to the liver echostructure[40-42]. Results showed that a nodular heterogeneous echostructure resulted in a relative risk estimate of up to 20.

In the STARHE study, a prospective multicentric study, we have developed an AI-based risk stratification model of hepatocarcinogenesis on ultrasound. This model could predict patients at high risk of developing a hepatocellular carcinoma with an odds ratio of 6.6. Although comparison with longitudinal studies is limited, an odds ratio of 6.6 is extremely promising for future risk-based personalised screening strategies. This study demonstrated that risk stratification of hepatocarcinogenesis can be achieved based on the deep learning analysis of ultrasound images of the liver parenchyma. This supports our hypothesis that non-tumour cirrhotic liver parenchyma is rich in structural information reflecting the severity of liver disease, its carcinogenic risk as well as the process of hepatocarcinogenesis. This study paves the way for a personalised screening program based on the risk of hepatocarcinogenesis predicted from liver imaging.

Furthermore, in the STARHE study, we have also developed an AI-based detection model for early-stage hepatocellular carcinoma. The developed object detection model achieved excellent performances in detecting very-early stage (< 2 cm) and early-stage hepatocellular carcinomas (overall rate of detected lesions = 68% and mAP10 = 0.67) on ultrasound cine clips. The confidence level in the predicted box of 70% should be tested in prospective longitudinal studies alongside radiologists' reading of the ultrasound images with two imaging outcomes: rates of detected early-stage stage hepatocellular carcinomas and rates of false positives, which could be a potential time hurdle for radiologists. This model could become a critical tool for radiologists and sonographers to improve the screening performance of ultrasound for early-stage hepatocellular carcinoma.

These results are reinforced by the strong methodology of the STARHE study. This is the first prospective multicentric study aiming to develop such models on ultrasound. Furthermore, the inclusion of patients eligible for hepatocellular carcinoma screening programs, the representation of the most common aetiologies of chronic liver disease (ALD, MASLD, controlled HBV and cured HVC) and the use of ultrasound cine clips mimic real-life practice of screening ultrasound, making the developed models applicable in clinical practice. On the other hand, the risk stratification and detection models have been developed following state-of-the-art AI methodology with a large training set and independent testing set, both stratified according to potential confounders (aetiology of liver disease, FASTRAK score, ultrasound manufacturer, hepatocellular carcinoma size and echogenicity). In addition, the testing set was designed to be representative of the targeted population with upstream sample size calculation based on previous reports[40-42]. The main limitation is the potential inclusion of patients at high risk of developing hepatocellular carcinoma in the low-risk group. Indeed, this limitation was only partially addressed by the follow-up at 1 year after the inclusion (e.g., hepatocellular carcinoma occurrence at 2 years after the inclusion). However, this would result in the underestimation of the performance of the risk stratification model. An alternative approach would have been to follow longitudinally only patients without hepatocellular carcinoma at baseline but this

approach would have been extremely challenging and expensive to implement. Also, to simplify the study design, there was no matched pair case-control, and this did not result in demographics groups imbalance. Finally, the risk stratification AI model was solely trained on the non-tumoral liver parenchyma preventing its training from any bias. In addition, only patients with early-stage hepatocellular carcinoma were included, none with more advanced stages, to prevent the inclusion of patients with infiltrative hepatocellular carcinoma.

The results of the MRI companion study will soon follow. If ultrasound is the modality of choice in clinical practice (availability, less expensive,...) and perfectly suited for an hepatocellular carcinoma-risk stratification deep learning model, it may not be contributory in all patients where visualisation of the liver is limited[13-15]. Therefore, developing equivalent models on ultrasound and MRI was advisable to address the applicability limitations of ultrasound. The MRI-based risk stratification model could also prove to be complementary to other risk stratification scores and further refine personalised prediction of hepatocarcinogenesis. The MRI-based detection model will definitely prove to be critical in assisting radiologists in the reading of aMRI. The choice of non-contrast aMRI could be criticised as it remains challenging to detect early-stage hepatocellular carcinoma on such a protocol and because the inter-reader agreement could be lower with a non-contrast protocol[24]. However, non-contrast aMRI has the advantages of the absence of contrast agent injection, simpler workflow, limited cost and the possibility to repeat poor quality acquisitions[27]. On the other hand, the challenges and limitations of dynamic contrast-enhanced and hepatobiliary aMRI are multiple: detection of inconclusive enhancing observations (need for recall examinations), injection of contrast, complex workflow with the need for intravenous access, and higher cost. Finally, the reported pooled sensitivity and specificity of these protocols are similar.

Alongside blood biomarkers such as FASTRAK score, these AI-based imaging models could provide the necessary tools to achieve personalised hepatocellular carcinoma screening and to significantly increase the number of patients who will benefit from the detection of hepatocellular carcinoma at an early stage, making them eligible for curative treatment with a

prognosis close to that of cancer-free liver disease. The next step will be to design a risk stratification-based personalised screening strategy integrating clinical, biochemical, and imaging risk stratification scores. In **Figure 11**, we propose a personalised screening strategy based on blood and imaging risk stratification, including ultrasound and abbreviated non-contrast MRI, defining 3 groups of different levels of risk with different screening modalities. Given the diagnostic performances of aMRI, this approach could significantly increase the number of patients who will benefit from the detection of hepatocellular carcinoma at an early stage. To date, the sensitivity of ultrasound is only 53% (and drops to 27.9% in early-stage hepatocellular carcinoma[11]), whereas the sensitivity of screening MRI is over 80%. This paradigm shift would have a considerable positive impact on the quality of patient care. This proposed strategy should be tested in a prospective clinical trial with hepatocellular carcinoma-related mortality as the main outcome and with cost-effectiveness analysis. This strategy could be refined over time with the discovery of new blood biomarkers. In addition, the AI-based imaging tools are also intended to be improved over the coming years with the inclusion of new patients in longitudinal follow-up cohorts.

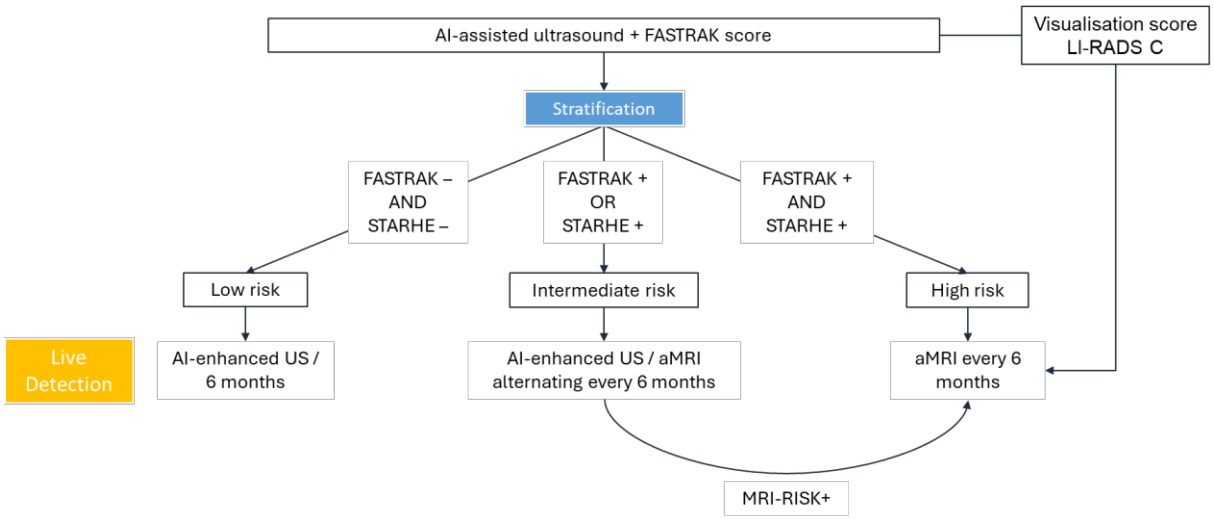


Figure 11 – Proposed risk-based personalised screening strategy.

Artificial intelligence is certainly not the only solution and innovative imaging can provide better understanding of chronic liver disease and help in their characterisation. In this thesis, we have studied the diagnostic capabilities of high-resolution 7T MRI in ex-vivo liver samples and developed an MR-derived METAVIR score in analogy with the histological staging criteria based on the presence and distribution of fibrosis. 7T high-resolution imaging demonstrated excellent performances (accuracy of 0.93) in accurately staging liver fibrosis compared to histopathology, highlighting its potential as an innovative surrogate tool for low-magnification histology. We showed that the physical capabilities of MRI can provide sufficient contrast between tissues to enable histopathology-like examination without the need for staining[58,70]. In addition, MRI has the undeniable advantage over histology of being able to acquire images in a relatively short time with immediate image interpretation (as with any routine MRI). The specimen, which remains intact, can still be processed for pathologic examination afterwards[58,70].

Reaching very high spatial resolution (~75µm), close to that of low-magnification histology, MRI provided a completely new insight into the imaging of the liver architecture, which would not have been visible with the spatial resolution of a clinical standard MRI (~ 1mm)[71,72]. Fibrotic changes to the liver parenchyma observed on Masson's trichrome stained histology slides were readily depictable on high-resolution MRI, appearing as hyperintense T2-weighted septa[58,71,72] together with fibrosis-related micro-architectural distortions at early stages[73]. Additionally, unlike histology, which is limited to 2D images, MRI provides images of the entire tissue sample as a volume. This volumetric analysis is also an advantage over traditional histology, particularly relevant for the study of fibrosis-related structural distortions in tissues.

Although the acquisition time is not yet suitable for extemporaneous examination, it is likely that, with rapid technological changes, the acquisition time can be further reduced, while maintaining sufficient spatial resolution and signal-to-noise ratio. Our study paved the way for further applications of the technique in the liver but also in other organs. Surgeries requiring extemporaneous analysis of tissues could benefit from this non-destructive imaging analysis to assess tumour margins. Similarly, the volumetric images of the entire specimen could also help

reduce sampling errors on macroscopic evaluation of the specimens, by providing more relevant mappings than macroscopic "naked eye" analysis, to select areas to be sectioned for further histological analysis.

Further research with a larger patient cohort and various aetiologies of liver disease is required to refine our understanding of 7T MRI liver semiology. Beyond fibrosis staging of liver tissue, MRI enables a cross-sectional volumetric exploration of the entire specimen, without cutting or destroying the sample. With short-time image acquisition and immediate image reading, high-resolution MRI could become a new modality for extemporaneous tissue analysis, especially in oncologic surgery. Future research developments should focus on bridging the gap with in-vivo MRI to impact clinical practice and provide true non-invasive microscopic histologic examination.

Perspectives

Multiple challenges remain. The personalised screening strategy based on blood and imaging risk stratification, including ultrasound and abbreviated non-contrast MRI, should be tested in a prospective clinical trial with hepatocellular carcinoma-related mortality as the main outcome and with cost-effectiveness analysis. Ultrasound will remain the cornerstone of chronic liver disease characterisation, including liver elastography and steatosis quantification, and hepatocellular carcinoma surveillance. It is therefore logical to maintain ultrasound-based tools as a first-line screening strategy alongside blood biomarkers. However, MRI biomarkers can be used to refine the screening strategy. In addition, quality control of ultrasound (LI-RADS visualisation score) should also be taken into consideration. In a retrospective cohort study, about 20% of patients with cirrhosis had moderately to severely limited ultrasound visualisation for hepatocellular carcinoma nodules, particularly those with obesity, ALD or MASLD cirrhosis[49]. Ultrasound quality was also shown to change between exams, including improvement in many patients with limited visualisation, encouraging longitudinal reassessment. The proposed risk-based personalised screening strategy could be refined over time with the discovery of new blood

biomarkers. In addition, the AI-based imaging tools are also intended to be improved over the coming years with the inclusion of new patients in longitudinal follow-up cohorts. Additional questions should be answered with longitudinal prospective screening studies: is this personalised screening strategy applicable to community centres? Is it universal and applicable to populations with different epidemiology? Is it acceptable for patients in terms of adherence and harm? What is the optimal frequency of abbreviated screening MRI?

The impact of this thesis in clinical practice could be major in the next years and contribute to improving patient care. The innovative AI-based imaging biomarkers developed in this study through a multidisciplinary collaboration could provide the necessary tools to achieve personalised hepatocellular carcinoma screening alongside blood biomarkers, allowing better detection of hepatocellular carcinoma at an early stage in a cost-effective approach. The proposed strategy based on blood, ultrasound, and MRI biomarkers should be tested in a prospective clinical trial. Imaging-based detection models are also likely to improve the diagnostic performance of screening modalities by assisting radiologists. The tools developed in this thesis are not theoretical constructs but are intended to integrate patient care in the short term.

Additional publications related to the thesis

CONVENTIONAL AND ARTIFICIAL INTELLIGENCE-BASED IMAGING FOR BIOMARKER DISCOVERY IN CHRONIC LIVER DISEASE

Chronic liver diseases, resulting from chronic injuries of various causes, lead to cirrhosis with life-threatening complications including liver failure, portal hypertension, and hepatocellular carcinoma. A key unmet medical need is robust non-invasive biomarkers to predict patient outcomes, stratify patients for risk of disease progression and monitor response to emerging therapies. Quantitative imaging biomarkers have already been developed, for instance, liver elastography for staging fibrosis or Proton Density Fat Fraction on Magnetic Resonance Imaging for liver steatosis. Yet, major improvements, in the field of image acquisition and analysis, are still required to be able to accurately characterise the liver parenchyma, monitor its changes and predict any pejorative evolution across disease progression. Artificial Intelligence has the potential to augment the exploitation of massive multi-parametric data to extract valuable information and achieve precision medicine. Machine learning algorithms have been developed to assess non-invasively certain histological characteristics of chronic liver diseases, including fibrosis and steatosis. Although still at an early stage of development, Artificial Intelligence-based imaging biomarkers provide novel opportunities to predict the risk of progression from early-stage chronic liver diseases towards cirrhosis-related complications, with the ultimate perspective of precision medicine.

Before developing new biomarkers in chronic liver diseases with a disruptive approach using artificial intelligence, we needed to establish a precise overview of already existing or emerging quantitative imaging techniques of diffuse liver diseases and provide an explanation of the different concepts of Artificial Intelligence. This work was published in Hepatology International.

Hepatology International
https://doi.org/10.1007/s12072-022-10303-0

REVIEW ARTICLE

Conventional and artificial intelligence-based imaging for biomarker discovery in chronic liver disease

Jérémy Dana[1,2,3,4] · Aïna Venkatasamy[2,5,6] · Antonio Saviano[1,3,7] · Joachim Lupberger[1,3] · Yujin Hoshida[8] · Valérie Vilgrain[9] · Pierre Nahon[10,11,12] · Caroline Reinhold[4,13,14] · Benoit Gallix[2,3,4] · Thomas F. Baumert[1,3,7]

Received: 5 November 2021 / Accepted: 17 January 2022

Abstract

Chronic liver diseases, resulting from chronic injuries of various causes, lead to cirrhosis with life-threatening complications including liver failure, portal hypertension, hepatocellular carcinoma. A key unmet medical need is robust non-invasive biomarkers to predict patient outcome, stratify patients for risk of disease progression and monitor response to emerging therapies. Quantitative imaging biomarkers have already been developed, for instance, liver elastography for staging fibrosis or proton density fat fraction on magnetic resonance imaging for liver steatosis. Yet, major improvements, in the field of image acquisition and analysis, are still required to be able to accurately characterize the liver parenchyma, monitor its changes and predict any pejorative evolution across disease progression. Artificial intelligence has the potential to augment the exploitation of massive multi-parametric data to extract valuable information and achieve precision medicine. Machine learning algorithms have been developed to assess non-invasively certain histological characteristics of chronic liver diseases, including fibrosis and steatosis. Although still at an early stage of development, artificial intelligence-based imaging biomarkers provide novel opportunities to predict the risk of progression from early-stage chronic liver diseases toward cirrhosis-related complications, with the ultimate perspective of precision medicine. This review provides an overview of emerging quantitative imaging techniques and the application of artificial intelligence for biomarker discovery in chronic liver disease.

Keywords Chronic liver disease · Histo-pathological features · Pejorative evolution · Quantitative biomarkers · Elastography · Machine learning · Radiomics · Deep learning

✉ Jérémy Dana
jeremy.dana@etu.unistra.fr

✉ Thomas F. Baumert
thomas.baumert@unistra.fr

1 Institut de Recherche sur les Maladies Virales et Hépatiques, Institut National de la Santé et de la Recherche Médicale (Inserm), U1110, 3 Rue Koeberlé, 67000 Strasbourg, France

2 Institut Hospitalo-Universitaire (IHU), Strasbourg, France

3 Université de Strasbourg, Strasbourg, France

4 Department of Diagnostic Radiology, McGill University, Montreal, Canada

5 Streinth Lab (Stress Response and Innovative Therapies), Inserm UMR_S 1113 IRFAC, Interface Recherche Fondamentale et Appliquée à la Cancérologie, 3 Avenue Moliere, Strasbourg, France

6 Department of Radiology Medical Physics, Faculty of Medicine, Medical Center-University of Freiburg, University of Freiburg, Killianstrasse 5a, 79106 Freiburg, Germany

7 Pôle Hépato-Digestif, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

8 Liver Tumor Translational Research Program, Division of Digestive and Liver Diseases, Department of Internal Medicine, Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, USA

9 Radiology Department, Hôpital Beaujon, Université de Paris, CRI, INSERM 1149, APHP. Nord, Paris, France

10 Liver Unit, Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpitaux Universitaires Paris Seine Saint-Denis, Bobigny, France

11 Université Sorbonne Paris Nord, 93000 Bobigny, France

12 Inserm, UMR-1138 "Functional Genomics of Solid Tumors", Paris, France

13 Augmented Intelligence and Precision Health Laboratory, Research Institute of McGill University Health Centre, Montreal, Canada

14 Montreal Imaging Experts Inc., Montreal, Canada

Published online: 09 February 2022

Introduction

Over the last decades, the prevalence of chronic liver diseases (CLD) and their associated morbidity and mortality markedly increased, especially with the rise of non-alcoholic fatty liver disease (NAFLD). A substantial proportion of patients will indeed ultimately develop liver fibrosis and eventually progress toward cirrhosis. Cirrhosis is the end-stage of disease with life-threatening complications (e.g., liver failure, portal hypertension, hepatocellular carcinoma), which accounts for approximately 1.8% of deaths [1]. When chronic liver injuries progress, decompensation of the disease (e.g., ascites, jaundice, gastrointestinal bleeding or hepatic encephalopathy) may occur, resulting in a dramatic decrease in the overall survival rate [2]. Currently, the clinical predictors of the risk of decompensation have a limited impact on the patients' management and we are unable to accurately monitor the changes or any pejorative evolution of liver parenchyma on imaging alone (e.g., using CT, MRI, ultrasound) [3–5]. The characterization of reference of CLD relies on invasive methods such as liver biopsy, to assess fibrosis, steatosis, and "activity" (i.e., inflammation) or trans-jugular catheterization for portal hypertension (i.e., measure of the hepatic venous pressure gradient). Such invasive and expensive gold standards are obviously inappropriate for screening and sequential monitoring. Additionally, liver biopsy is also prone to risks of under-sampling [6] and/or inter-reader variability [7]. All this leads to a necessary and ongoing transition toward non-invasive assessment of CLD progression and prognosis. Image-based biomarkers can provide a quantitative and reproducible representation of the liver parenchyma including pathogenesis, molecular and genetic pathways and, particularly, of its evolution [8, 9]. Indeed, they can be used at initial diagnosis or at any time during the evolution of the disease, creating the opportunity to impact clinical management.

Several image-based quantitative biomarkers have already emerged. For instance, fibrosis can be non-invasively assessed by ultrasound- or MRI-based elastography techniques. Approaches have been developed to estimate steatosis, exploiting the attenuation of ultrasonic waves or employing advanced MRI acquisitions techniques (e.g., multi-echo DIXON, spectroscopy). Several bio-clinical scoring systems based on routine parameters and liver elastography have proven valuable to predict the first liver-related event and overall survival in patients with cirrhosis [10–12].

Recently, artificial intelligence (AI) has gained spectacular popularity in the scientific community, suggesting that we are at the dawn of a revolution in patients' care and management. The major strength of AI is its potential to augment the exploitation of massive multi-parametric data, often non-structured and unexploited, to extract valuable information and achieve personalized clinical decisions for patients [9, 13]. AI has the potential to go beyond the human eye and previously cited tools, to finally make biopsy outdated. This review article aims to provide a precise overview of quantitative imaging techniques of diffuse liver diseases, together with an explanation of the different concepts of artificial intelligence, with short- and long-term potential clinical applications for risk stratification and early diagnosis.

Artificial intelligence in imaging

Artificial intelligence (AI), a subfield of computer science, is a "fancy" term gathering different concepts including among others radiomics and machine learning. More precisely, machine learning is the umbrella term referring to the approaches seeking to learn from data without explicit programming. Machine learning can achieve tasks of classification, prediction, segmentation, detection, or images optimization (e.g., faster image acquisition, increased signal-to-noise ratio, etc.). The tasks of segmentation, detection and optimization of images will not be discussed in this article as not directly related to the characterization of CLD.

To achieve classification or prediction of clinical outcomes, different approaches exist, according to the available data and the objectives. The machine can learn from labeled data (e.g., tumor types) to pursue a defined objective (e.g., tumor type classification) or from unlabeled data to reveal unknown structural patterns across data. These approaches are respectively called supervised and unsupervised.

Therefore, AI-based imaging models, or the machine learning process, will seek to identify and combine new imaging biomarkers, inaccessible to the human eye, in a mathematical model [14]. It aims to provide predictive and/or prognostic information about patients and their pathologies, based on sophisticated statistical analysis [8]. Such imaging biomarkers are called radiomics. They are image descriptors reflecting the tissue heterogeneity and indirectly its molecular and genetic substrate [15]. They are reproducible, quantitative, human-engineered (so-called "traditional") or free (so-called "deep" and automatically calculated). "Traditional" radiomics should be extracted, selected, and combined using a classifier (support vector machine, random forest, etc.) through a high-quality pipeline to ensure its robustness. These key steps should always be detailed to ensure the repeatability of the work. To provide reproducible and standardized processing workflow, but also consistency between studies, the image biomarker standardization initiative (IBSI) proposed biomarkers definitions and reporting guidelines for radiomics studies, including among

other data description, image processing or image biomarker computation [16]. Compared to "traditional" radiomics, "deep" radiomics are free of human design, inaccessible to our understanding, and may highlight the most informative imaging markers to solve research hypotheses.

Deep learning (DL) [13, 17], a subclass of machine learning, refers to the deep convolutional neural network (CNN), named by analogy with human neurons (Fig. 1). Input data are weighted based on their importance and undergo a non-linear transformation, called activation function, to result in an output. During the training process, these input weights, or parameters, are computed and optimized, to allow the model to reach the highest diagnostic performances by minimizing the loss error function through a process called back-propagation. Upstream of the adjustment of the weights, designing a DL model requires the right choices regarding the most appropriate CNN architecture and hyper-parameters (number of hidden layers and units, weights initialization, type of activation function, learning rate, regularization technique to prevent overfitting, etc.) for the specific task to be achieved [13, 18]. The more hidden layers there are, the deeper a convolutional neural network is and the more complex the network is. Complexifying a neural network allows the identification and pooling of images features of a higher level of abstraction, thus increasing its performances. However, convolutional neural networks can be so powerful that they can perfectly adjust to a specific dataset, so-called overfitting, resulting in very high diagnostic performances on the training dataset, but preventing its external validation.

It is therefore evident that the quality of data has a major impact on the performance and reliability of AI-based models. First, training datasets should represent the wide spectrum of disease expression. As training datasets are usually built from retrospective data, selection bias should be of critical concern. Because available data in medicine are limited and positive cases of the disease are usually the minority class, techniques of data augmentation (i.e., simple geometric transformations of images or artificial creation of fake data from the original dataset using DL techniques—e.g., generative adversarial network) or transfer learning (i.e., pre-training on larger—not necessarily medical-related—datasets leading to pre-trained model parameters and, faster

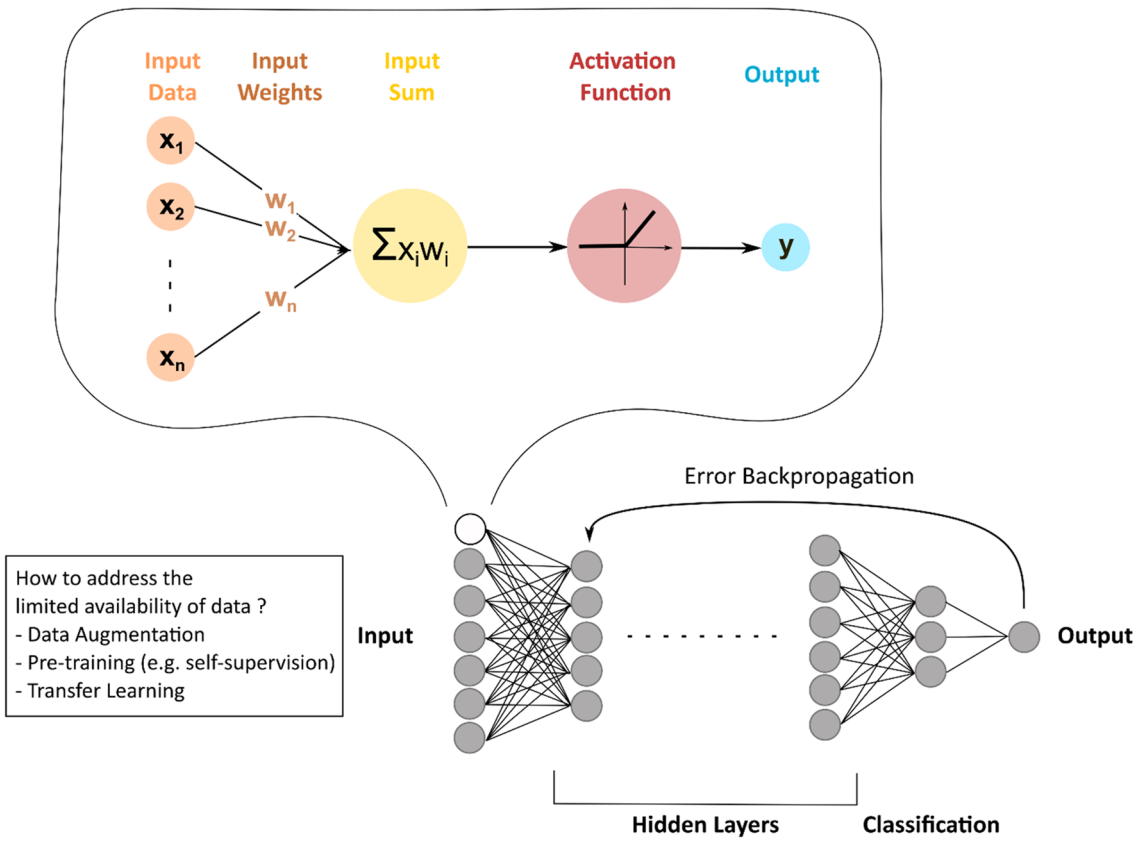


Fig. 1 Concepts of deep learning. By analogy to human neurons, deep learning generally refers to neural networks. Input data are weighted based on their importance and undergo a non-linear transformation, called activation function, to result in an output. These input weights, or parameters, are computed and optimized to allow the model to reach the highest diagnostic performances by minimizing the loss error function through a process called back-propagation

and more effective training) have been developed. However, techniques of data augmentation should be used with great caution in the field of medical research and only to optimize model training, not to test it. They should be applied by considering disease prevalence and heterogeneity. Furthermore, creating fake images can result in impaired model training. Therefore, samples of such images should be checked by experts. Secondly, testing a model, preferably prospectively, on an independent dataset, as a safety check, is crucial to evaluate its true performance and ensure that there is no overfitting. Moreover, a special focus should be made on the exhaustivity of the representation of the disease spectrum in the testing dataset. These quality steps are of critical importance because machine learning algorithms can be difficult, if not impossible, to understand. To facilitate the quality assessment of AI-related studies, quality scores have been proposed (e.g., radiomics quality score [19], simplified and reproducible AI quality score [20]). Unfortunately, a significant number of research studies does not respect these quality pre-requisites. This explains the limited number of studies cited in this review. However, this cannot fully explain the discrepancy between the considerable number of publications, increasingly following standardized reporting guidelines (e.g., IBSI, CLAIM [16, 21]), and their poor implementation in clinical routine. Multiple obstacles arise. As previously explained, neural networks rely on complex non-linear interactions using hidden factors, making the concepts of transparency, explainability, intelligibility and provability inaccessible, although critical for their acceptability by physicians and patients [22]. The presence of different manufacturers (i.e., vendors), the heterogeneity of imaging acquisition protocols (e.g., different times of contrast injection between centers; use of conventional CT or on dual-energy reconstruction; different T2-weighted MR images depending on the center-specific and non-consensual choice of echo time and repetition time; etc.), and the absence of large and free databases are a direct limitation to the robustness and safety of AI-based models that are even more crucial in medicine. Furthermore, AI-based models should be time-efficient to reach the clinical routine. For instance, manual segmentation cannot be seriously considered and implies developing robust and reproducible automatic segmentation algorithms. Besides these technical limitations, ethical and legal considerations are at stake: how should the patient be informed that medical decisions are enhanced by AI-based algorithms? How can patient privacy be preserved once data are shared with AI developer partners? Who owns the intellectual property of the AI model, computer scientists or data owners? Who will be accountable and responsible for decision-making with AI including potential errors and harm? [23]

Imaging biomarkers for liver fibrosis

Over the past years, significant efforts have been made to develop new quantitative imaging biomarkers, aiming to replace liver biopsy to assess fibrosis, steatosis, iron overload and inflammation (Table 1).

Fibrosis is the inevitable consequence of all progressive CLD. It is mainly caused by sustained liver insults, resulting in pathological deposition of collagen extracellular matrix and, ultimately, parenchyma and vascular distortion with regenerative nodules [24]. Currently, conventional imaging modalities fail to properly characterize liver fibrosis. Accurate detection of early-stage fibrosis is necessary because appropriate therapeutic management could stop the evolution of fibrosis to cirrhosis. Morphological changes, irregular contours and coarse texture have limited sensitivity to predict significant fibrosis (≥ METAVIR F2) and poorly correlate with fibrosis stages [25]. Quantitative measurement of the liver surface nodularity may improve consistency for the imaging diagnosis of cirrhosis (i.e., METAVIR F4 stage) [26]. This biomarker has also been associated with the detection of portal hypertension and has proven relevant for preoperative assessment of operative risks in patients with resectable hepatocellular carcinoma (HCC) [27, 28]. Gadoxetic acid-enhanced MRI could also be used to estimate

Table 1 Non-invasive conventional quantitative imaging methods for assessing liver histo-pathological features

Histo-pathological features	Methods
Fibrosis	Elastography: transient elastography, point and 2D shearwave elastography, magnetic resonance elastography Liver surface nodularity Enhancement of liver parenchyma on MR hepatobiliary phase
Steatosis	Controlled attenuation parameter (CAP) on transient elastography Ultrasound-based: attenuation coefficient, hepato-renal B-mode ratio, sound speed MRI-based: DIXON method (in/out phase), multi-echo DIXON method (proton density fat fraction), spectroscopy
Iron	MRI-based: multi-echo DIXON method
Activity	Ultrasound-based: shear wave dispersion MRI-based: damping ratio (complex shear modulus), enhancement of liver parenchyma on hepatobiliary phase, T1 relaxation time, proton-decoupled phosphorus 31 MR spectroscopy

fibrosis stages, as relative enhancement on the hepato-biliary phase (at 20 min) negatively correlated with fibrosis [29].

Non-invasive assessment of liver fibrosis has undergone a breakthrough with the rise of elastography techniques (quantitative methods). Liver stiffness, based on the elastic properties of liver tissue, is the non-invasive biomarker of choice for the diagnosis of liver fibrosis, even at an early stage. Elastography techniques demonstrated higher staging performance than serum fibrosis indexes [30], such as the aspartate aminotransferase to platelet index (APRI) or the Fibrosis-4 index (FIB-4), even if their association may be beneficial and complementary [31–33]. Different modalities are available, including vibration-controlled transient elastography (Fibroscan, Echosens, Paris, France), ultrasound-guided elastography [such as point shear wave elastography (pSWE) and 2D shear wave elastography (2D SWE)] and magnetic resonance elastography (MRE) [34, 35]. Transient elastography (TE) estimates the liver stiffness by measuring the speed of a shear wave propagating through the liver parenchyma using pulsed echo ultrasound acquisition. It has been exhaustively evaluated and shown to be effective in predicting advanced fibrosis stages (AUC > 0.72 for ≥ F2 stages and AUC > 0.90 for F4 stage) [36–43]. As it is not associated with any imaging modality, this technique will not be further discussed in this review. However, the same technology has been applied to ultrasound, allowing targeted measurements guided by the imaging abnormalities of the liver parenchyma, with at least equivalent diagnostic performances [44–51]. Unlike pSWE, which only enables a focal measurement of liver stiffness, 2D SWE provides a real-time 2D color mapping of liver stiffness. Unfortunately, each manufacturer providing ultrasound-guided elastography has its specificities, preventing cross-comparison and complexifying the use of cut-off values. In addition, the reproducibility of measurements is affected by the experience of the operator [52]. If 2D SWE is perfectly suited for clinical practice, as it is performed during a conventional ultrasound, its diagnostic performances have been outperformed by MRE, which quickly became the surrogate biomarker of liver fibrosis [53]. This technology is based on shear waves emitted by an external acoustic driver. Indeed, the wave propagation speed and the damping of shear waves are impacted by the frequency vibration due to the dispersion of elastic waves in soft tissues [54]. 3D MRE should be based on multi-frequency excitations because it increases the consistency and reproducibility of the measurements. MRE demonstrated higher accuracy than 2D SWE or TE, especially for early-stage liver fibrosis, with strong reliability for longitudinal follow-up and without inter-observer variability [53, 55–63]. Furthermore, it can be performed in the presence of ascites and measurements are not affected by steatosis [64]. However, MRE is not recommended in routine clinical practice given its cost and limited availability [33]. In addition, 2D SWE also demonstrated high accuracy in predicting first liver-related event, all-cause mortality and infection requiring hospitalization [11].

If the association of conventional imaging and non-invasive assessment of liver stiffness is powerful for grading fibrosis, artificial intelligence can maximize the diagnostic performances of these techniques, by identifying new features (Table 2). In patients at risk of advanced CLD, liver ultrasound is the first imaging modality performed, because of its advantages (i.e., available, non-invasive, radiation-free, less expensive) compared to other techniques (CT or MRI) and the possibility of performing shear wave elastography during the same examination. The same reasons should provide strong relevance for ultrasound-based AI models in clinical routine. A recent study demonstrated high accuracy for the prediction of fibrosis stages using a Deep Convolutional Neural Network trained on B-mode gray-scale ultrasonography images [65]. In an external testing dataset, consistent with acceptable generalizability, the accuracy of the model to predict significant fibrosis (≥ F2) or cirrhosis (F4) was 0.87 and 0.86 respectively. Moreover, applying radiomics analysis to the images of the 2D SWE color mapping could further improve its diagnostic performances. Wang et al. reported increased diagnostic performances of a deep learning model using 2D shear wave elastography (2D SWE) images in predicting liver fibrosis stages [66]. The diagnostic performances of this AI model in predicting significant fibrosis (≥ F2) were higher (AUC = 0.85) than that of 2D SWE alone (AUC = 0.77) or biomarkers, such as APRI (AUC = 0.60) or FIB-4 (AUC = 0.62). Finally, B-mode and 2D SWE images could prove complementary in the training of DL models as suggested by Xue et al. [67].

MRI is the most performant, exhaustive and reproducible imaging modality. This explains the predominance of this modality in the AI literature on this problematic. Different studies demonstrated strong diagnostic performances of radiomics [68] and deep learning models [69, 70], either on T2-weighted or on post-contrast sequences. Hectors et al. developed a deep learning model on hepatobiliary phase images, with similar performances to MRE (AUC = 0.91 for predicting significant fibrosis) [69]. However, these encouraging results should be tempered by the constraints of clinical practice. MRI remains a time-consuming and costly technique, especially when compared to the existing efficient ultrasound-based AI models.

Finally, it is interesting to note that the deep learning approach can also be extremely performant using CT images, whereas CT-scan is not the modality of choice in liver imaging. A deep learning algorithm trained on 7461 portal venous phase CT scans with pathologically confirmed liver fibrosis largely outperformed radiologists' reading and fibrosis biomarkers (APRI and FIB-4) [71]. This model achieved high diagnostic performances, regardless of the

Table 2 Artificial intelligence-based biomarkers for grading fibrosis

Article	Model type	Imaging technique	Reference standards	Sample size (*n*)				Performance metrics on testing dataset
				Training	Validation	Testing		
						Internal	External	
Lee European Radiology 2020	Deep learning	B-mode US	METAVIR stages (biopsy or transient elastography)	3446	263	266	572	AUC = F4: 0.86
Wang Gut 2019	Deep learning	2D shear wave elastography	METAVIR stages (liver biopsy)	266		132		AUC = F4: 0.97 ≥ F3: 0.98 ≥ F2: 0.85
Xue European Radiology 2020	Deep learning	B-mode US and 2D shear wave elastography	Scheuer scoring system (hepatectomy)	364		102		AUC = F4: 0.95 ≥ F3: 0.93 ≥ F2: 0.93
He American Journal of Roentgenology 2019	"Traditional" Radiomics	T2 FSE weighted MRI Clinical factors	MR elastography Two-class classification (cut-off = 3 kPa)	225			84	AUC 0.80
Hectors European Radiology 2020	Deep learning	Gadoxetic acid-enhanced hepatobiliary phase MRI	METAVIR stages (liver biopsy)	178	123	54		AUC = F4: 0.85 ≥ F3: 0.90 ≥ F2: 0.91
Yasaka Radiology 2018	Deep learning	Gadoxetic acid-enhanced hepatobiliary phase MRI	METAVIR stages (liver biopsy)	534		100		AUC = F4: 0.84 = F3: 0.84 = F2: 0.85
Choi Radiology 2018	Deep learning	Portal venous phase CT	METAVIR stages (liver biopsy)	7461		421	470	AUC = F4: 0.95 ≥ F3: 0.97 ≥ F2: 0.96

For comparative purposes, AUROC values of transient elastography, shear wave elastography and magnetic resonance elastography for detecting advanced fibrosis (≥ stage 3) were 0.88, 0.95 and 0.96 in a meta-analysis published by Xiao et al. [21]

etiology of liver disease, with an AUC value of 0.96 for predicting significant fibrosis (≥ F2) in a large testing dataset of 891 patients.

Imaging biomarkers for liver overload

In the context of the increasing prevalence of overweight and type 2 diabetes mellitus, the prevalence of NAFLD is expanding [72]. It covers a wide spectrum of diseases, ranging from isolated liver steatosis to nonalcoholic steatohepatitis, resulting in severe complications including cirrhosis, liver failure, portal hypertension and hepatocellular carcinoma [73–75]. Liver steatosis can be easily assessed. Indeed, steatosis can be evaluated either by TE using the controlled attenuation parameter, ultrasound (e.g., hepato-renal B-mode ratio, attenuation coefficient [76], sound speed [77, 78]) or MRI (e.g., Dixon method with in- and out-of-phases, spectroscopy, proton density fat fraction—PDFF [79]). If the quantification of steatosis on a non-enhanced CT-scan appears easy, as there is a linear correlation between liver attenuation and steatosis, enabling quantitative CT liver fat measurements, CT scan is not a suitable modality for steatosis assessment, due to its poorer diagnostic performance and its ionizing aspect [80]. The controlled attenuation parameter (CAP) and the attenuation coefficient (AC) are the most routinely performed biomarkers with the hepato-renal B-mode ratio. They are based on the same principle of ultrasonic attenuation of the echo wave by the steatotic liver. Measurement of this attenuation allows the estimation of steatosis. The diagnostic performance of CAP and AC in predicting any grade of steatosis, or moderate to severe steatosis (grade 2 and 3), is good, with AUC values of 0.93 versus 0.81 and 0.76 versus 0.89, respectively [77, 81]. The hepato-renal B-mode ratio is defined as the ratio of the echogenicity of the liver parenchyma to the renal cortex. Moret et al. demonstrated that the diagnostic performances of the B-mode ratio and the CAP were not significantly different in the same population [81]. Lastly, ultrasonic adaptive sound speed estimation, decreased in the presence of steatosis, has

been proposed but is still at a preliminary stage of evaluation [77, 78]. Alternatively to the CAP and the AC, MRI offers multiple techniques to estimate steatosis, with unequal diagnostic performances. The Dixon method (1984) is a chemical-shift imaging method using the in-phase/out-of-phase cycling of fat and water, due to different rates of precession. In the presence of steatosis, the signal intensity of the liver drops on the out-of-phase sequence. However, this method is highly subject to inter-reader variability and does not allow quantification of the steatosis. Quantification of steatosis has been achieved by the application of multi-echo Dixon, which compensates for multiple confounders including the T2* relaxation effects and the spectral complexity of fat [82]. This method is named proton density fat fraction (PDFF), defined as the fraction of mobile protons (H1) linked to the triglyceride relative to those of water [e.g., IDEAL IQ (General Electrics), mDixon-Quant (Philips) and Multi-echo VIBE Dixon (Siemens)]. Finally, steatosis can also be assessed by MR spectroscopy, which directly measures the relative proton quantity from water and triglycerides signals. However, this method is limited by the delicate spectral analysis of data and its sampling volume. PDFF-MRI is considered the method of reference, as it allows quantification of steatosis in the entire liver and because it is easy to perform and analyze [82]. It should also be noted that thresholds for grading steatosis differ between PDFF-MRI (6.4%, 17.4%, and 22.1%) and histological analysis (5%, 33%, and 66%). Indeed, the methods of evaluation of steatosis are different [82]. PDFF-MRI considers the proportion of mobile protons within fat molecules in a three-dimensional voxel, whereas histological analysis evaluates the fractions of hepatocytes with fat vacuoles in a two-dimensional plane.

However, if MRI-PDFF is the non-invasive gold standard to assess steatosis, it gathers the limitations of MRI in routine clinical practice (time-consuming and costly technique). The most original and interesting AI-based approach has been published by Han et al. [83]. They developed a one-dimensional deep learning model using raw radiofrequency ultrasound data to diagnose NAFLD and quantify the hepatic fat fraction. If inaccessible to the medical framework, raw radiofrequency ultrasound signals are richer in information than gray-scale B-mode images. It allowed a strong correlation between the ultrasound-based predicted fat fraction and MRI-PDFF (Pearson $r = 0.85$; $p < 0.001$), with excellent accuracy (96%) for NAFLD diagnosis in the test cohort. However, the diagnostic performances of the model decreased when MRI-PDFF was greater than 18%.

Besides steatosis, MRI is also the non-invasive gold standard to detect and quantify liver iron concentration [84, 85]. It is a reliable method based on multi-echo gradient-echo sequences, available on every device, either 1.5 or 3-Tesla MRI. Liver iron overload results in lower liver intensity due to T2 and T2* relaxation time shortening. Quantification can be obtained by the computation of T2* (or R2*), by measuring the liver to muscle signal intensity ratio (SIR) or by combining both methods. It can be coupled with that of liver steatosis by the DIXON method. As iron deposition in the liver is responsible for toxicity, monitoring liver iron overload could become a prognostic factor of progression of CLD [86, 87]. Neither ultrasound nor CT scan, whether or not enhanced by AI, has proven valuable for the detection or quantification of iron overload.

Pejorative evolution of chronic liver diseases

The prediction of the risk of progression of CLD is in the spotlight. It is particularly true with the increasing prevalence of NAFLD as its progression to steatohepatitis (NASH) predisposes to cirrhosis and HCC. NASH is characterized by the presence of steatosis with lobular inflammation and hepatocyte ballooning, leading to necrosis, apoptosis, increased collagen extracellular matrix and ultimately, fibrosis. In this field, the development of non-invasive markers is at its preliminary stages. Recently, 2D SWE demonstrated greater capabilities in liver characterization than solely stiffness assessment using shear waves speed. Indeed, shear waves disperse as they pass through the liver. Such dispersion can be estimated using a mathematical parameter called the dispersion slope. Sugimoto et al. suggested that this parameter was indirectly impacted by lobular inflammation, which could be helpful to detect and grade inflammation [88, 89], gold standard but could also be biased as it correlates with liver fibrosis [90, 91]. The combination of this parameter with the assessment of steatosis (using the attenuation coefficient) and fibrosis (using shear wave elasticity) could become an acceptable substitute to the pathology gold standard in NASH. These concepts have also been explored using MR elastography. The damping ratio, derived from the complex shear modulus, could discriminate NASH, even without advanced fibrosis, raising the possibility of reflecting inflammation [92]. Besides fibrosis as mentioned above, Bastati et al. also showed that gadoxetic acid-enhanced MRI could be used to distinguish NAFLD from NASH, as the relative enhancement on the hepato-biliary phase (at 20 min) negatively correlated with the degree of lobular inflammation and ballooning, but not with steatosis [29]. Other MRI-derived parameters, such as the T1 relaxation time, have also proven valuable in identifying NASH when combined with fat fraction and liver stiffness [93, 94]. Proton-decoupled phosphorus 31 MR spectroscopy may also help because of the changes in metabolites concentrations in NASH including NADPH (reduced form of nicotinamide adenine dinucleotide phosphate), a marker of inflammation and fibrinogenic activity in the liver [95]. The relative failure

of accurately assessing non-invasively liver inflammation may be overcome by AI-based techniques in future.

In compensated advanced CLD, as the problem is no longer to predict the risk of progression of fibrosis, it is relevant to focus on the patient survival cliff which is characterized by the occurrence of HCC, portal hypertension decompensation or liver failure.

If HCC risk stratification models based on clinical-biological (age, sex, diabetes, AST/ALT, albumin, platelets, etc.) parameters exist [96–102], they cannot consider the direct analysis of the liver parenchyma, which is the pathophysiological substrate of hepato-carcinogenesis. In the 1990s, several authors studied the incidence of HCC according to the liver echostructure [103–105]. They concurred on the excess risk of a nodular heterogeneous echostructure with an estimated relative risk of up to 20 [103]. Unfortunately, this did not lead to the development of reliable imaging risk stratification models.

Clinically significant portal hypertension (CSPH), defined by a hepatic venous pressure gradient (HVPG) $\geq$ 10 mmHg, is critical for CLD prognosis. If the definition relies on the invasive measurement of the HPVG, different non-invasive liver-based approaches have been developed. As previously discussed, liver stiffness is a robust biomarker of liver fibrosis. It was therefore expected to observe a correlation between liver stiffness and HPVG and a capability to discriminate patients with CSPH [106–110]. Furthermore, liver stiffness was also proven to have a prognostic value for portal hypertension-related complications including clinical decompensation and variceal bleeding [111–113]. As liver surface nodularity is also a biomarker of cirrhosis, it has been shown to have a similar performance to liver stiffness for the detection of CSPH [114]. However, such diagnostic performances are only true in portal hypertension secondary to cirrhosis, not in pre- or post-sinusoidal portal hypertension. In contrast, spleen stiffness could prove to be a promising technique for monitoring HPVG [115]. On the other hand, an innovative approach consisted of the development of a computational model for estimating HVPG based on CT angiographic images [116]. Recently, AI-based models, either traditional radiomics or deep learning, have been developed for CT and/or MRI, with very high diagnostic performances [117–120]. More precisely, Liu et al. developed two DL CNN models (CT- and MRI-based) on liver and spleen images that achieved strong diagnostic performances for identifying patients with CSPH with an AUC value of 0.93 (CT) and 0.94 (MRI) on an independent testing dataset. These models outperformed liver stiffness (AUC = 0.73) [118].

Predicting liver failure is also crucial for patient management, either during the natural course of CLD or preoperatively before major hepatectomy. To this date, the decision of portal vein embolization before major hepatectomy still relies on the simple measurement of the liver volume. Accurate prediction of postoperative liver failure is still not implemented in the clinical routine. This could beneficiate from the development of AI-based models. Indeed, several studies reported performant predictive radiomics models for identifying patients at risk of liver failure after major hepatectomy [121–123].

Future perspectives

To date, reproducible quantitative imaging biomarkers are available to assess liver fibrosis, steatosis, and iron overload. If MR imaging with elastography is the most exhaustive modality to assess CLD, as these biomarkers can be evaluated during a single examination, without the need of contrast agents (Fig. 2), liver ultrasound with the concomitant performance of US-guided elastography during the same examination seems the most relevant and time-efficient first-line technique for clinical routine. Artificial intelligence has already proven valuable to create new biomarkers [124] and/or increase the diagnostic performances of the existing ones [66], but has not integrated routine clinical practice yet. As MRI cannot be extensively recommended in routine clinical practice given its cost and limited availability, AI could help maximize diagnostic performances of ultrasound. AI may also help in automatizing time-consuming tasks such as measurement of the liver volume using deep CNN [125, 126], a simple prognostic biomarker of the pejorative outcome of acute liver failure [127]. But to become clinical tools, AI models should be developed following a high-standard process to achieve generalizability and transferability including training on datasets representing the wide spectrum of the disease expression to avoid selection biases, and independent and prospective testing to avoid overfitting [20].

Furthermore, despite the remarkable rise of quantitative imaging biomarkers for the prediction of pathological features, some decisive clinical needs remain unmet. The assessment of the short- and long-term risk of progression of CLD toward a pejorative outcome (e.g., liver failure, portal hypertension decompensation or HCC [96–102]) still requires the development of reliable non-invasive tools. This absence can be explained by the difficulty of implementing studies that would need to be exhaustive and prospective over a long period to collect a large number of pejorative events. If fibrosis and steatosis appear as decisive markers for the characterization of CLD, they fail to accurately predict the progression of early-stage CLD to cirrhosis-related complications. Assessing the disease activity, or inflammation, would better reflect the risk of progressive fibrosis and thus its complications. Refining risk stratification of progressive disease from initial diagnosis would majorly impact

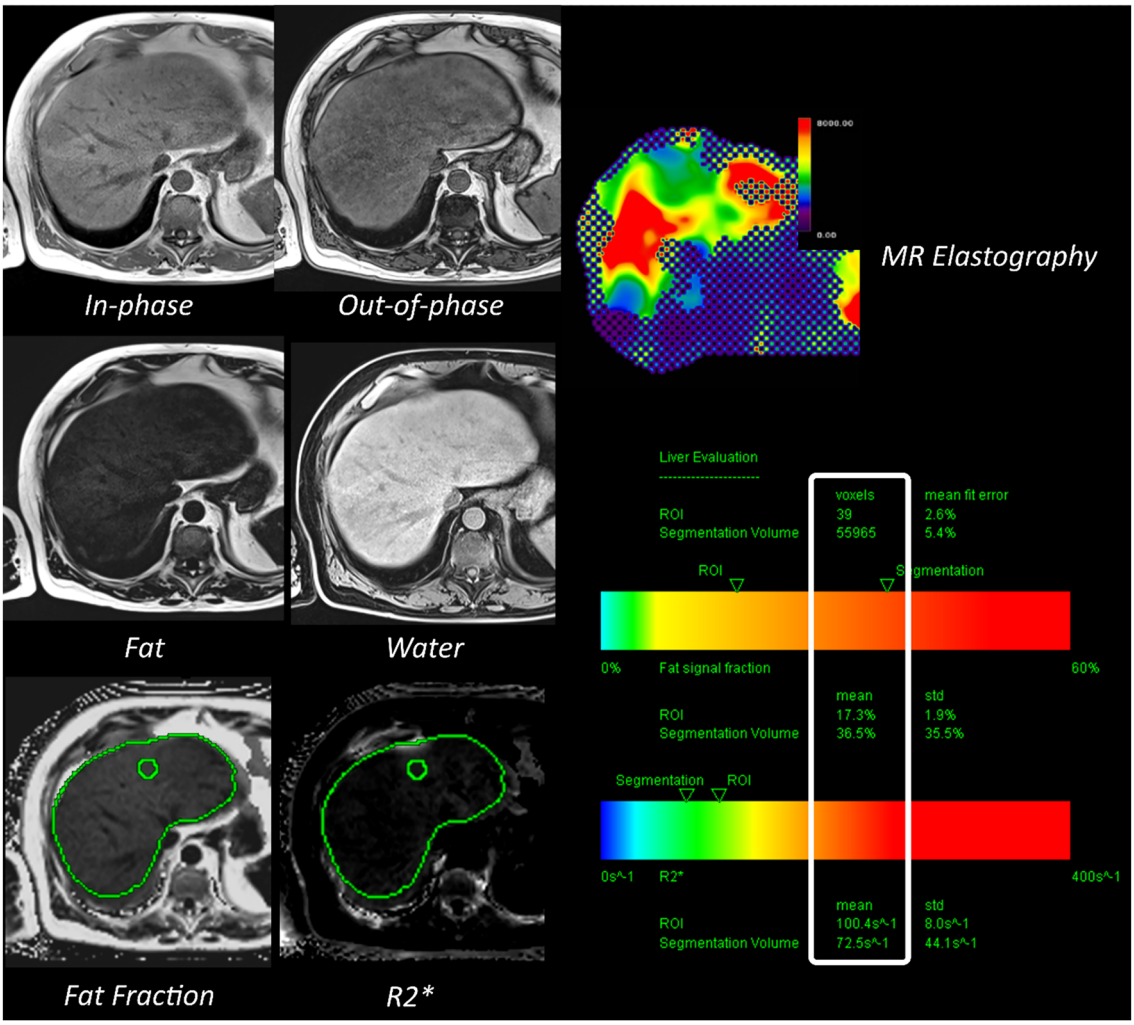


Fig. 2 Magnetic resonance imaging-based quantitative biomarkers for steatosis (fat fraction), iron overload (R2*) and fibrosis (MR elastography)

therapeutic management. Unfortunately, at the date of this review article, only preliminary research tools exist, without currently clinical transfer and applicability, and none was based on AI techniques. Stratification of the disease progression is crucial for the accurate selection of patients who will most benefit from treatment, therefore avoiding side effects if no benefit is expected, to achieve the best clinical outcome/cost ratio.

Conclusion

As varied as they are, image-based biomarkers can provide a comprehensive representation of the liver parenchyma at the time of initial diagnosis, or at any time during the disease, creating the opportunity to outdate invasive gold standards and impact on clinical management. Artificial intelligence provides opportunities to revolutionize liver imaging, by creating novel reproducible and quantitative imaging biomarkers and augmenting human intelligence to improve decision-making and operational processes. It aims to be part of personalized care, from diagnosis to treatment, as it learns without explicit programming. To achieve this goal, certain limitations need to be overcome. Extensive work is still required to substantiate AI by pathology, molecular and genetic substrate. Precision medicine may ultimately be achieved by integrating clinical, biological (such as single-cell RNA-seq, exome sequencing), serological (such as blood-based biomarkers) and imaging data.

A future challenge for meeting the clinical needs of CLD is the stratification of the risk of disease progression to pejorative outcomes, aiming at identifying patients who will most benefit from treatments. In this regard, it is of paramount importance that AI models will be developed with the concept of a future integration as part of the clinical routine enabling their widespread application.

Funding This work was supported by French state funds managed within the "Plan Investissements d'Avenir" and by the ANR (ANR-10-IAHU-02 to B.G), by the ARC, Paris and Institut Hospitalo-Universitaire, Strasbourg (TheraHCC1.0 and 2.0 IHUARC IHU201301187 and IHUARC2019 to T.F.B.), the European Union (ERC-AdG-2014-671231-HEPCIR to T.F.B. and Y.H., ERC-AdG-2020-667273-FIBCAN to T.F.B. and Y. H.), ANRS, Paris (ECTZ171594 to J.L., ECTZ131760 to J.L. and P.N., ECTZ160436 and ECTZ103701 to T.F.B), NIH (DK099558 and CA233794 to Y.H., CA209940 and R03AI131066 to T.F.B.), Cancer Prevention and Research Institute of Texas (RR180016 to Y.H), US Department of Defense (W81XWH-16-1-0363 to T.F.B. and Y.H.), the Irma T. Hirschl/Monique Weill-Caulier Trust (Y.H.) and the Foundation of the University of Strasbourg (HEPKIN to T.F.B.) and the Institut Universitaire de France (IUF; T.F.B.). This work has been published under the framework of the LABEX ANR-10-LABX-0028_ HEPSYS and Inserm Plan Cancer and benefits from funding from the state managed by the French National Research Agency as part of the Investments for the future program.

Declarations

Conflict of interest TFB is founder, shareholder and advisor of Alentis Therapeutics. He is an inventor on patent applications of the University of Strasbourg, Inserm and IHU for liver disease therapeutics and biomarkers. PN has relationships with AstraZeneca, Bayer, Bristol-Myers Squibb, EISAI, Ipsen, Roche. JD, AV, AS, JL, YH, VV, CR and BG declare no conflict of interest with this publication.

References

1. WHO | Projections of mortality and causes of death, 2016 to 2060. World Health Organization. World Health Organization; 2016.
2. D'Amico G, Garcia-Tsao G, Pagliaro L. Natural history and prognostic indicators of survival in cirrhosis: a systematic review of 118 studies. J Hepatol. 2006;44:217–231
3. Hu K-Q, Tong MJ. The long-term outcomes of patients with compensated hepatitis C virus–related cirrhosis and history of parenteral exposure in the united states. Hepatology. 1999;29:1311–1316
4. Benvegnù L, Gios M, Boccato S, Alberti A. Natural history of compensated viral cirrhosis: a prospective study on the incidence and hierarchy of major complications. Gut. 2004;53:744–749
5. Jepsen P, Ott P, Andersen PK, Sørensen HT, Vilstrup H. Clinical course of alcoholic liver cirrhosis: a Danish population-based cohort study. Hepatology. 2010;51:1675–1682
6. Regev A, Berho M, Jeffers LJ, Milikowski C, Molina EG, Pyrsopoulos NT, et al. Sampling error and intraobserver variation in liver biopsy in patients with chronic HCV infection. Am J Gastroenterol. 2002;97:2614–2618
7. Merriman RB, Ferrell LD, Patti MG, Weston SR, Pabst MS, Aouizerat BE, et al. Correlation of paired liver biopsies in morbidly obese patients with suspected nonalcoholic fatty liver disease. Hepatology. 2006;44:874–880
8. Dana J, Agnus V, Ouhmich F, Gallix B. Multimodality imaging and artificial intelligence for tumor characterization: current status and future perspective. Semin Nucl Med. 2020. http://www.sciencedirect.com/science/article/pii/S000129982030074X. Accessed 2 Aug 2020.
9. Aerts HJWL. The potential of radiomic-based phenotyping in precision medicine: a review. JAMA Oncol. 2016;2:1636–1642
10. López SA, Manzano ML, Gea F, Gutiérrez ML, Ahumada AM, Devesa MJ, et al. A model based on noninvasive markers predicts very low hepatocellular carcinoma risk after viral response in hepatitis C virus-advanced fibrosis. Hepatology. 2020;72:1924–1934
11. Rasmussen DN, Thiele M, Johansen S, Kjærgaard M, Lindvig KP, Israelsen M, et al. Prognostic performance of seven biomarkers compared to liver biopsy in early alcohol-related liver disease. J Hepatol. 2021. https://www.sciencedirect.com/science/article/pii/S0168827821004116. Accessed 14 Jul 2021.
12. Hoshida Y, Villanueva A, Sangiovanni A, Sole M, Hur C, Andersson KL, et al. Prognostic gene expression signature for patients with hepatitis C-related early-stage cirrhosis. Gastroenterology. 2013;144:1024–1030
13. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep learning: a primer for radiologists. Radiographics. 2017;37:2113–2131
14. Savadjiev P, Chong J, Dohan A, Vakalopoulou M, Reinhold C, Paragios N, et al. Demystification of AI-driven medical image interpretation: past, present and future. Eur Radiol. 2019;29:1616–1624
15. Savadjiev P, Chong J, Dohan A, Agnus V, Forghani R, Reinhold C, et al. Image-based biomarkers for solid tumor quantification. Eur Radiol. 2019;29:5431–5440
16. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. 2020;295:328–338
17. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–444
18. Cheng PM, Montagnon E, Yamashita R, Pan I, Cadrin-Chênevert A, Perdigón Romero F, et al. Deep learning: an update for radiologists. Radiographics. 2021;41:1427–1445
19. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14:749–762
20. Lecointre L, Dana J, Lodi M, Akladios C, Gallix B. Artificial intelligence-based radiomics models in endometrial cancer: a systematic review. Eur J Surg Oncol. 2021;47:2734–2741
21. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2:e200029
22. Hinton G. Deep learning-a technology with the potential to transform health care. JAMA. 2018;320:1101–1102
23. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. 2021.
24. Hernandez-Gea V, Friedman SL. Pathogenesis of liver fibrosis. Annu Rev Pathol. 2011;6:425–456
25. Vilgrain V, Lagadec M, Ronot M. Pitfalls in liver imaging. Radiology. 2015;278:34–51
26. Smith AD, Branch CR, Zand K, Subramony C, Zhang H, Thaggard K, et al. Liver surface nodularity quantification from routine CT images as a biomarker for detection and evaluation of cirrhosis. Radiology. 2016;280:771–781
27. Sartoris R, Rautou P-E, Elkrief L, Pollorsi G, Durand F, Valla D, et al. Quantification of liver surface nodularity at CT: utility for detection of portal hypertension. Radiology. 2018;289:698–707
28. Hobeika C, Cauchy F, Sartoris R, Beaufrère A, Yoh T, Vilgrain V, et al. Relevance of liver surface nodularity for preoperative risk assessment in patients with resectable hepatocellular carcinoma. Br J Surg. 2020;107:878–888
29. Bastati N, Feier D, Wibmer A, Traussnigg S, Balassy C, Tamandl D, et al. Noninvasive differentiation of simple steatosis and steatohepatitis by using gadoxetic acid-enhanced MR imaging in patients with nonalcoholic fatty liver disease: a proof-of-concept study. Radiology. 2014;271:739–747

30. Xiao G, Zhu S, Xiao X, Yan L, Yang J, Wu G. Comparison of laboratory tests, ultrasound, or magnetic resonance elastography to detect fibrosis in patients with nonalcoholic fatty liver disease: a meta-analysis. Hepatology. 2017;66:1486–1501
31. Calvopina DA, Noble C, Weis A, Hartel GF, Ramm LE, Balouch F, et al. Supersonic shear-wave elastography and APRI for the detection and staging of liver disease in pediatric cystic fibrosis. J Cyst Fibros. 2019;19:449–454
32. Lewindon PJ, Puertolas-Lopez MV, Ramm LE, Noble C, Pereira TN, Wixey JA, et al. Accuracy of transient elastography data combined with APRI in detection and staging of liver disease in pediatric patients with cystic fibrosis. Clin Gastroenterol Hepatol. 2019;17:2561-2569.e5
33. Berzigotti A, Tsochatzis E, Boursier J, Castera L, Cazzagon N, Friedrich-Rust M, et al. EASL clinical practice guidelines on non-invasive tests for evaluation of liver disease severity and prognosis – 2021 update. J Hepatol. 2021. https://www.sciencedirect.com/science/article/pii/S0168827821003986. Accessed 22 Jun 2021.
34. Tang A, Cloutier G, Szeverenyi NM, Sirlin CB. Ultrasound elastography and MR elastography for assessing liver fibrosis: part 1, principles and techniques. Am J Roentgenol. 2015;205:22–32
35. Tang A, Cloutier G, Szeverenyi NM, Sirlin CB. Ultrasound elastography and MR elastography for assessing liver fibrosis: part 2, diagnostic performance, confounders, and future directions. Am J Roentgenol. 2015;205:33–40
36. Foucher J, Chanteloup E, Vergniol J, Castéra L, Le Bail B, Adhoute X, et al. Diagnosis of cirrhosis by transient elastography (FibroScan): a prospective study. Gut. 2006;55:403–408
37. Poynard T, Vergniol J, Ngo Y, Foucher J, Munteanu M, Merrouche W, et al. Staging chronic hepatitis C in seven categories using fibrosis biomarker (FibroTest™) and transient elastography (FibroScan®). J Hepatol. 2014;60:706–714
38. Rajakannu M, Coilly A, Adam R, Samuel D, Vibert E. Prospective validation of transient elastography for staging liver fibrosis in patients undergoing hepatectomy and liver transplantation. J Hepatol. 2018;68:199–200
39. Castéra L, Vergniol J, Foucher J, Le Bail B, Chanteloup E, Haaser M, et al. Prospective comparison of transient elastography, Fibrotest, APRI, and liver biopsy for the assessment of fibrosis in chronic hepatitis C. Gastroenterology. 2005;128:343–350
40. Ziol M, Handra-Luca A, Kettaneh A, Christidis C, Mal F, Kazemi F, et al. Noninvasive assessment of liver fibrosis by measurement of stiffness in patients with chronic hepatitis C. Hepatology. 2005;41:48–54
41. Ganne-Carrié N, Ziol M, de Ledinghen V, Douvin C, Marcellin P, Castera L, et al. Accuracy of liver stiffness measurement for the diagnosis of cirrhosis in patients with chronic liver diseases. Hepatology. 2006;44:1511–1517
42. de Lédinghen V, Douvin C, Kettaneh A, Ziol M, Roulot D, Marcellin P, et al. Diagnosis of hepatic fibrosis and cirrhosis by transient elastography in HIV/hepatitis C virus-coinfected patients. J Acquir Immune Defic Syndr. 2006;41:175–179
43. Castera L, Forns X, Alberti A. Non-invasive evaluation of liver fibrosis using transient elastography. J Hepatol. 2008;48:835–847
44. Cassinotto C, Boursier J, Paisant A, Guiu B, Irles-Depe M, Canivet C, et al. Transient versus 2-dimensional shear-wave elastography in a multistep strategy to detect advanced fibrosis in NAFLD. Hepatology. http://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/hep.31655. Accessed 14 May 2021.
45. Gao Y, Zheng J, Liang P, Tong M, Wang J, Wu C, et al. Liver fibrosis with two-dimensional US shear-wave elastography in participants with chronic hepatitis B: a prospective multicenter study. Radiology. 2018;289:407–415
46. Yoneda M, Thomas E, Sclair SN, Grant TT, Schiff ER. Supersonic shear imaging and transient elastography with the XL probe accurately detect fibrosis in overweight or obese patients with chronic liver disease. Clin Gastroenterol Hepatol. 2015;13:1502-1509.e5
47. Leung VY, Shen J, Wong VW, Abrigo J, Wong GL, Chim AM, et al. Quantitative elastography of liver fibrosis and spleen stiffness in chronic hepatitis B carriers: comparison of shear-wave elastography and transient elastography with liver biopsy correlation. Radiology. 2013;269:910–918
48. Friedrich-Rust M, Lupsor M, de Knegt R, Dries V, Buggisch P, Gebel M, et al. Point shear wave elastography by acoustic radiation force impulse quantification in comparison to transient elastography for the noninvasive assessment of liver fibrosis in chronic hepatitis C: a prospective international multicenter study. Ultraschall Med. 2015;36:239–247
49. Ferraioli G, Tinelli C, Bello BD, Zicchetti M, Filice G, Filice C. Accuracy of real-time shear wave elastography for assessing liver fibrosis in chronic hepatitis C: a pilot study. Hepatology. 2012;56:2125–2133
50. Zhuang Y, Ding H, Zhang Y, Sun H, Xu C, Wang W. Two-dimensional shear-wave elastography performance in the noninvasive evaluation of liver fibrosis in patients with chronic hepatitis B: comparison with serum fibrosis indexes. Radiology. 2016;283:873–882
51. Zheng J, Guo H, Zeng J, Huang Z, Zheng B, Ren J, et al. Two-dimensional shear-wave elastography and conventional US: the optimal evaluation of liver fibrosis and cirrhosis. Radiology. 2015;275:290–300
52. Ferraioli G, Tinelli C, Zicchetti M, Above E, Poma G, Di Gregorio M, et al. Reproducibility of real-time shear wave elastography in the evaluation of liver elasticity. Eur J Radiol. 2012;81:3102–3106
53. Lefebvre T, Wartelle-Bladou C, Wong P, Sebastiani G, Giard J-M, Castel H, et al. Prospective comparison of transient, point shear wave, and magnetic resonance elastography for staging liver fibrosis. Eur Radiol. 2019;29:6477–6488
54. Asbach P, Klatt D, Hamhaber U, Braun J, Somasundaram R, Hamm B, et al. Assessment of liver viscoelasticity using multifrequency MR elastography. Magn Reson Med. 2008;60:373–379
55. Dyvorne HA, Jajamovich GH, Bane O, Fiel MI, Chou H, Schiano TD, et al. Prospective comparison of magnetic resonance imaging to transient elastography and serum markers for liver fibrosis detection. Liver Int. 2016;36:659–666
56. Chen J, Yin M, Talwalkar JA, Oudry J, Glaser KJ, Smyrk TC, et al. Diagnostic performance of MR elastography and vibration-controlled transient elastography in the detection of hepatic fibrosis in patients with severe to morbid obesity. Radiology. 2016;283:418–428
57. Imajo K, Kessoku T, Honda Y, Tomeno W, Ogawa Y, Mawatari H, et al. Magnetic resonance imaging more accurately classifies steatosis and fibrosis in patients with nonalcoholic fatty liver disease than transient elastography. Gastroenterology. 2016;150:626-637.e7
58. Loomba R, Wolfson T, Ang B, Hooker J, Behling C, Peterson M, et al. Magnetic resonance elastography predicts advanced fibrosis in patients with nonalcoholic fatty liver disease: a prospective study. Hepatology. 2014;60:1920–1928
59. Shi Y, Guo Q, Xia F, Dzyubak B, Glaser KJ, Li Q, et al. MR elastography for the assessment of hepatic fibrosis in patients with chronic hepatitis B infection: does histologic necroinflammation influence the measurement of hepatic stiffness? Radiology. 2014;273:88–98
60. Chang W, Lee JM, Yoon JH, Han JK, Choi BI, Yoon JH, et al. Liver fibrosis staging with MR elastography: comparison of diagnostic performance between patients with chronic hepatitis B and those with other etiologic causes. Radiology. 2016;280:88–97

61. Cui J, Heba E, Hernandez C, Haufe W, Hooker J, Andre MP, et al. Magnetic resonance elastography is superior to acoustic radiation force impulse for the diagnosis of fibrosis in patients with biopsy-proven nonalcoholic fatty liver disease: a prospective study. Hepatology. 2016;63:453–461
62. Yin M, Talwalkar JA, Glaser KJ, Manduca A, Grimm RC, Rossman PJ, et al. Assessment of hepatic fibrosis with magnetic resonance elastography. Clin Gastroenterol Hepatol. 2007;5:1207-1213.e2
63. Huwart L, Sempoux C, Vicaut E, Salameh N, Annet L, Danse E, et al. Magnetic resonance elastography for the noninvasive staging of liver fibrosis. Gastroenterology. 2008;135:32–40
64. Venkatesh SK, Yin M, Ehman RL. Magnetic resonance elastography of liver: technique, analysis, and clinical applications. J Magn Reson Imaging. 2013;37:544–555
65. Lee JH, Joo I, Kang TW, Paik YH, Sinn DH, Ha SY, et al. Deep learning with ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network. Eur Radiol. 2020;30:1264–1273
66. Wang K, Lu X, Zhou H, Gao Y, Zheng J, Tong M, et al. Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study. Gut. 2019;68:729–741
67. Xue L-Y, Jiang Z-Y, Fu T-T, Wang Q-M, Zhu Y-L, Dai M, et al. Transfer learning radiomics based on multimodal ultrasound imaging for staging liver fibrosis. Eur Radiol. 2020;30:2973–2983
68. He L, Li H, Dudley JA, Maloney TC, Brady SL, Somasundaram E, et al. Machine learning prediction of liver stiffness using clinical and T2-weighted MRI radiomic data. AJR Am J Roentgenol. 2019;213:592–601
69. Hectors SJ, Kennedy P, Huang K-H, Stocker D, Carbonell G, Greenspan H, et al. Fully automated prediction of liver fibrosis using deep learning analysis of gadoxetic acid-enhanced MRI. Eur Radiol. 2020;31:3805–3814
70. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Liver fibrosis: deep convolutional neural network for staging by using gadoxetic acid-enhanced hepatobiliary phase MR images. Radiology. 2018;287:146–155
71. Choi KJ, Jang JK, Lee SS, Sung YS, Shim WH, Kim HS, et al. Development and validation of a deep learning system for staging liver fibrosis by using contrast agent-enhanced CT images in the liver. Radiology. 2018;289:688–697
72. Younossi ZM. Non-alcoholic fatty liver disease – a global public health perspective. J Hepatol. 2019;70:531–544
73. Nguyen VH, Le MH, Cheung RC, Nguyen MH. Differential clinical characteristics and mortality outcomes in persons with NAFLD and/or MAFLD. Clin Gastroenterol Hepatol. 2021. https://www.sciencedirect.com/science/article/pii/S1542356521005670X. Accessed 25 Aug 2021.
74. Natarajan Y, Kramer JR, Yu X, Li L, Thrift AP, El-Serag HB, et al. Risk of cirrhosis and hepatocellular cancer in patients with NAFLD and normal liver enzymes. Hepatology. 2020;72:1242–1252
75. Mendes FD, Suzuki A, Sanderson SO, Lindor KD, Angulo P. Prevalence and indicators of portal hypertension in patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol. 2012;10:1028-1033.e2
76. Zhang B, Ding F, Chen T, Xia L-H, Qian J, Lv G-Y. Ultrasound hepatic/renal ratio and hepatic attenuation rate for quantifying liver fat content. World J Gastroenterol. 2014;20:17985–17992
77. Dioguardi Burgio M, Ronot M, Reizine E, Rautou P-E, Castera L, Paradis V, et al. Quantification of hepatic steatosis with ultrasound: promising role of attenuation imaging coefficient in a biopsy-proven cohort. Eur Radiol. 2020;30:2293–2301
78. Imbault M, Burgio MD, Faccinetto A, Ronot M, Bendjador H, Deffieux T, et al. Ultrasonic fat fraction quantification using in vivo adaptive sound speed estimation. Phys Med Biol. 2018;63:215013
79. Runge JH, Smits LP, Verheij J, Depla A, Kuiken SD, Baak BC, et al. MR spectroscopy–derived proton density fat fraction is superior to controlled attenuation parameter for detecting and grading hepatic steatosis. Radiology. 2017;286:547–556
80. Guo Z, Blake GM, Li K, Liang W, Zhang W, Zhang Y, et al. Liver fat content measurement with quantitative CT validated against MRI proton density fat fraction: a prospective study of 400 healthy volunteers. Radiology. 2020;294:89–97
81. Moret A, Boursier J, Debry PH, Riou J, Crouan A, Dubois M, et al. Evaluation of the hepatorenal B-mode ratio and the "controlled attenuation parameter" for the detection and grading of steatosis. Ultraschall Med. 2020. http://www.thieme.connect.de/DOI/DOI?10.1055/a-1233-2290. Accessed 15 May 2021.
82. Tang A, Tan J, Sun M, Hamilton G, Bydder M, Wolfson T, et al. Nonalcoholic fatty liver disease: MR imaging of liver proton density fat fraction to assess hepatic steatosis. Radiology. 2013;267:422–431
83. Han A, Byra M, Heba E, Andre MP, Erdman JW, Loomba R, et al. Noninvasive diagnosis of nonalcoholic fatty liver disease and quantification of liver fat with radiofrequency ultrasound data using one-dimensional convolutional neural networks. Radiology. 2020;295:342–350
84. d'Assignies G, Paisant A, Bardou-Jacquet E, Boulic A, Bannier E, Lainé F, et al. Non-invasive measurement of liver iron concentration using 3-Tesla magnetic resonance imaging: validation against biopsy. Eur Radiol. 2018;28:2022–2030
85. Henninger B, Alustiza J, Garbowski M, Gandon Y. Practical guide to quantification of hepatic iron with MRI. Eur Radiol. 2020;30:383–393
86. Czaja AJ. Iron disturbances in chronic liver diseases other than haemochromatosis – pathogenic, prognostic, and therapeutic implications. Aliment Pharmacol Ther. 2019;49:681–701
87. Pietrangelo A. Hereditary hemochromatosis: pathogenesis, diagnosis, and treatment. Gastroenterology. 2010;139:393-408.e2
88. Sugimoto K, Moriyasu F, Oshiro H, Takeuchi H, Abe M, Yoshimasu Y, et al. The role of multiparametric US of the liver for the evaluation of nonalcoholic steatohepatitis. Radiology. 2020;296:532–540
89. Sugimoto K, Moriyasu F, Oshiro H, Takeuchi H, Yoshimasu Y, Kasai Y, et al. Viscoelasticity measurement in rat livers using shear-wave US elastography. Ultrasound Med Biol. 2018;44:2018–2024
90. Deffieux T, Gennisson J-L, Bousquet L, Corouge M, Cosconea S, Amroun D, et al. Investigating liver stiffness and viscosity for fibrosis, steatosis and activity staging using shear wave elastography. J Hepatol. 2015;62:317–324
91. Chen S, Sanchez W, Callstrom MR, Gorman B, Lewis JT, Sanderson SO, et al. Assessment of liver viscoelasticity by using shear waves induced by ultrasound radiation force. Radiology. 2013;266:964–970
92. Allen AM, Shah VH, Therneau TM, Venkatesh SK, Mounajjed T, Larson JJ, et al. The role of three-dimensional magnetic resonance elastography in the diagnosis of nonalcoholic steatohepatitis in obese patients undergoing bariatric surgery. Hepatology. 2020;71:510–521
93. Kim JW, Lee Y-S, Park YS, Kim B-H, Lee SY, Yeon JE, et al. Multiparametric MR index for the diagnosis of non-alcoholic steatohepatitis in patients with non-alcoholic fatty liver disease. Sci Rep. 2020;10:2671

94. Ding Y, Rao S-X, Meng T, Chen C, Li R, Zeng M-S. Usefulness of T1 mapping on Gd-EOB-DTPA-enhanced MR imaging in assessment of non-alcoholic fatty liver disease. Eur Radiol. 2014;24:959–966
95. Sevastianova K, Hakkarainen A, Kotronen A, Cornér A, Arkkila P, Arola J, et al. Nonalcoholic fatty liver disease: detection of elevated nicotinamide adenine dinucleotide phosphate with in vivo 3.0-T 31P MR spectroscopy with proton decoupling. Radiology. 2010;256:466–473
96. Ioannou GN. HCC surveillance after SVR in patients with F3/F4 fibrosis. J Hepatol. 2021;74:458–465
97. Ioannou GN, Green P, Kerr KF, Berry K. Models estimating risk of hepatocellular carcinoma in patients with alcohol or NAFLD-related cirrhosis for risk stratification. J Hepatol. 2019;71:523–533
98. Ioannou GN, Tang W, Beste LA, Tincopa MA, Su GL, Van T, et al. Assessment of a deep learning model to predict hepatocellular carcinoma in patients with hepatitis C cirrhosis. JAMA Netw Open. 2020;3:e2015626–e2015626
99. Sharma SA, Kowgier M, Hansen BE, Brouwer WP, Maan R, Wong D, et al. Toronto HCC risk index: a validated scoring system to predict 10-year risk of HCC in patients with cirrhosis. J Hepatol. 2018;68:92–99
100. Papatheodoridis G, Dalekos G, Sypsa V, Yurdaydin C, Buti M, Goulis J, et al. PAGE-B predicts the risk of developing hepatocellular carcinoma in Caucasians with chronic hepatitis B on 5-year antiviral therapy. J Hepatol. 2016;64:800–806
101. Fan R, Papatheodoridis G, Sun J, Innes H, Toyoda H, Xie Q, et al. aMAP risk score predicts hepatocellular carcinoma development in patients with chronic hepatitis. J Hepatol. 2020;73:1368–1378
102. Audureau E, Carrat F, Layese R, Cagnot C, Asselah T, Guyader D, et al. Personalized surveillance for hepatocellular carcinoma in cirrhosis – using machine learning adapted to HCV status. J Hepatology. 2020. https://www.journal-of-hepatology.eu/article/S0168-8278(20)30394-9/abstract. Accessed 30 Jun 2020.
103. Kitamura S, Iishi H, Tatsuta M, Ishikawa H, Hiyama T, Tsukuma H, et al. Liver with hypoechoic nodular pattern as a risk factor for hepatocellular carcinoma. Gastroenterology. 1995;108:1778–1784
104. Tarao K, Hoshino H, Shimizu A, Ohkawa S, Harada M, Nakamura Y, et al. Patients with ultrasonic coarse-nodular cirrhosis who are anti-hepatitis C virus-positive are at high risk for hepatocellular carcinoma. Cancer. 1995;75:1255–1262
105. Caturelli E, Castellano L, Fusilli S, Palmentieri B, Niro GA, del Vecchio-Blanco C, et al. Coarse nodular US pattern in hepatic cirrhosis: risk for hepatocellular carcinoma. Radiology. 2003;226:691–697
106. Kitson MT, Roberts SK, Colman JC, Paul E, Button P, Kemp W. Liver stiffness and the prediction of clinically significant portal hypertension and portal hypertensive complications. Scand J Gastroenterol. 2015;50:462–469
107. Elkrief L, Rautou P-E, Ronot M, Lambert S, Dioguardi Burgio M, Francoz C, et al. Prospective comparison of spleen and liver stiffness by using shear-wave and transient elastography for detection of portal hypertension in cirrhosis. Radiology. 2015;275:589–598
108. Elkrief L, Ronot M, Andrade F, Dioguardi Burgio M, Issoufaly T, Zappa M, et al. Non-invasive evaluation of portal hypertension using shear-wave elastography: analysis of two algorithms combining liver and spleen stiffness in 191 patients with cirrhosis. Aliment Pharmacol Ther. 2018;47:621–630
109. Ronot M, Lambert S, Elkrief L, Doblas S, Rautou P-E, Castera L, et al. Assessment of portal hypertension and high-risk oesophageal varices with liver and spleen three-dimensional multifrequency MR elastography in liver cirrhosis. Eur Radiol. 2014;24:1394–1402
110. Choi S-Y, Jeong WK, Kim Y, Kim J, Kim TY, Sohn JH. Shear-wave elastography: a noninvasive tool for monitoring changing hepatic venous pressure gradients in patients with cirrhosis. Radiology. 2014;273:917–926
111. Grgurević I, Bokun T, Mustapić S, Trkulja V, Heinzl R, Banić M, et al. Real-time two-dimensional shear wave ultrasound elastography of the liver is a reliable predictor of clinical outcomes and the presence of esophageal varices in patients with compensated liver cirrhosis. Croat Med J. 2015;56:470–481
112. Merchante N, Rivero-Juárez A, Téllez F, Merino D, Ríos-Villegas MJ, Ojeda-Burgos G, et al. Liver stiffness predicts variceal bleeding in HIV/HCV-coinfected patients with compensated cirrhosis. AIDS. 2017;31:493–500
113. Robic MA, Procopet B, Métivier S, Péron JM, Selves J, Vinel JP, et al. Liver stiffness accurately predicts portal hypertension related complications in patients with chronic liver disease: a prospective study. J Hepatol. 2011;55:1017–1024
114. Souhami A, Sartoris R, Rautou P-E, Cauchy F, Bouattour M, Durand F, et al. Similar performance of liver stiffness measurement and liver surface nodularity for the detection of portal hypertension in patients with hepatocellular carcinoma. JHEP Rep. 2020;2:100147
115. Takuma Y, Nouso K, Morimoto Y, Tomokuni J, Sahara A, Takabatake H, et al. Portal hypertension in patients with liver cirrhosis: diagnostic accuracy of spleen stiffness. Radiology. 2016;279:609–619
116. Qi X, An W, Liu F, Qi R, Wang L, Liu Y, et al. Virtual hepatic venous pressure gradient with CT angiography (CHESS 1601): a prospective multicenter study for the noninvasive diagnosis of portal hypertension. Radiology. 2019;290:370–377
117. Lin Y, Li L, Yu D, Liu Z, Zhang S, Wang Q, et al. A novel radiomics-platelet nomogram for the prediction of gastroesophageal varices needing treatment in cirrhotic patients. Hepatol Int. 2021;15:995–1005
118. Liu Y, Ning Z, Örmeci N, An W, Yu Q, Han K, et al. Deep convolutional neural network-aided detection of portal hypertension in patients with cirrhosis. Clin Gastroenterol Hepatol. 2020;18:2998-3007.e5
119. Tseng Y, Ma L, Li S, Luo T, Luo J, Zhang W, et al. Application of CT-based radiomics in predicting portal pressure and patient outcome in portal hypertension. Eur J Radiol. 2020;126:108927
120. Meng D, Wei Y, Feng X, Kang B, Wang X, Qi J, et al. CT-based radiomics score can accurately predict esophageal variceal rebleeding in cirrhotic patients. Front Med. 2021;8:745931
121. Zhu W-S, Shi S-Y, Yang Z-H, Song C, Shen J. Radiomics model based on preoperative gadoxetic acid-enhanced MRI for predicting liver failure. World J Gastroenterol. 2020;26:1208–1220
122. Chen Y, Liu Z, Mo Y, Li B, Zhou Q, Peng S, et al. Prediction of post-hepatectomy liver failure in patients with hepatocellular carcinoma based on radiomics using Gd-EOB-DTPA-enhanced MRI: the liver failure model. Front Oncol. 2021;11:605296
123. Cai W, He B, Hu M, Zhang W, Xiao D, Yu H, et al. A radiomics-based nomogram for the preoperative prediction of posthepatectomy liver failure in patients with hepatocellular carcinoma. Surg Oncol. 2019;28:78–85
124. Xu X, Zhang H-L, Liu Q-P, Sun S-W, Zhang J, Zhu F-P, et al. Radiomic analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma. J Hepatol. 2019;70:1133–1144
125. Chlebus G, Schenk A, Moltz JH, van Ginneken B, Hahn HK, Meine H. Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. Sci Rep. 2018;8:15497
126. Chlebus G, Meine H, Thoduka S, Abolmaali N, van Ginneken B, Hahn HK, et al. Reducing inter-observer variability and

interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections. PLoS One. 2019;14:e0217228

127. Zabron A, Quaglia A, Fatourou E, Peddu P, Lewis D, Heneghan M, et al. Clinical and prognostic associations of liver volume determined by computed tomography in acute liver failure. Liver Int. 2018;38:1592–1601

MULTIMODALITY IMAGING AND ARTIFICIAL INTELLIGENCE FOR TUMOUR CHARACTERIZATION: CURRENT STATUS AND FUTURE PERSPECTIVE

Further explanations of the different concepts of Artificial Intelligence along with the requirements to develop clinically relevant and safe models were discussed in "*Multimodality Imaging and Artificial Intelligence for Tumour Characterization: Current Status and Future Perspective*" published in Seminars in Nuclear Medicine. We also introduced the potential applications in oncology. Indeed, Research in medical imaging has yet to achieve precision oncology. Over the past 30 years, only the simplest imaging biomarkers (RECIST, SUV,...) have become widespread clinical tools. This may be due to our inability to accurately characterise tumours and monitor intra-tumoral changes in imaging. Artificial intelligence, through machine learning and deep learning, opens a new path in medical research because it can bring together a large amount of heterogeneous data into the same analysis to reach a single outcome. Supervised or unsupervised learning may lead to new paradigms by identifying unrevealed structural patterns across data. Deep learning will provide human-free, undefined upstream, reproducible, and automated quantitative imaging biomarkers. Since tumour phenotype is driven by its genotype and thus indirectly defines tumoral progression, tumour characterisation using machine learning and deep learning algorithms will allow us to monitor molecular expression noninvasively, anticipate therapeutic failure, and lead therapeutic management. To follow this path, quality standards have to be set: standardization of imaging acquisition as it has been done in the field of biology, transparency of the model development as it should be reproducible by different institutions, validation, and testing through a high-quality process using large and complex open databases and better interpretability of these algorithms.

ELSEVIER

Seminars in
NUCLEAR
MEDICINE

Multimodality Imaging and Artificial Intelligence for Tumor Characterization: Current Status and Future Perspective

Jérémy Dana, M.D.,[*,†,‡] Vincent Agnus, Ph.D.,[*,§] Farid Ouhmich, M.Sc.,[*,§] and Benoit Gallix, M.D., Ph.D.[*,§,‖,¶]

Research in medical imaging has yet to do to achieve precision oncology. Over the past 30 years, only the simplest imaging biomarkers (RECIST, SUV,. . .) have become widespread clinical tools. This may be due to our inability to accurately characterize tumors and monitor intratumoral changes in imaging. Artificial intelligence, through machine learning and deep learning, opens a new path in medical research because it can bring together a large amount of heterogeneous data into the same analysis to reach a single outcome. Supervised or unsupervised learning may lead to new paradigms by identifying unrevealed structural patterns across data. Deep learning will provide human-free, undefined upstream, reproducible, and automated quantitative imaging biomarkers. Since tumor phenotype is driven by its genotype and thus indirectly defines tumoral progression, tumor characterization using machine learning and deep learning algorithms will allow us to monitor molecular expression noninvasively, anticipate therapeutic failure, and lead therapeutic management. To follow this path, quality standards have to be set: standardization of imaging acquisition as it has been done in the field of biology, transparency of the model development as it should be reproducible by different institutions, validation, and testing through a high-quality process using large and complex open databases and better interpretability of these algorithms.
Semin Nucl Med 50:541-548

Introduction

Artificial intelligence (AI) is a widely spread term referring to different fields and leading to different objectives. Medicine and patient care are at the dawn of a revolution. Future is personalized medicine, from diagnosis to treatment, and machine learning (ML) will be part of it because it can learn without explicit programming.[1] Because of the obvious relatively large amount of images and its impact in oncology, research in medical imaging has been one of the first to explore this new tool. Indeed, it is known that tumor phenotype is driven by its genotype and can be assessed by the multiple imaging modalities, morphologic and functional. ML will help medical imaging analysis for tumor detection, segmentation, characterization, treatment, and follow-up. Certainly, human assessment is precious and can evaluate different tumor, qualitative, and semiquantitative features (size, shape, calcifications, necrosis, etc.). These features are part of the medical lexicon and are called "semantic." However, semantic features are time-consuming and tend to be subjective and poorly reproducible. Therefore, their use remains limited. As opposed to semantic features, ML-based imaging biomarkers are quantitative, reproducible, and automatically measurable. With the emergence of molecular targeting therapeutics, we urgently need accurate tools to propose the most suited treatments. Furthermore, molecular expression in tumor can change under treatment. Multimodal imaging may help in monitoring molecular expression noninvasively, anticipating therapeutic failure, and leading therapeutic management.

To reach this objective, AI algorithms will help in identifying new image biomarkers, inaccessible to human eyes. These new

*IHU of Strasbourg, Strasbourg, France.
†Inserm & University of Strasbourg UMR-S1110, Strasbourg, France.
‡Faculty of Medicine, University of Paris, Paris, France.
§Icube Laboratory, University of Strasbourg, Strasbourg, France.
‖Faculty of Medicine, University of Strasbourg, Strasbourg, France.
¶Faculty of Medicine, McGill University, Montreal, Quebec, Canada.
Address reprint requests to Benoit Gallix, IHU of Strasbourg, 1 place de l'Hôpital, Strasbourg, 67000, France E-mail: Benoit.gallix@ihu-strasbourg.eu

https://doi.org/10.1053/j.semnuclmed.2020.07.003
0001-2998/

image biomarkers will reflect imaging tumor phenotype and indirectly its genotype. We aim to expose in this paper the state-of-the-art and future perspectives of AI driven by multimodality imaging.

Complex and Reproducible Imaging Biomarkers Beyond Visible

In the era of personalized treatments (molecular-targeted therapies, immunotherapy, locoregional therapy, etc.), the international consensus in oncology still advocates monitoring solid neoplasia, as varied as they are, according to the Response Evaluation Criteria in Solid Tumors (RECIST) criteria. These criteria are based on the sum of measurements of the longest axis of tumor lesions, chosen arbitrarily, without considering volume, shape, morphology, vascularity, and even less the internal structure reflecting cellular metabolism. If nuclear medicine may be an added value in the follow-up of solid neoplasia (Standardize Uptake Value), its use remains limited. This may be due to our inability to accurately characterize tumors and monitor intratumoral changes in imaging. We need new parameters, human-engineered or free, to go beyond the visible image and more sophisticated statistical analysis to choose the most impacting features. ML meets all these requirements. This approach consists of extracting patterns from a set of data in order to make predictions based on statistics (Fig. 1). In the case of medical image analysis, these patterns are called imaging biomarkers.[2] This extraction can be performed based on hand-crafted descriptive mathematical models or directly learned from the images without any human intervention. These imaging biomarkers can be used in two types of ML algorithms, the supervised and the nonsupervised.

Supervised learning consists in building a predictive model thanks to outputs already known and labeled by the physician. One of the main applications lies in classification issues, which in oncology and tumor characterization can include the prediction of tumor grade (or tumor differentiation), molecular expression, risk of recurrence, or even survival. Specific molecular expressions are impacting prognosis and treatment by inducing resistance profile or allowing targeted treatments (HER, KRAS, IDH,. . .). Different classifiers exist and can be more or less efficient according to the hypothesis-driven research. The most popular classifiers are random forests, support vector machines, and convolutional neural networks (CNN). On the contrary, unsupervised learning will apply to unlabeled data. The purpose of unsupervised

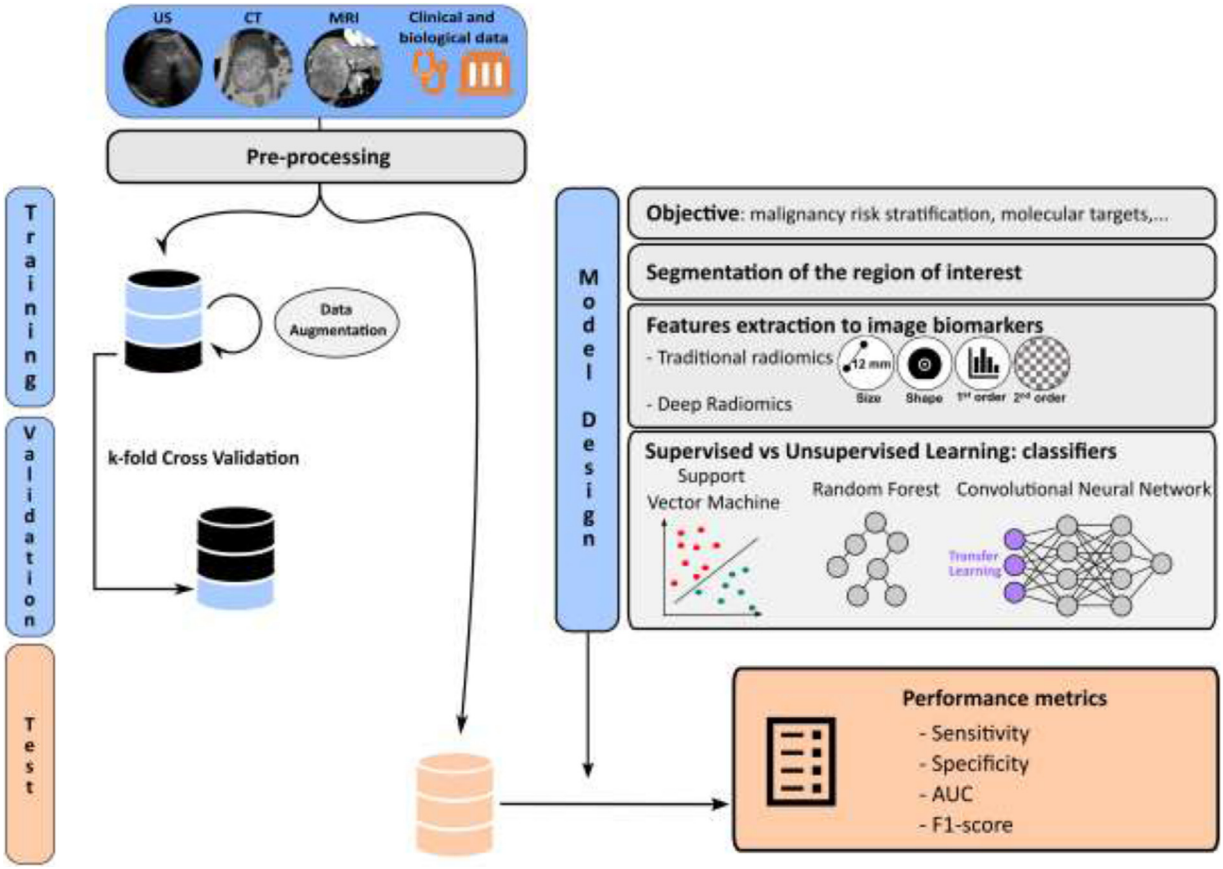


Figure 1 Artificial intelligence (machine learning and deep learning) processing according to quality standards. Clinical, biological, and imaging data should be divided into two strictly distinct sets: training/validation and test. Training/validation dataset should be used to design the model. Only then will the performance metrics of this AI model be evaluated on the test dataset. AUC: Area Under the Curve

learning will be to reveal new medical paradigms by identifying new structures in the data. A common application lies in clustering data and/or estimates its probability density. Unsupervised learning may help to group patients with different expressions of a same disease which can lead to a better understanding of it. There is a wide variety of algorithms, from the most classical K-means, through self-organizing maps to neural networks (auto-encoder).

Radiomics

Radiomics are noninvasive, reproducible, and automatically calculated quantitative features that are supposed to reflect the heterogeneity of the tumor phenotype and thus indirectly its genotype. Radiomics correspond to human-engineered and mathematically defined image descriptors either simple, such as size or shape, to more complex: first-order features based on intensity voxel histogram statistics, second-order or texture features (gray-level co-occurrence matrix, gray-level run-length matrix,...) reflecting the spatial relationship between voxel values or higher order statistics (fractals, wavelets,...) representing more complex patterns.

Recently, a new approach allowed us to extract new human-free image biomarkers, mathematically undefined upstream, and sometimes inaccessible to our understanding.[3] Some authors have even used the term "deep radiomics" by analogy to deep learning and the use of complex neural networks.[4] On the contrary of "traditional" handcrafted radiomics, they are free of human intervention and can identify new representations and the most informative properties of the image to solve research hypotheses. Deep learning is one of the aspects of ML using nonlinear transformations based on CNN imagined from the human neurons. Hidden layers are used to complexify the CNN model in order to extract and pool neural features with different levels of data abstraction. As the human brain, neural network can adjust its parameters to optimize its predictions by reducing the loss function (or error). This process is called back-propagation. The explainability of deep learning-based features is very limited for the moment.

How to Meet the Need of a Large Amount of Data?

Paradoxically, ML may be limited in medical imaging by the lack of data. Indeed, simple models such as linear models or support vector machines can be built with few parameters. They are therefore easier to learn and require less data. On the other hand, these models are often too simple to describe class distributions. This complexity can be learned with neural networks which contain millions of parameters. Their learning requires a large amount of data. These millions of parameters can exactly record the complete dataset, which is called overfitting (Fig. 2). It is therefore necessary during the training to make sure that there are enough data to avoid this phenomenon.

Several techniques exist to overcome the data limitation. First, a widespread computer science technique called "Data Augmentation" can be applied. This corresponds to the artificial creation of new data from the original dataset. The difficulty lies in respecting the original data. Especially in oncology and medical imaging, it can be particularly hazardous to create new data. Therefore, only simple geometric transformations of the image should be recommended. These can be rotated, mirrored, or translated. A second technique lies in the existence of common images features. A second type of data augmentation is the generation of synthetic data using neural networks known as Generative Adversarial Network.

Because deep learning requires large databases, it can be smart to pretrain a learning model. It is not necessary to use medical imaging databases and large databases of nature or animal images can be used. As an example, ImageNet, a publicly accessible database, is commonly used to pretrain deep neural networks. The re-use of this pretrained model is called transfer learning. The first learning layers will come from the pretraining, mimicking the visual primitive system, and this pretrained model will be re-trained, or fine-tuned, on the study database.

Validation and Testing

Two essential steps in the development of diagnostic and characterization algorithms are validation and testing (or external validation). The objective of these steps is to

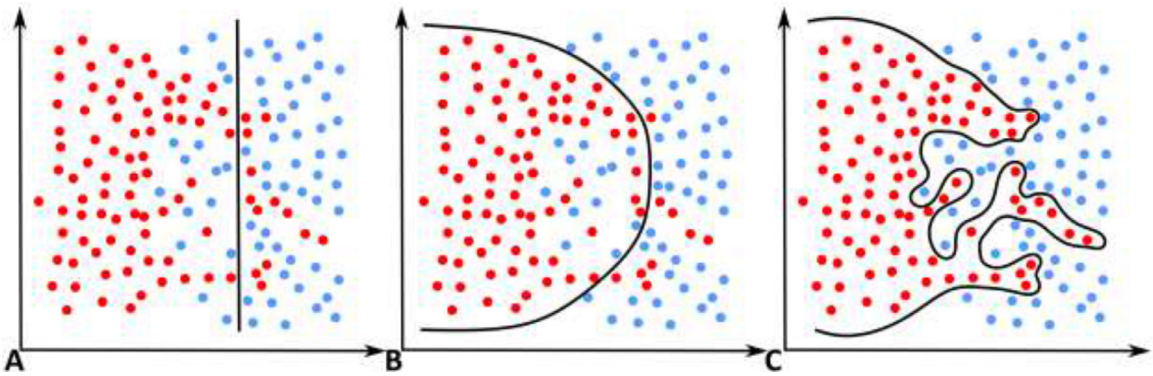


Figure 2 Illustration of different two-dimensional decision boundaries: from (A) too simple or underfitting through (B) well-balanced complexity to (C) too complex or overfitting.

optimize the training of a model, not to obtain the best diagnostic performance on the training dataset, but to allow a generalization of the model to patient populations different from the training population.

The validation step allows the parameters of the training model to be adjusted to avoid overfitting the training data. It is obvious that reaching high diagnostic performances on the training set alone is quite simple. At best, these performances will be overestimated. At worst, it will simply be wrong. In order not to sacrifice part of the training data at the validation step, it is common to use a technique called cross-validation. This technique consists in dividing the training set into k groups (usually 5 or 10), selecting one of these groups, training the model on the other k-1 groups and then validating on the selected group. This operation can be repeated k time. Especially in medicine, when creating these subgroups, it is necessary to ensure that percentages of classes are the same as that of the overall population.

The test step allows to evaluate the performances of the selected parameters. To do this, the dataset allocated to the test must be independent of the first dataset, ideally from an external database. It should never be used to train the model and it is therefore imperative to keep a strict separation between the training and test datasets. In a practical way, the model has to be locked before being tested on the external base without the possibility of a new learning iteration.

Tumor Characterization by Multimodality Imaging

Medical imaging is rich in its diversity (X-ray, CT scan, ultrasound, MRI, metabolic imaging) and each technique produces multiple and complementary image dataset. If the diversity and complementarity of its modalities have proven their importance in patient care, they are different information for human but also for ML. The analysis, during the same diagnostic or therapeutic process, of multiple imaging modalities repeated over time, is a difficult objective to achieve. To do so, harmonization between the imaging techniques must be reached.

Tumor progression over time and response to treatment by automated methods should provide valuable information. This unstructured temporal integration of data is a new dimension for ML.

Figure 3 illustrates the potential impacts of AI on patient management.

Malignancy Risk Assessment

First objective in oncology is usually to determine the benignity or malignancy of a lesion. It may be a very challenging task with sometimes limited answers and unnecessary follow-ups. As already mentioned, semantic features

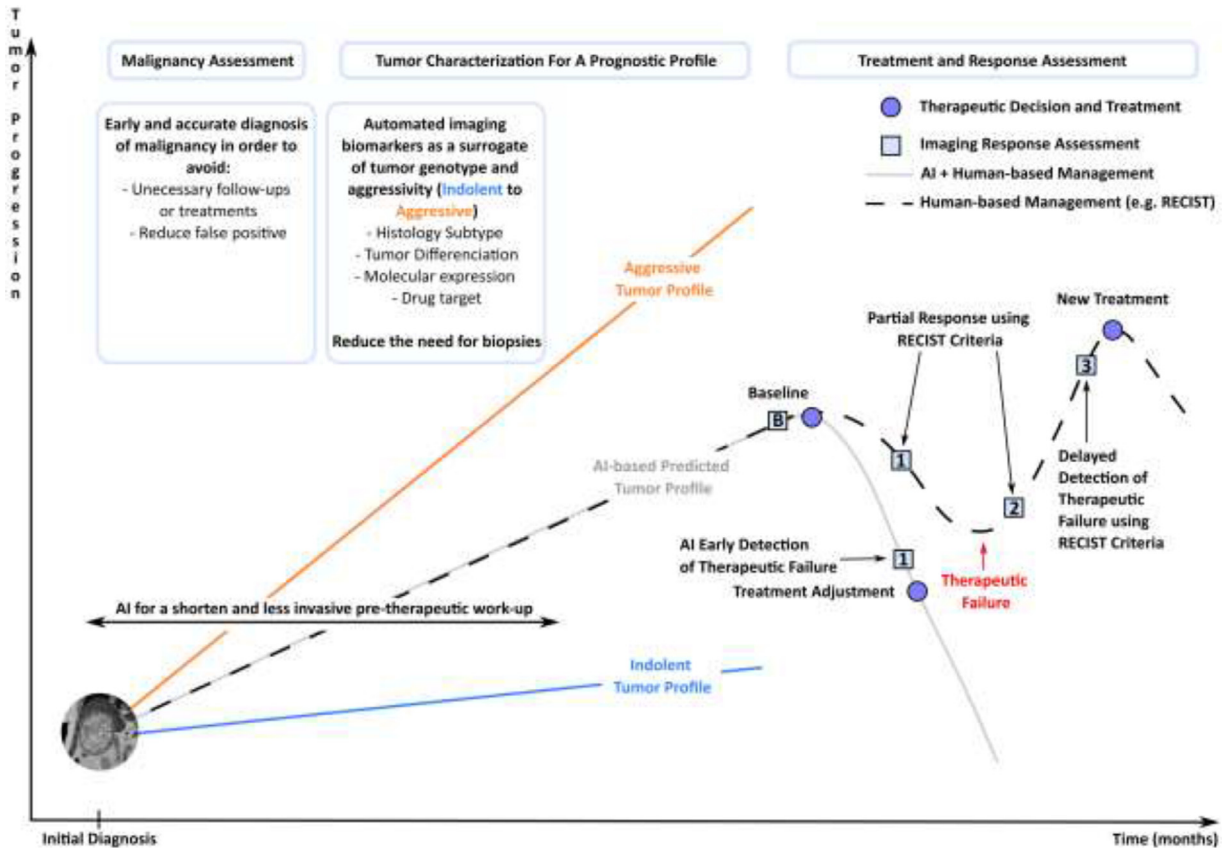


Figure 3 Tumor characterization by AI impacts oncology patient management. Compared to human-based management, AI can predict the tumor progression profile by stratifying its prognosis, optimize treatment modalities for better performance and allows early detection of bad and good responders to treatment.

are subjective and poorly reproducible leading to imperfect "Reporting and Data Systems" for assessing the risk of malignancy. Using "data augmentation," transfer learning and neural features extraction techniques from ultrasound images, Chi et al[5] proposed an algorithm based on a random forest classifier with higher performance than TI-RADS, a US-based widespread malignancy risk stratification system, in distinguishing benign from malignant thyroid nodules. Thoracic oncology has also been a major field of interest. Indeed, diagnostic issues are encountered for infra-centimetric pulmonary nodules while nonenhanced CT lung contrast offers interesting possibilities for ML. These diagnostic issues are encountered since metabolic imaging techniques such as PET scan are noncontributory and growth over time is the only diagnostic tool available. However, follow-up modalities and screening programs are still debated in terms of reduction of mortality.[6] In this context, Lui et al[7] proposed a semantic (density and margins) and radiomics features-based nomogram to assess the risk of malignancy of small lung nodules. Semantic features slightly improved the diagnostic performance of the radiomics signature. Finally, it can be noted that interest has also been shown in pancreatic intraductal papillary mucinous neoplasms[8] or in liver masses[9] to achieve the same objective of assessing the risk of malignancy.

It is interesting to highlight that, so far, many ML studies have focused on this objective without achieving sufficient clinical impact to be used in practice.

Tumor Characterization for Precision Oncology

Accurate noninvasive tumor characterization is the key for stratifying prognosis, predicting treatment response and optimizing patient management. It is supported by the proven assumption that imaging tumor phenotype is driven by its genotype and indirectly determines its evolution and our management.

It starts by defining the histologic subtype. This task was evaluated by Yasaka et al[9] between hepatocellular carcinoma and other liver malignancies using CT scan. As it is the case from the physician's point of view, MRI should be superior to CT scan to accomplish this task using ML. Hamm et al[10] developed a deep learning model based on multiphasic contrast-enhanced T1-weighted imaging to discriminate hepatocellular carcinoma, intrahepatic cholangiocarcinoma, colorectal metastasis, focal nodular hyperplasia, hemangioma, and cyst achieving an accuracy of 0.92. The diagnostic performance of this model outperformed two radiologists' review (respective accuracy of 0.80 and 0.85). Hepatocellular carcinoma was diagnosed by the model with a sensitivity of 0.94 and a specificity of 0.98.

Patient management can also start by the discovery of metastatic lesions without the knowledge of the primary neoplasia. If it is sometimes possible to guide the diagnostic investigations of the primary lesion, they are usually exhaustive due to the absence of diagnostic orientation leading to a delayed therapeutic management. In a retrospective study including patients with brain metastases secondary to breast, lung, gastrointestinal cancers, and melanoma, Kniep et al[11] designed a multiparametric MRI-based (T1-weighted contrast material-enhanced, T1-weighted nonenhanced, and fluid-attenuated inversion recovery) model to discriminate the tumor type of the brain metastases and guide the diagnostic investigations of the primary lesion. Using radiomics features and a random forest classifier, they achieved higher diagnostic performance than radiologist but still relatively modest with an Area Under the Curve (AUC) value of 0.64 for non−small-cell lung cancer. The highest diagnostic performance was observed for melanoma with an AUC value of 0.82 with a statistically significant difference compared to radiologist. The explication of such results for melanoma may lie in the presence of increased T1 signal areas (melanin) insufficiently identified by the radiologist. Nonenhanced T1 weighted imaging first order maximum was in the top 10 most important radiomics features for melanoma. This could reflect the superiority of automated and reproducible quantitative radiomics features over semantic features.

Second, it is known that prognosis is well correlated with tumor grade. Thus, rapidity of treatment initiation should take tumor grade into account. When biopsy is not yet available, anticipating tumor grade with noninvasive imaging could impact therapeutic management and, for example, indicate neoadjuvant therapy. It is also important to note that tissue sample biopsy may not appreciate certain higher grade focal areas within the tumor. ML may supplement this limitation. For example, if tumor grade of neuroendocrine pancreatic tumor is still defined according to ki-67 index, it can be well predicted by a ML nomogram.[12]

Whether treatment in oncology relies more and more on targeted therapies, pathology of tissue sample is still required to evaluate the expression of these targets within the tumor. Yet, tumor may be heterogeneous with different areas with distinct molecular characteristics. Knowing the genetic status of the tumor through a virtual biopsy could aid pretreatment decision-making. Genetic characterization of tumor by radiomics is called radiogenomics. Aerts et al were among the first to identify CT-based radiomics features in lung and head-and-neck carcinoma associated with the underlying gene-expression patterns by reflecting tumoral heterogeneity.[13] These radiogenomics features outperformed TNM classification for predicting survival and may impact therapeutic management. Similar results have been reported in non−small-cell lung cancer using a CT-based deep learning model in comparison with standard methods such as TNM classification.[14] In thoracic oncology, several molecular-targeted therapies exist such as tyrosine kinase inhibitor (TKI)-sensitive mutations of the epidermal growth factor receptor (EGFR), ALK, ROS1, or MET genes. Presence of TKI-sensitive mutations of EGFR has already been the focus of studies which have shown that it can be predicted with high accuracy by ML algorithms.[15] In Jia et al's study,[15] predicting TKI-sensitive mutations of EGFR using CT-based ML algorithms had benefited from clinical and biological data. When clinical features (sex and smoking history) were added to the model, diagnostic performance was slightly improved.

Thoracic oncology has not been the only field of interest. In Neuro-oncology, glioma can present with different molecular profiles impacting prognosis and driving patient management and treatments. Predicting tumor grade and mutational status of 1p19q, IDH1, MGMT, and ATRX may be achieved with ^{18}F-FET PET/MRI-based ML algorithms.[16]

Finally, from tumor characterization results a prognosis and a probability of response to treatment. In addition to prognosis, risk of local recurrence can also be appreciated by ML models such as in patients with hepatocellular carcinoma on cirrhosis after local treatment.[17]

Obviously, therapeutic choices also depend on the extension of the tumor, from local invasion to lymph node status. Bladder cancer prognosis clearly correlates with bladder muscle invasion requiring a radical cystectomy instead of a transurethral resection. Preoperative accurate assessment of muscular invasion would prevent under or overtreatment. Combining clinical with radiomics features from T2-weighted MRI, Zheng et al[18] developed a highly performant nomogram for the preoperative assessment of muscular invasiveness with an Area Under the Curve (AUC) value of 0.88. On another note, some authors demonstrated that noninvasive imaging can also accurately predict deep myometrial invasive and lympho-vascular space invasion of endometrial carcinoma using MRI-based ML algorithms.[19] The same is true for microvascular invasion in hepatocellular carcinoma, a difficult preoperative assessment, using contrast-enhanced US[20] or MRI-based[21,22] algorithms.

The preoperative prediction of lymph nodes status has also been a major field of interest as much in breast cancer using ultrasound[23] as in colorectal carcinoma with CT-scan.[24] While lymph node metastasis and extranodal extension may change operative planification or indicate adjuvant treatments in locally advanced cancer, preoperative assessment of extranodal extension remains poor. In response to this issue, Kann et al[25] developed and validated across different institutions, a deep radiomics CT-based algorithm achieving high performance in predicting extranodal extension in head-and-neck squamous cell carcinoma with an AUC of 0.84.

Response to Treatment

Prediction of response is another challenging objective. As in breast cancer, pathologic complete response to neoadjuvant chemotherapy is a major prognostic factor in oncology. Pretherapeutic prediction would be a significant added value for patient management. This task was performed by Li et al[26] focusing on tumor volume in a retrospective study of breast cancer patients prior to neoadjuvant chemotherapy. Peritumoral environment should also provide relevant prognostic information as it is illustrated by different immune score quantification developed in nonsquamous non−small-cell lung,[27] colon,[28] or gastric cancer.[29] Jiang et al designed a radiomics CT-based model predictor of the immuno-score of gastric cancer that was significantly associated with disease-free and overall survival.[30] Peritumoral T-cell immune environment targeting specific antigens at the surface of the tumor cells is also the key of the efficacity of immunotherapy. Sun et al designed a radiomics model to predict CD8 T-cell infiltration as an image biomarker for good response to immunotherapy.[31] A high baseline radiomics score was associated with improved overall survival. These results are consistent with other radiomics model developed in patients with metastatic melanoma and non−small-cell lung cancer treated by PD-L1 immunotherapy for predicting response.[32,33]

As explained above, RECIST criteria are still advocated despite its limitation. ML should help to detect treatment failure at an early stage by monitoring intratumoral changes reflecting genotypic modifications. In patients with unresectable hepatic metastases of colorectal cancer treated with FOLFIRI and bevacizumab, Dohan et al[34] developed a radiomics score for early prediction of good responders. It supplemented standard evaluation as it was able to predict a poor outcome at 2 months with the same performance as RECIST 1.1 at 6 months. Other authors reported interesting results for the prediction of complete pathologic response in triple-negative breast cancer at pretreatment MRI based on Kurtosis, a traditional radiomics feature.[35]

Furthermore, response evaluation of solid tumor under immunotherapy is particularly challenging in imaging with the concept of pseudo-progression that we are unable to differentiate from real progression, resulting in a delay in therapeutic management and the continuation of ineffective treatment. Tumor characterization for identifying pseudo-progression still needs to be studied.

Like the concept of pseudo-progression under immunotherapy, glioblastoma under Temozolomide, an alkylating agent, in association with radiotherapy may demonstrate pseudo-progression up to several weeks after the end of treatment. This pseudo-progression mimics true progression and diagnosis is usually made on spontaneous improvement or stabilization of imaging findings over several months. Therapeutic consequences can be important. According to Akbari et al, pseudo-progression has distinctive MRI-based radiomics features that could help for patient management.[36]

Radiation therapy may also cause radiation injury regardless of the underlying tumor type resulting in new contrast enhancement. Thus, differential diagnosis with tumor recurrence may be challenging and impacts patient management. Subject to the small number of patients (52) and the lack of external testing, Lohmann et al[37] support the contribution of radiomics features in distinguishing radiation injury from recurrent brain metastasis. The radiomics model using combined contrast-enhanced MRI and O-(2-[18F]fluoroethyl)-L-tyrosine PET-based features outperformed single-modality models (PET or MRI) reinforcing the interest of associating morphologic and functional imaging modalities.

Perspectives

To be an integral part of medical imaging and patient management, several challenges remain. As explained, medical imaging data remain rare for deep learning requirements.[38] As long as large and complex databases from different institutions are not available, “Data augmentation/generation” and “Transfer Learning” will serve as powerful tools but hardly compensate for the lack of data. Thus, it seems obvious that

the quality of a ML model depends on the training database. It is therefore essential to use high-quality images[39] and standardize imaging acquisition protocols which is even more necessary for multimodal algorithms. Also, training and validation databases must represent the full spectrum of the disease in order to make the algorithms generalizable and robust. The evaluation of the diagnostic performance of the model must therefore be carried out in an external and independent population to ensure this generalizability. However, an external and independent test of the model is frequently missing. Second, ML algorithms should be reproducible. Authors should use open source code packages (eg, Pyradiomics) to standardize the extraction of data from medical images. Other research institutions should be able to reproduce any of the published ML models. The trained network or a network of identical architecture with the same training database and the same initialization parameters should be shared. Unfortunately, deep learning algorithms are not easily reproducible. Unlike traditional radiomics features, mathematically defined, deep radiomics features are usually represented by the concept of "black box." They result from nonlinear transformations and the deeper the convolutional neural network, the more difficult it becomes to interpret the deep radiomics from a physio-pathologic point of view. A final condition for bringing ML into tomorrow's medicine is to reliably automate the segmentation of lesions. Segmentation using CNN (as U-net or cascaded architecture) are already proposed.[40] Most published studies rely on manual segmentation, implicating interobserver variability or, at best, semiautomatic segmentation.

To meet reference quality standards, Lambin et al[41] proposed a radiomics quality score. In a recent review of radiomics in hepatocellular carcinoma using this score, all studies but one were scored below 18/36 (50%). Main reasons were the retrospective design, the lack of validation, and open-access scientific data resources. To guide authors, a Checklist for Artificial Intelligence in Medical Imaging (CLAIM) has also been proposed.[42]

On another note, ML would benefit from an exhaustive exploitation of multimodal imaging techniques. Indeed, ultrasound remains the least studied imaging technique. A few reasons can be advanced. The main limitation lies in the complete absence of standardized acquisition. However, ultrasound brings real-time kinetics, elastography, or Doppler data. Regarding the published MRI-based algorithms, they usually do not associate different sequences losing multiple information on tissue characterization. Combining metabolic and molecular imaging modalities (such as ^{18}FET PET scanner) with conventional CT scan or MRI will also contribute to tumor characterization and better understanding of underlying molecular mechanisms. As an example, comparison of metabolic imaging with diffusion-weighted MRI is already useful for assessing cell density and proliferation. Furthermore, ML will learn from the different metabolic radiotracers reflecting intratumoral metabolism and heterogeneity.

At the difference of physicians who benefit from the interpretation of previous examinations, no studies have included temporality in the ML algorithms. However, changes in size, shape, limitation, enhancement, and heterogeneity are crucial information for tumor characterization.

In our opinion, ML will mostly prove helpful in the assessment of tumor response. Early detection of nonresponse is crucial to rapidly adapt therapeutic management and propose new treatments. The tumor phenotype, driven by the tumor genome, and molecular expression define the indications for targeted treatments and immunotherapy. As these tumor characteristics can change under treatment, ML can be used to monitor the expression of tumor targets, detect phenotypic changes, and thus adapt treatments early. The contribution of medical imaging to personalized medicine will rely largely on the automation of image analysis through ML methods.

Conclusion

For nearly 30 years, multimodal cross-sectional imaging has been attempting to design reproducible and high-performance biomarkers. Only the simplest imaging biomarkers (RECIST criteria, SUV,. . .) have become widespread clinical tools. The advent of AI brings new paradigms by identifying structural patterns across large and heterogeneous data. However, automated image analysis using these new biomarkers will only become part of the clinical practice under several conditions: standardization of imaging acquisition as it has been done in the field of biology, transparency of the model development as it should be reproducible by different institutions, validation and testing through a high-quality process using large and complex open databases, and better interpretability of these algorithms.

This work was supported by the French Government research program « Investissements d'avenir » managed by "Agence Nationale de la Recherche" [ANR-10-IAHU-02]

References

1. Hosny A, Parmar C, Quackenbush J, et al: Artificial intelligence in radiology. Nat Rev Cancer 18:500-510, 2018. https://doi.org/10.1038/s41568-018-0016-5
2. Savadjiev P, Chong J, Dohan A, et al: Image-based biomarkers for solid tumor quantification. Eur Radiol 2019. https://doi.org/10.1007/s00330-019-06169-w
3. Chartrand G, Cheng PM, Vorontsov E, et al: Deep learning: A primer for radiologists. Radiographics 37:2113-2131, 2017. https://doi.org/10.1148/rg.2017170077
4. LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 521:436-444, 2015. https://doi.org/10.1038/nature14539
5. Chi J, Walia E, Babyn P, et al: Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. J Digit Imaging 30:477-486, 2017. https://doi.org/10.1007/s10278-017-9997-y
6. Oudkerk M, Devaraj A, Vliegenthart R, et al: European position statement on lung cancer screening. Lancet Oncol 18:e754-e766, 2017. https://doi.org/10.1016/S1470-2045(17)30861-6
7. Liu Q, Huang Y, Chen H, et al: The development and validation of a radiomic nomogram for the preoperative prediction of lung adenocarcinoma. BMC Cancer 20:533, 2020. https://doi.org/10.1186/s12885-020-07017-7

8. Hanania AN, Bantis LE, Feng Z, et al: Quantitative imaging to evaluate malignant potential of IPMNs. Oncotarget 7:85776-85784, 2016. https://doi.org/10.18632/oncotarget.11769
9. Yasaka K, Akai H, Abe O, et al: Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: A preliminary study. Radiology 286:887-896, 2017. https://doi.org/10.1148/radiol.2017170706
10. Hamm CA, Wang CJ, Savic LJ, et al: Deep learning for liver tumor diagnosis part I: Development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 29:3338-3347, 2019. https://doi.org/10.1007/s00330-019-06205-9
11. Kniep HC, Madesta F, Schneider T, et al: Radiomics of brain MRI: Utility in prediction of metastatic tumor type. Radiology 290:479-487, 2019. https://doi.org/10.1148/radiol.2018180946
12. Liang W, Yang P, Huang R, et al: A combined nomogram model to preoperatively predict histologic grade in pancreatic neuroendocrine tumors. Clin Cancer Res 25:584-594, 2019. https://doi.org/10.1158/1078-0432.CCR-18-1305
13. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:4006, 2014. https://doi.org/10.1038/ncomms5006
14. Hosny A, Parmar C, Coroller TP, et al: Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med 15:e1002711. https://doi.org/10.1371/journal.pmed.1002711, 2018
15. Jia T-Y, Xiong J-F, Li X-Y, et al: Identifying EGFR mutations in lung adenocarcinoma by noninvasive imaging using radiomics features and random forest modeling. Eur Radiol 29:4742-4750, 2019. https://doi.org/10.1007/s00330-019-06024-y
16. Haubold J, Demircioglu A, Gratz M, et al: Non-invasive tumor decoding and phenotyping of cerebral gliomas utilizing multiparametric 18F-FET PET-MRI and MR fingerprinting. Eur J Nucl Med Mol Imaging 47:1435-1445, 2020. https://doi.org/10.1007/s00259-019-04602-2
17. Ji G-W, Zhu F-P, Xu Q, et al: Radiomic features at contrast-enhanced CT predict recurrence in early stage hepatocellular carcinoma: A multi-institutional study. Radiology 294:568-579, 2020. https://doi.org/10.1148/radiol.2020191470
18. Zheng J, Kong J, Wu S, et al: Development of a noninvasive tool to preoperatively evaluate the muscular invasiveness of bladder cancer using a radiomics approach. Cancer 125:4388-4398, 2019. https://doi.org/10.1002/cncr.32490
19. Ueno Y, Forghani B, Forghani R, et al: Endometrial carcinoma: MR imaging-based texture model for preoperative risk stratification—A preliminary analysis. Radiology 284:748-757, 2017. https://doi.org/10.1148/radiol.2017161950
20. Hu H-T, Wang Z, Huang X-W, et al: Ultrasound-based radiomics score: A potential biomarker for the prediction of microvascular invasion in hepatocellular carcinoma. Eur Radiol 29:2890-2901, 2019. https://doi.org/10.1007/s00330-018-5797-0
21. Banerjee S, Wang DS, Kim HJ, et al: A computed tomography radiogenomic biomarker predicts microvascular invasion and clinical outcomes in hepatocellular carcinoma. Hepatology 62:792-800, 2015. https://doi.org/10.1002/hep.27877
22. Xu X, Zhang H-L, Liu Q-P, et al: Radiomic analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma. J Hepatol 70:1133-1144, 2019. https://doi.org/10.1016/j.jhep.2019.02.023
23. Zheng X, Yao Z, Huang Y, et al: Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nat Commun 11:1236, 2020. https://doi.org/10.1038/s41467-020-15027-z
24. Huang Y-Q, Liang C-H, He L, et al: Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. J Clin Oncol 34:2157-2164, 2016. https://doi.org/10.1200/JCO.2015.65.9128
25. Kann BH, Hicks DF, Payabvash S, et al: Multi-institutional validation of deep learning for pretreatment identification of extranodal extension in head and neck squamous cell carcinoma. J Clin Oncol 38:1304-1311, 2020. https://doi.org/10.1200/JCO.19.02031
26. Li P, Wang X, Xu C, et al: 18F-FDG PET/CT radiomic predictors of pathologic complete response (pCR) to neoadjuvant chemotherapy in breast cancer patients. Eur J Nucl Med Mol Imaging 47:1116-1126, 2020. https://doi.org/10.1007/s00259-020-04684-3
27. Li B, Cui Y, Diehn M, et al: Development and validation of an individualized immune prognostic signature in early-stage nonsquamous non−small cell lung cancer. JAMA Oncol 3:1529-1537, 2017. https://doi.org/10.1001/jamaoncol.2017.1609
28. Pagès F, Mlecnik B, Marliot F, et al: International validation of the consensus immunoscore for the classification of colon cancer: A prognostic and accuracy study. Lancet 391:2128-2139, 2018. https://doi.org/10.1016/S0140-6736(18)30789-X
29. Jiang Y, Zhang Q, Hu Y, et al: ImmunoScore signature: A prognostic and predictive tool in gastric cancer. Ann Surg 267:504-513, 2018. https://doi.org/10.1097/SLA.0000000000002116
30. Jiang Y, Wang H, Wu J, et al: Noninvasive imaging evaluation of tumor immune microenvironment to predict outcomes in gastric cancer. Ann Oncol 31:760-768, 2020. https://doi.org/10.1016/j.annonc.2020.03.295
31. Sun R, Limkin EJ, Vakalopoulou M, et al: A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: An imaging biomarker, retrospective multicohort study. Lancet Oncol 19:1180-1191, 2018. https://doi.org/10.1016/S1470-2045(18)30413-3
32. Trebeschi S, Drago SG, Birkbak NJ, et al: Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. Ann Oncol 30:998-1004, 2019. https://doi.org/10.1093/annonc/mdz108
33. Mu W, Tunali I, Gray JE, et al: Radiomics of 18F-FDG PET/CT images predicts clinical benefit of advanced NSCLC patients to checkpoint blockade immunotherapy. Eur J Nucl Med Mol Imaging 47:1168-1182, 2020. https://doi.org/10.1007/s00259-019-04625-9
34. Dohan A, Gallix B, Guiu B, et al: Early evaluation using a radiomic signature of unresectable hepatic metastases to predict outcome in patients with colorectal cancer treated with FOLFIRI and bevacizumab. Gut 69:531-539, 2020. https://doi.org/10.1136/gutjnl-2018-316407
35. Chamming's F, Ueno Y, Ferré R, et al: Features from computerized texture analysis of breast cancers at pretreatment MR imaging are associated with response to neoadjuvant chemotherapy. Radiology 286:412-420, 2018. https://doi.org/10.1148/radiol.2017170143
36. Akbari H, Rathore S, Bakas S, et al: Histopathology-validated machine learning radiographic biomarker for noninvasive discrimination between true progression and pseudo-progression in glioblastoma. Cancer 126:2625-2636, 2020. https://doi.org/10.1002/cncr.32790
37. Lohmann P, Kocher M, Ceccon G, et al: Combined FET PET/MRI radiomics differentiates radiation injury from recurrent brain metastasis. Neuroimage Clin 20:537-542, 2018. https://doi.org/10.1016/j.nicl.2018.08.024
38. Savadjiev P, Chong J, Dohan A, et al: Demystification of AI-driven medical image interpretation: Past, present and future. Eur Radiol 29:1616-1624, 2019. https://doi.org/10.1007/s00330-018-5674-x
39. Sabottke CF, Spieler BM: The effect of image resolution on deep learning in radiography. Radiology 2:e190015. https://doi.org/10.1148/ryai.2019190015, 2020
40. Ouhmich F, Agnus V, Noblet V, et al: Liver tissue segmentation in multiphase CT scans using cascaded convolutional neural networks. Int J Comput Assist Radiol Surg 14:1275-1284, 2019. https://doi.org/10.1007/s11548-019-01989-z
41. Lambin P, Leijenaar RTH, Deist TM, et al: Radiomics: The bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749-762, 2017. https://doi.org/10.1038/nrclinonc.2017.141
42. Mongan J, Moy L, Kahn CE: Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. Radiology 2:e200029. https://doi.org/10.1148/ryai.2020200029, 2020

PROGNOSTIC STRATIFICATION IN EARLY-STAGE HEPATOCELLULAR CARCINOMA: IMAGING BIOMARKERS ARE NEEDED

In this editorial published in Liver International, I discussed the crucial importance of developing prognosis stratification biomarkers that could be used to optimize the curative treatment approach in hepatocellular carcinoma and to identify patients with early-stage hepatocellular carcinoma that might benefit from adjuvant therapies after tumour ablation. The main challenge of preliminary studies developing such tools is obviously clinical transferability. To be implemented in the clinical routine, the computation and use of imaging-based biomarkers must be as simple as possible. It is unrealistic to think that, with the increasing number of imaging studies, radiologists will have the time to make manual annotations on images. Imaging-based biomarkers should be just a click away. This requires certain technical issues to be resolved, such as the misregistration of sequences mainly due to inconsistent breathing, or automated segmentation of the tumour.

Received: 31 January 2024 | Accepted: 2 February 2024

DOI: 10.1111/liv.15869

Liver INTERNATIONAL WILEY

EDITORIAL

Prognostic stratification in early-stage hepatocellular carcinoma: Imaging biomarkers are needed

Early-stage hepatocellular carcinoma (HCC), defined as a single tumour (stage 0 if ≤2 cm; stage A if >2 cm) or ≤3 tumours ≤3 cm (stage A) as per the Barcelona Clinic Liver Cancer (BCLC) classification,[1] is eligible for curative treatment including liver transplantation, surgical resection, and percutaneous ablation. Although percutaneous ablation has been shown to offer overall survival rates comparable to those of surgical resection for small lesions, the latter remains the standard therapeutic option for larger lesions (>3 cm) provided that the patient is a good surgical candidate.[2,3] When both surgical resection and percutaneous ablation are feasible, the choice between the two possibilities mainly relies on liver function, portal pressure, age, comorbidities, and local expertise and, except for alpha-fetoprotein level, no factors related to the tumour aggressiveness are routinely taken into account. For obvious reasons, the presence of microvascular invasion cannot be currently assessed in tumours treated by ablation. However, tumour differentiation is accessible by percutaneous biopsy, which should reinforce its systematic performance. It is therefore urgent to develop pretherapeutic biomarkers to capture the heterogeneity of early-stage HCC, risk stratify their prognosis to personalize the treatment and identify those that could potentially benefit from adjuvant treatment. In this issue of *Liver International*, Wang et al. aimed to develop a state-of-the-art deep learning model to predict microvascular invasion (MVI) using pre-treatment magnetic resonance imaging (MRI) in patients with solitary tumours ≤3 cm.[4] The originality of this work was to use a cohort of patients with surgically resected HCC to train the deep learning model to predict the presence of MVI before testing the model in a cohort of patients with ablated HCC with recurrence-free survival and overall survival as the primary outcomes.

With a very large dataset of 696 patients and a limited imbalance between positive (28.5%) and negative cases of MVI to train the model, the latter achieved high performance in the validation cohort of surgically resected HCC (AUC of .901 and .816 in BCLC A and 0 HCC, respectively). Interestingly, when tested in the ablation cohort, the imbalance between cases at high risk of MVI and those at low risk was similar (30.6%) with no significant difference in size between HCCs with high risk of MVI and those without. In the ablation cohort, the recurrence-free survival rates of patients with high MVI risk were 57.1% at 1 year, 30.7% at 2 years, 13.1% at 3 years, and 2.6% at 5 years, which were significantly lower than those of patients predicted without MVI (87.8% at 1 year, 80.4% at 2 years, 71.3% at 3 years, and 56.0% at 5 years, $p < .001$). The 1-, 3-, and 5-year overall survival rates were 90.9%, 68.2%, and 49.1% for patients with high MVI risk, which were also significantly lower than 98.4%, 92.2%, and 81.5% for patients predicted without MVI, respectively ($p < .001$). Using a stepwise multivariate Cox regression analysis, this MVI biomarker was shown to be an independent risk factor for a lower recurrence-free survival rate, in addition to alpha-fetoprotein >20 ng/mL and unfavourable tumour location. The latter is most likely explained by technical difficulties and the heat-sink effect as it was the only independent risk factor for local tumour progression. Interestingly, the MVI biomarker was significantly associated with intrahepatic distance recurrence (32.3% vs. 9.6% at 1 year, 57.3% vs. 13.0% at 2 years, 71.0% vs. 20.1% at 3 years, 76.3% vs. 32.8% at 5 years), which could reinforce its clinical relevance as a companion biomarker of adjuvant therapies. If no significant risk factor was found for extrahepatic metastasis, this can be explained by the extremely limited number of positive cases (5/180). The overall C-index of the multivariate Cox regression model for evaluating recurrence-free survival was .73.

In addition to the large size of the datasets and the robustness of the deep learning methodology, another strength reinforcing the generalisability of the model is the relative heterogeneity of the included MR images. Although all MR scanners in the training centre were from the same vendor, the acquisition parameters of the different MRI sequences were heterogeneous enough to ensure the generalisability of the model to another dataset of MR images from a different vendor. Furthermore, this study reinforces the importance of exploiting the full potential of MRI to capture all the tumour specificities. In the same way, radiologists analyse HCCs from all MRI sequences, the deep learning model developed in this study performed better when integrating a multiphase approach than a single-phase approach (AUC of .883 vs. .685–.763 in the validation cohort).

Clinical transferability is the main future challenge of this study, which can also be stated for most of the artificial intelligence studies. To be implemented in the clinical routine, the computation and use of imaging-based biomarkers must be as simple as possible. It is unrealistic to think that, with the increasing number of imaging studies, radiologists will have the time to make manual annotations on all the MRI phases. Imaging-based biomarkers should be just a click away. This requires certain technical issues to be resolved, such as the misregistration of sequences mainly due to inconsistent breathing, or automated

segmentation of the tumour. It should also be remembered that one objective of developing this type of biomarker is to provide prognosis stratification for the accurate selection of patients with early-stage HCC who will most benefit from adjuvant therapies, therefore avoiding side effects if no benefit is expected, to achieve the best clinical outcome/cost ratio. Therefore, although relevant from a computational perspective, the best clinical threshold of stratification biomarkers may not be found by maximizing sensitivity and specificity as it is done in this study. Finally, a classic limitation to the applicability of such deep learning models to populations with different epidemiology is the high prevalence of patients with chronic hepatitis B virus (93.0%–96.1%) and the absence of cirrhosis in almost half of the patients in the training cohort. It can be noted that a major clinical difference between the ablation cohort and the surgical training cohort was the higher prevalence of cirrhosis in the ablation cohort (72.2% vs. 44.8%). Nevertheless, it is usual to find such differences as there is a systematic selection bias in patients treated with ablation (older age, higher total bilirubin, lower albumin, etc.).

To return to everyday clinical challenges, the stratification of patients' prognosis at the initial diagnosis of their tumour remains an unmet need. For instance, the deep learning model proposed here, based on imaging features of tumour aggressiveness, has the potential to assist physicians in choosing the most appropriate treatment based on the predicted MVI risk. Indeed, surgery may be a more appropriate treatment for small HCCs (<3 cm) in surgical candidates than ablation if MVI is likely to be present around the tumour. Additionally, for nonsurgical candidates or unresectable tumours, detecting features of MVI on pre-therapeutic imaging could lead to more aggressive interventional radiology treatments (such as multipolar ablation or combined ablation + transarterial chemoembolization) to ensure wider ablation margins and reduce the risk of loco-regional recurrence (Figure 1).

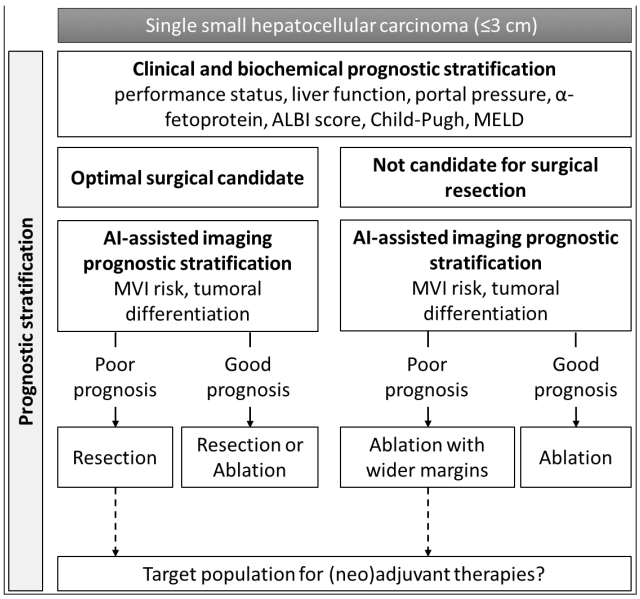


FIGURE 1 Proposal of an artificial intelligence-assisted management algorithm for single small hepatocellular carcinoma (≤3 cm). AI, artificial intelligence; MVI, microvascular invasion.

However, both curative-intent treatments are associated with a high risk of intrahepatic or distant recurrence, reported up to 70% overall at 5 years.[2,5] In this context, multiple clinical trials have been conducted to evaluate the impact of local (i.e., transarterial chemoembolization) and systemic (i.e., sorafenib or more recently immunotherapies) adjuvant therapies on progression-free survival and overall survival.[6] Recently, an interim analysis of the randomized phase III clinical trial (IMbrave050) of adjuvant atezolizumab + bevacizumab for patients at high risk of recurrence following resection or ablation has demonstrated a significant improvement in recurrence-free survival.[7] The high risk of recurrence was well defined in a surgical cohort based on the size, number of resected tumours, the poor differentiation of the tumour and the presence of micro- or macrovascular invasion on the surgical specimen. Interestingly, in the subgroup of patients treated by ablation, the benefit of the adjuvant treatment was less clear, and the criteria to define the high risk of recurrence were different, relying solely on tumour size and number. Pending the results of other ongoing phase II/III trials, no international recommendation currently endorses the use of (neo)adjuvant treatment in the setting of percutaneous ablation especially for single small tumours which is the specific population studied in the Wang et al. study. Hence, these approaches based on imaging biomarkers are relevant and can help identify a target population for future studies where adjuvant therapies would be tested in the context of percutaneous ablation.

In conclusion, this study is a promising step forward in developing prognosis stratification biomarkers that could be used to optimize the curative treatment approach and to identify patients with early-stage HCC that might benefit from adjuvant therapies after tumour ablation. Prospective clinical trials are needed to refine and test this imaging-based biomarker before it can be implemented in clinical practice.

CONFLICT OF INTEREST STATEMENT

The authors do not have any disclosures to report.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Jérémy Dana[1,2,3,4]
Olivier Sutter[5,6]

[1]*Department of Diagnostic Radiology, McGill University Health Centre, Montreal, Quebec, Canada*
[2]*Augmented Intelligence & Precision Health Laboratory (AIPHL), McGill University Health Centre Research Institute, Montreal, Quebec, Canada*
[3]*Institut Hospitalo-Universitaire (IHU) Strasbourg, Université de Strasbourg, Strasbourg, France*
[4]*Inserm U1110, Institut de Recherche sur les Maladies Virales et Hépatiques, Université de Strasbourg, Strasbourg, France*
[5]*Interventional Radiology Unit, Hôpital Avicenne, Hôpitaux*

Universitaires Paris Seine-Saint-Denis, APHP, Bobigny, France
[6]*Team MONC, Inria, CNRS UMR 5251, Bordeaux INP, Université de Bordeaux, Bordeaux, France*

Correspondence
Jérémy Dana, Department of Diagnostic Radiology, McGill University Health Centre, 1001 Décarie Blvd, Montreal, QC H4A 3J1, Canada.
Email: jeremy.dana@mail.mcgill.ca

ORCID

Jérémy Dana https://orcid.org/0000-0003-0352-4128

REFERENCES

1. Reig M, Forner A, Rimola J, et al. BCLC strategy for prognosis prediction and treatment recommendation: the 2022 update. *J Hepatol*. 2022;76(3):681-693.
2. Takayama T, Hasegawa K, Izumi N, et al. Surgery versus radiofrequency ablation for small hepatocellular carcinoma: a randomized controlled trial (SURF trial). *Liver Cancer*. 2022;11(3):209-218.
3. Mohkam K, Dumont PN, Manichon AF, et al. No-touch multibipolar radiofrequency ablation vs. surgical resection for solitary hepatocellular carcinoma ranging from 2 to 5 cm. *J Hepatol*. 2018;68(6):1172-1180.
4. Wang W, Wang Y, Song D, et al. A transformer-based microvascular invasion classifier enhances prognostic stratification in HCC following radiofrequency ablation. *Liver Int*. 2024. doi:10.1111/liv.15846
5. Doyle A, Gorgen A, Muaddi H, et al. Outcomes of radiofrequency ablation as first-line therapy for hepatocellular carcinoma less than 3 cm in potentially transplantable patients. *J Hepatol*. 2019;70(5):866-873.
6. Nahon P, Vibert E, Nault JC, Ganne-Carrie N, Ziol M, Seror O. Optimizing curative management of hepatocellular carcinoma. *Liver Int*. 2020;40(suppl 1):109-115.
7. Qin S, Chen M, Cheng A-L, et al. Atezolizumab plus bevacizumab versus active surveillance in patients with resected or ablated high-risk hepatocellular carcinoma (IMbrave050): a randomised, open-label, multicentre, phase 3 trial. *Lancet*. 2023;402(10415):1835-1847.

R2* IMPACT ON HEPATIC STEATOSIS QUANTIFICATION WITH A COMMERCIAL SINGLE VOXEL TECHNIQUE AT 1.5 AND 3T

Over the last decades, the prevalence of chronic liver diseases and their associated morbidity and mortality markedly increased, especially with the rise of metabolic associated steatotic liver disease (MASLD). Steatotic liver disease is a liver condition characterised by abnormal fat accumulation in more than 5% of hepatocytes. It is presumed to represent the most prevalent liver disease worldwide due to its association the metabolic syndrome, obesity and type 2 diabetes. However, as highlighted in the most recent multi-society consensus statement on new fatty liver disease nomenclature[74], SLD does not only refer to MASLD but encompasses the whole spectrum of causes of hepatic steatosis from alcohol-associated liver disease to cryptogenic causes. Therefore, it is critical to accurately quantify hepatic steatosis in a non-invasive approach. Multiple non-invasive methods have been developed to evaluate SLD such as transient elastography and ultrasound. The current non-invasive gold standard is magnetic resonance (MR)-based multi-echo Dixon. This technique measures the proton density fat fraction (PDFF), corresponding to the ratio of the fat protons signal to the signal of water and fat protons. PDFF has shown an excellent correlation with fat content from biopsies. However, Dixon-based techniques are limited by fat-water swaps, where the fat signal is incorrectly assigned to the water signal and vice versa, leading to inaccurate quantification. A previous study investigated the prevalence of fat-water swaps and showed that 8% of cases suffered from fat-water swaps at 3.0 T.

MR spectroscopy measures the chemical composition of fat and is an alternative method to multi-echo Dixon PDFF. However, MR spectroscopy techniques offer limited spatial coverage and are sensitive to T2-signal decay, leading to inaccurate quantification. Nonetheless, MR spectroscopy is useful for fat quantification as it does not suffer from fat-water swaps.

In this study published in the Canadian Association of Radiology Journal (impact factor in 2002 of 3.1), we aimed to validate the commercial HISTO fat quantification accuracy at 3.0 T using multi-echo Dixon as the reference method, to establish its robustness to R2* variations, and to compare the results to measurements performed at 1.5 T.

Original Research

R_2* Impact on Hepatic Fat Quantification With a Commercial Single Voxel Technique at 1.5 and 3.0 T

Canadian Association of Radiologists Journal
1–9

DOI: 10.1177/08465371241255896
journals.sagepub.com/home/caj

Véronique Fortier[1,2,3,4], Ahmed Mohamed[5], Evan McNabb[1], Jérémy Dana[1], Rita Zakarian[6], Ives R. Levesque[3,4,6], and Caroline Reinhold[1,2,6,7]

Abstract

Rationale and Objectives: Fat quantification accuracy using a commercial single-voxel high speed T_2-corrected multi-echo (HISTO) technique and its robustness to R_2* variations at 3.0 T, such as those introduced by iron in liver, has not been fully established. This study evaluated HISTO at 3.0 T and sought to reproduce results at 1.5 T. **Methods:** Phantoms were prepared with a range of fat content and R_2*. Data were acquired at 1.5 T and 3.0 T, using HISTO and a Dixon technique. Fat quantification accuracy was evaluated as a function of R_2*. The patient study included 239 consecutive patients. Data were acquired at 1.5 T or 3.0 T, using HISTO and Dixon techniques. The techniques were compared using Bland-Altman plots. Bias significance was evaluated using a one-sample *t*-test. **Results:** In phantoms, HISTO was accurate within 10% up to a R_2* of 100 s^{-1} at both field strengths, while Dixon was accurate within 10% where R_2* was accurately quantified (up to 350 s^{-1} at 1.5 T, and 550 s^{-1} at 3.0 T). In patients, where R_2* was $<100\ s^{-1}$, fat quantification from both techniques agreed at 1.5 T ($P = .71$), but not at 3.0 T ($P = .007$), with a bias $<1\%$. **Conclusion:** Results suggest that HISTO is reliable when R_2* is $<100\ s^{-1}$, corresponding to patients with at most mild liver iron overload, and that it should be used with caution when R_2* is $>100\ s^{-1}$. Dixon should be preferred for hepatic fat quantification due to its robustness to R_2* variations.

Résumé

Justificatif et objectifs : La précision de la mesure de tissu adipeux au moyen d'une technique multiécho à haute vitesse et à correction t2 pour un seul voxel (HISTO) et sa robustesse aux variations R_2* à 3,0 T, telles que celles causées par le fer intrahépatique, n'a pas été établie. Cette étude a évalué HISTO à 3,0 T et a cherché à reproduire les résultats à 1,5 T. **Méthodes :** Des fantômes ont été préparés avec un large éventail de contenu en graisse et R_2*. Les données ont été acquises à 1,5 T et à 3,0 T en utilisant HISTO et une technique Dixon. La précision de la quantification de graisse a été évaluée sous forme de fonction de R_2*. L'étude sur les patients a inclus 239 patients consécutifs. Les données ont été acquises à 1,5 T et à 3,0 T au moyen des techniques HISTO et Dixon. Les techniques ont été comparées au moyen des graphiques de Bland-Altman. La signification d'un biais a été évaluée en utilisant un test *t* à un échantillon. **Résultats :** Dans les fantômes, HISTO a été exact dans une limite de 10 % et jusqu'à un R_2* de 100 s-1 pour les deux forces de champ, tandis que la technique Dixon a été exacte dans une limite de 10 % où R_2* était précisément quantifié (jusqu'à 350 s-1 à 1,5 T et 550 s-1 à 3,0 T). Chez les patients, quand R_2* était <100 s-1, la mesure de graisse avec les deux techniques concordait à 1,5 T ($P = 0{,}71$), mais pas à 3,0 T ($P = 0{,}007$), et il y avait présence d'un biais $< 1\%$. **Conclusion :** Les résultats semblent indiquer que la technique

[1] Department of Medical Imaging, McGill University Health Centre, Montreal, QC, Canada
[2] Diagnostic Radiology, McGill University, Montreal, QC, Canada
[3] Gerald Bronfman Department of Oncology, McGill University, Montreal, QC, Canada
[4] Medical Physics Unit, McGill University, Montreal, QC, Canada
[5] Radiology Department, National Cancer Institute, Cairo University, Cairo, Egypt
[6] Research Institute of the McGill University Health Centre, Montreal, QC, Canada
[7] Montreal Imaging Experts Inc., Montreal, QC, Canada

Corresponding Author:
Véronique Fortier, Department of Medical Imaging, McGill University Health Centre, Glen Site, C02.5505, 1001 boul. Décarie, Montréal, QC H4A 3J1, Canada.
Email: veronique.fortier@mcgill.ca

HISTO est fiable quand R_2* est $<$ 100 s-1, correspondant à des patients ayant tout au plus une surcharge modérée en fer et elle ne doit être utilisée qu'avec prudence quand R_2* est $>$ 100 s-1. La technique Dixon doit être préférée pour la mesure de la graisse hépatique en raison de sa robustesse envers les variations de R_2*.

Keywords
liver, fat quantification, HISTO, iron, R_2

Introduction

Steatotic liver disease (SLD)[1] is a liver condition characterized by abnormal fat accumulation in more than 5% of hepatocytes.[2,3] Histopathological analysis of tissue samples obtained by percutaneous biopsy is the historical gold standard to quantify hepatic fat. However, it is subject to sampling error in heterogeneous hepatic fat.

Multiple non-invasive methods have been developed to evaluate SLD such as transient elastography and ultrasound.[4] The current non-invasive gold-standard is magnetic resonance (MR)-based multi-echo Dixon. This technique measures the proton density fat fraction (PDFF), corresponding to the ratio of the fat protons signal to the signal of water and fat protons.[5] PDFF has shown excellent correlation with fat content from biopsies.2,3 However, Dixon-based techniques are limited by fat-water swaps, where the fat signal is incorrectly assigned to the water signal and vice versa, leading to inaccurate quantification.[6] A previous study investigated the prevalence of fat-water swaps and showed that 8% of cases suffered from fat-water swaps at 3.0 T.[7]

MR spectroscopy (MRS) measures the chemical composition of fat and is an alternative method to multi-echo Dixon PDFF.[2] However, MRS techniques offer limited spatial coverage and are sensitive to T_2-signal decay, leading to inaccurate quantification.[8] Nonetheless, MRS is useful for fat quantification as it does not suffer from fat-water swaps.

The high speed T_2-corrected multi-echo (HISTO) technique has been proposed to address some of the limitations of MRS approaches.[9] HISTO is an MRS technique that enables fat quantification from a single breath-hold and includes a correction for T_2-signal decay.[8] HISTO is commercially available, and studies have demonstrated its accuracy for hepatic fat quantification at 1.5 T.[8-10]

HISTO fat quantification accuracy has also been demonstrated at 3.0 T in fat-water phantoms.[11] Its accuracy at 3.0 T has also been investigated in a few clinical indications. Fat fractions estimated with HISTO and Dixon were notably consistent with a strong positive correlation in thigh muscles.[12] A study performed in liver also demonstrated consistent results between HISTO and Dixon for hepatic fat quantification.[13]

Hepatic fat quantification from HISTO, however, requires further investigation since iron accumulation can decrease its accuracy. Different conditions can lead to iron accumulation in liver, such as hereditary haemochromatosis and repeated blood transfusions.[14] Fat and iron can also simultaneously accumulate in the liver. Notably, previous studies have shown that up to a third of patients with non-alcoholic fatty liver disease also had elevated iron content, which has been related to adverse outcomes.[15,16] Iron accumulation increases both R_2 and R_2* relaxation rates, which can bias the fat quantification.[8] R_2 and R_2* increase with field strength, thus leading to a higher potential for bias at 3.0 T. While a previous study performed at 1.5 T has shown accurate fat quantification with HISTO in the presence of an iron-based agent in phantoms,[8] this has not been demonstrated at 3.0 T. Therefore, the objectives of this work were to validate the commercial HISTO fat quantification accuracy at 3.0 T using multi-echo Dixon as the reference method, to establish its robustness to R_2* variations, and to compare the results to measurements performed at 1.5 T.

Methods

Phantom

Fat-water phantoms were designed to mimic a range of liver fat and iron content, using a diluted fat-water emulsion with an R_2* modulator.[17-19] The phantoms were based on a commercial safflower oil emulsion (Microlipid, Nestle Inc.) with 50% fat mass by volume. This emulsion is stable in its liquid form at room temperature, has a controlled fat content, and its fat resonances are consistent with those measured in vivo in the liver.[20] Additionally, it has shown good correlation between fat mass fractions and PDFF estimates in a previous study.[21] The phantom design is described in Supplemental Materials and is illustrated in Figure 1.

Phantoms were scanned using 2 MRI scanners, 1.5 T (Siemens Aera, software VE11C) and 3.0 T (Siemens Skyra, software VE11C). The standard 20-channel head coil provided by the vendor was used. Phantoms were stored in the scanner room for 24 hours before imaging and temperature was measured in the large water compartment before and after imaging.

Phantoms were scanned with 3 different commercial sequences: 2 HISTO sequences performed for each vial, and 1 multi-echo Dixon sequence (qDixon). The HISTO sequences used different echo times (TEs). The first sequence was performed with the default TEs,[8] while the second used shorter TEs, specifically designed to measure fat content in the presence of iron. This technique, known as HISTO Iron Overload, is provided by the vendor as an alternative in

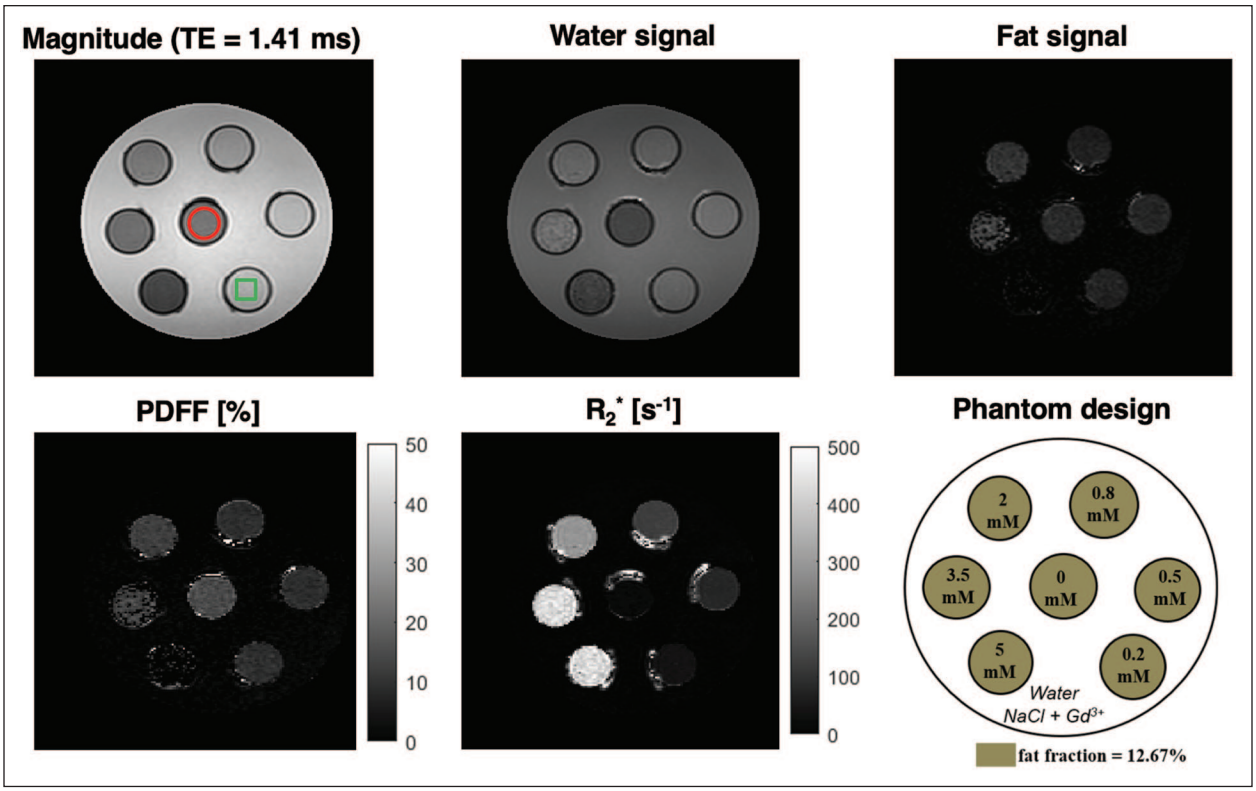


Figure 1. Example images and maps obtained with qDixon at 3.0 T for the phantom with one of the 4 fat fractions (fat mass fraction = 12.67%) used in this work. The schematic of the phantom is shown on the lower right panel, where each vial corresponds to a different [$MnCl_2$]. Note the variations in fat signal and PDFF with increasing [$MnCl_2$] (increasing R_2*). The red circle in the top left panel shows an example of the ROI that was used in all phantom vials for qDixon PDFF and R_2* measurements. The green square in the same panel illustrates typical placement of the HISTO voxel for data acquisition in phantom vials. The signal distortions visible around the vials are caused by the ink used to identify the graduations on the centrifuge tubes and were avoided in ROI placement.
Note. ROI = regions of interest; PDFF = proton density fat fraction; HISTO = high speed T_2-corrected multi-echo.

Table 1. MRI and MRS Pulse Sequence Parameters at 1.5 T and 3.0 T.

Parameters		HISTO (1.5 T, 3.0 T)	HISTO Iron Overload (1.5 T, 3.0 T)	qDixon (1.5 T)	qDixon (3.0 T)
Voxel size (mm^3)	*Phantoms*	15 × 15 × 15	15 × 15 × 15	1.25 × 1.59 × 7	1.25 × 1.59 × 7
	Patients	30 × 30 × 30	—	2.80 × 3.13 × 7	2.80 × 3.13 × 7
Matrix size		Single voxel	Single voxel	160 × 126 × 64	160 × 126 × 64
TE (ms)		12, 24, 36, 48, 72	12, 15, 18, 21, 24	2.38, 4.76, 7.14, 9.52, 11.90, 14.28	1.41, 2.72, 4.03, 5.34, 6.65, 7.96
Bandwidth (Hz/px)		1200	1200	1080	1080
TR (ms)		3000	3000	15.8	9.76
Parallel imaging (CAIPIRINHA)		—	—	2 × 2	2 × 2
Flip angle (°)		90	90	4	4
Mixing time (ms)		10	10	—	—
Scan time (s)	*Phantoms*	15	15	24	14
	Patients	15	15	15-20	12-18

instances where iron overload is suspected. Table 1 details the parameters of these sequences.

Quantitative measurements were obtained using the vendor's automatic post-processing tools (Liver Lab, Siemens Healthineers). PDFF and R_2* maps were automatically generated for qDixon using a signal model that includes a single R_2* and a multi-resonances fat spectrum. PDFF estimates were obtained automatically for HISTO, along with water and fat R_2.

The PDFF accuracy for HISTO and qDixon was evaluated at both field strengths. Circular regions of interest (ROI) with an area of 2.5 cm^2 were drawn in each vial using RadiAnt

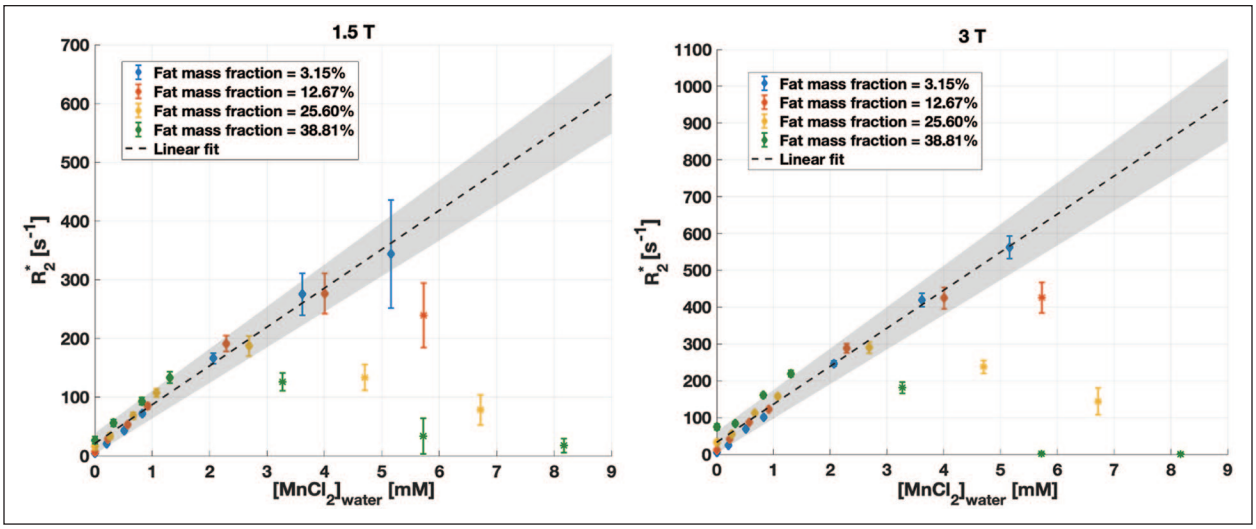


Figure 2. R_2* in phantoms from qDixon as a function of the $[MnCl_2]_{water}$ for all fat mass fractions, at 1.5 T (left) and 3.0 T (right), showing the mean ± standard deviation measured in circular regions of interest. The variation is linear in vials with lower fat content and/or lower $[MnCl_2]_{water}$, but deviates from this trend for high fat content and $[MnCl_2]_{water}$ because of the inaccurate estimate of R_2*. Vials shown with a star marker (*) were excluded from the data analysis. The same 6 vials were excluded at both field strengths.

(Medixant, Poznan, Poland, https://www.radiantviewer.com) on the middle slice of qDixon datasets. The mean and standard deviation of the PDFF and R_2* were measured within all ROIs. R_2* variations were examined as a function of $[MnCl_2]$ adjusted for the water volume ($[MnCl_2]_{water}$), which assumes that $MnCl_2$ is only dissolved in the water phase of the emulsion.[22] A robust linear regression using bi-square weight[23] (Matlab function "robustfit" with the option "bisquare") was performed between R_2* and $[MnCl_2]_{water}$ using software implemented in MATLAB (R2019a, The MathWorks, Natick, MA). The PDFF accuracy was investigated as a function of R_2*. Variations of HISTO water R_2 were also evaluated as a function of $[MnCl_2]_{water}$, as detailed in Supplemental Materials.

Patient Study

A retrospective review of consecutive abdominal MRI examinations was performed under the approval of the local Research Ethics Board (REB). Written informed consent for participants was waived. Inclusion criteria were defined as any abdominal MRI exams performed in adults (age ≥18 years old), for which HISTO and qDixon were acquired. Exams were performed between July and September 2020 at 1.5 T, and between May and August 2020 at 3.0 T, using the same 2 scanners as for the phantoms. All exams with qDixon fat-water swaps were excluded.

A combination of a vendor-provided 18-channel phased array coil and a 32-channel spine coil was used. HISTO and qDixon were each acquired during a single inspiration breath hold. Sequences' parameters are described in Table 1. The HISTO single voxel was positioned in the liver tissue, avoiding blood vessels.

Quantitative measurements were obtained using the vendor-provided automatic post-processing, as for the phantom. HISTO PDFF was compared to the mean qDixon PDFF estimated over the whole liver,[13] obtained using the vendor-provided automated liver segmentation tool. The whole liver mean R_2* from qDixon and the fat and water R_2 from HISTO were also recorded.

Statistical analysis was performed using Matlab. Bland-Altman plots were generated to compare PDFF measurements from both techniques at both field strengths. Bias was tested using a one sample, two-tailed *t*-test. *P*-values $<.05$ were considered statistically significant.

Results

Phantom

R_2* was linear with $[MnCl_2]_{water}$ for vials with low fat fractions and/or low $[MnCl_2]$. In vials with large fat fraction and large $[MnCl_2]$, the mean R_2* deviated from the expected linear behaviour as a function of $[MnCl_2]_{water}$ at both field strengths (Figure 2). This suggested a failure of qDixon for very large R_2*. These vials were excluded from further data analysis. Six vials were excluded at both field strengths, corresponding to the 3 largest $[MnCl_2]$ for a fat fraction of 38.81%, the 2 largest $[MnCl_2]$ for a fat fraction of 25.60%, and the largest $[MnCl_2]$ for a fat fraction of 12.67%. After exclusion, a range of R_2* from 5 to 350 s^{-1} was measured at 1.5 T, and from 5 to 550 s^{-1} at 3.0 T. An R_2* relaxivity for $MnCl_2$ of 66 s^{-1}/mM was measured at 1.5 T, and of 103 s^{-1}/mM at 3.0 T (R_2 = 0.87 for both regressions).

The difference between qDixon PDFF and fat mass fraction increased as a function of the fat content and of the

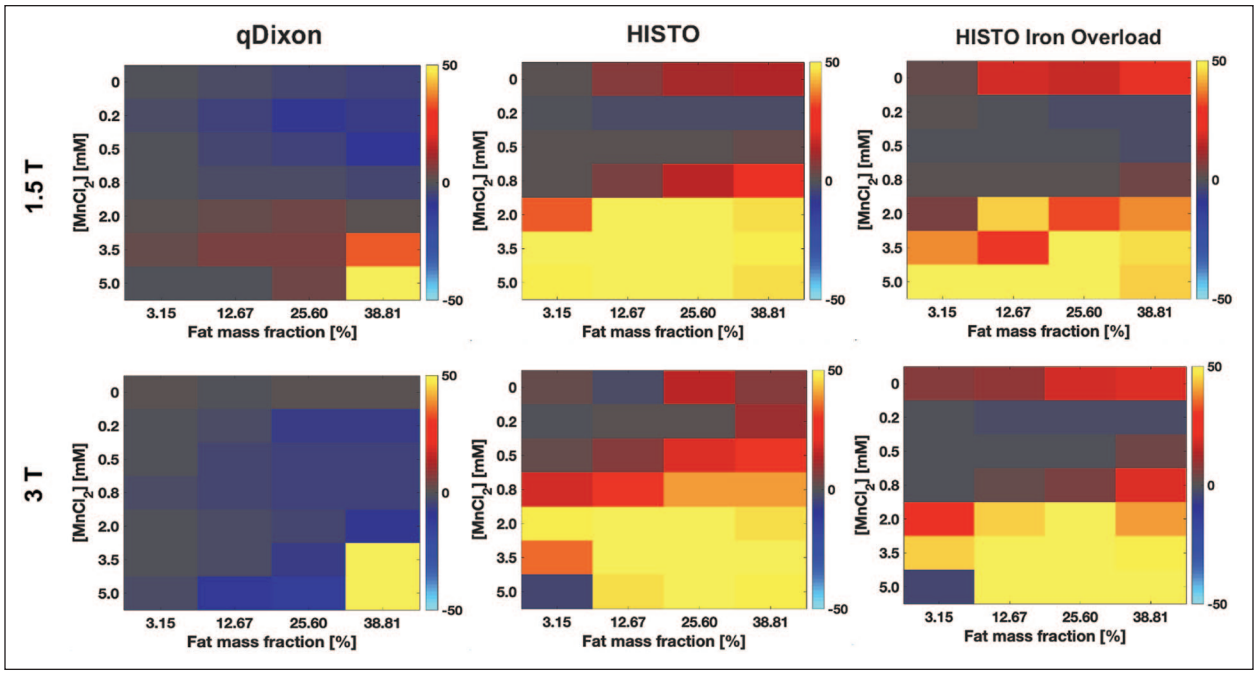


Figure 3. Relative difference (%) between the proton density fat fraction (PDFF) estimated with qDixon (left), HISTO (middle), and HISTO Iron Overload (right) in phantoms and the fat mass fraction, at 1.5 T (top) and 3.0 T (bottom), shown as a function of the concentration of manganese chloride in total solution ([$MnCl_2$]) and of the fat mass fraction. For HISTO and HISTO Iron overload, the PDFF inaccuracy increases significantly as the [$MnCl_2$] increases, with an error larger than 10% for [$MnCl_2$] larger or equal to 2 mM. Both HISTO sequences were also inaccurate in the absence of $MnCl_2$. qDixon was much more accurate over the range of [$MnCl_2$] studied.
Note. HISTO = high speed T_2-corrected multi-echo.

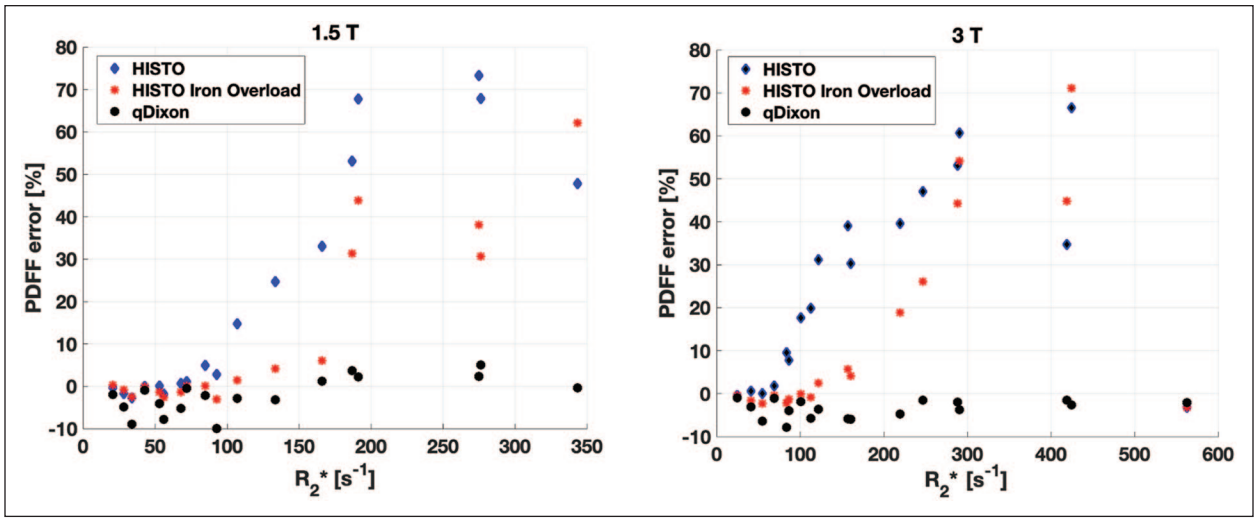


Figure 4. Proton density fat fraction (PDFF) error with HISTO and qDixon in phantoms, at 1.5 T (left) and 3.0 T (right), as a function of the mean R_2* measured based on qDixon measurements. The PDFF error is defined as the difference between the PDFF estimate and the fat mass fraction. Very large PDFF errors were obtained with both HISTO sequences for R_2* larger than 100 s^{-1}. qDixon accuracy was much more stable across the range of R_2* studied, with a PDFF error always under 10%.

[$MnCl_2$]. At 1.5 T, qDixon showed a PDFF error larger than 10% only for the highest fat fraction (38.81%) for 3 of the [$MnCl_2$] (5, 3.5, and 0.5 mM) (Figure 3). At 3.0 T, qDixon showed PDFF errors smaller than 10% for all vials except for the highest [$MnCl_2$] combined with a fat fraction of 12.67%, 25.60%, and 38.81%, and for the highest fat fraction combined with a [$MnCl_2$] of 3.5 mM (Figure 3).

HISTO sequences demonstrated large PDFF inaccuracies (Figure 3). At 1.5 T, HISTO PDFF was accurate within 10% for all fat fractions only for [$MnCl_2$] of 0.2 and 0.5 mM, while at 3.0 T, this was observed for a [$MnCl_2$] of 0.2 mM. With HISTO Iron Overload, the PDFF accuracy was slightly improved at 1.5 T, where the PDFF error was smaller than 10% for a [$MnCl_2$] between 0.2 and 0.8 mM for all fat

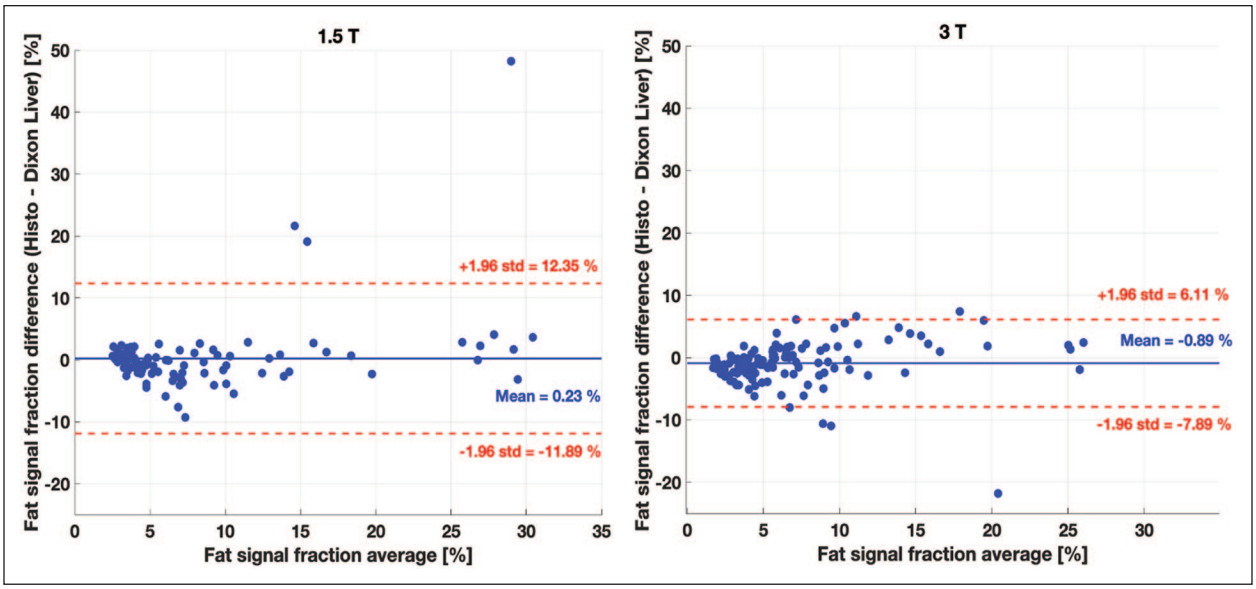


Figure 5. Comparison between the hepatic fat quantification in patients with HISTO and qDixon (mean over whole liver) at 1.5 T (left) and 3.0 T (right). The bias between the 2 techniques is very small (under 1%) at both field strengths. The bias was significant (*P*-value $<.05$) at 3.0 T only.

fractions. This slight improvement in the PDFF accuracy with HISTO Iron Overload was also observed at 3.0 T, but only for $[MnCl_2]$ of 0.2 and 0.5 mM. Both HISTO sequences showed PDFF errors larger than 10% at both field strengths in the absence of $MnCl_2$ ($[MnCl_2]=0$ mM). Vials without $MnCl_2$ were excluded from further data analysis.

HISTO PDFF accuracy was strongly dependant on R_2* (Figure 4). HISTO was accurate to 10% up to a maximum R_2* of 100 s^{-1} (water R_2 ~75 s^{-1}). HISTO Iron Overload was slightly more accurate as a function of R_2*, where a maximum R_2* of 170 s^{-1} (water R_2 ~120 s^{-1}) was associated with a PDFF error smaller than 10%. qDixon was more robust to R_2* variations, resulting in less than 10% PDFF error over the entire range of R_2* that was properly quantified.

Patient Study

Two hundred thirty nine abdominal MRI exams were retrieved, including 110 exams performed at 1.5 T and 129 at 3.0 T. Twelve datasets (10.9%) imaged at 1.5 T were excluded due to qDixon fat-water swaps, resulting in 98 patients (66 women and 32 men, mean age of 61 years old ± 14, median 64 years old). At 3.0 T, 7 datasets (5.4%) were excluded due qDixon fat-water swaps, resulting in 122 exams (59 women and 63 men, mean age of 62 ± 15 years old, median 66 years old).

PDFF, R_2, and R_2* distributions were skewed toward low values. Supporting Figure S2 illustrates the qDixon PDFF and R_2* distributions that were measured at both field strengths, while Supporting Figure S3 shows the HISTO PDFF and R_2 distributions. At 1.5 T, the qDixon PDFF range was between 1.5% and 31.0% (median = 6.0%). At 3.0 T, it was from 1.9% to 37.9% (median = 5.8%). The range of the mean R_2* measured over the whole liver with qDixon was from 20 to 155 s^{-1} (median = 35 s^{-1}) at 1.5 T, while it was between 33 and 160 s^{-1} (median = 54 s^{-1}) at 3.0 T. The vast majority of R_2* measurements were below 100 s^{-1} (99% at 1.5 T and 94% at 3.0 T).

HISTO PDFF was comparable to qDixon PDFF (Figure 5). Both techniques demonstrated an absolute bias smaller than 1% at both field strengths. The bias was not significant at 1.5 T (*P*-value = .71), whereas at 3.0 T, HISTO significantly underestimated the PDFF on average by 0.89% (*P*-value = .007).

Discussion

HISTO PDFF results obtained in this work were accurate up to 10% for R_2* smaller than 100 s^{-1} in phantoms. HISTO results at 1.5 T disagree with a previous study[8] performed in phantoms doped with an iron-based contrast agent, where a maximum PDFF error of 4.4% was measured for a water R_2 up to 140 s^{-1}. Phantom results also showed a much larger PDFF error than previous studies that were performed in vivo outside of the liver,[12] or in phantoms without R_2* modulations.[11] This discrepancy may be explained by the larger range of R_2* studied in this work. PDFF inaccuracies might also be due to a bias in water R_1 introduced by $MnCl_2$. The r_1/r_2 is larger for $MnCl_2$ compared to ferritin,[24,25] resulting in a larger R_1 for a given concentration. Based on previous work,[8] it appears that HISTO may neglect R_1 contributions, which could lead to PDFF inaccuracies.

HISTO and qDixon provided comparable PDFF with a bias smaller than 1% in patients at 1.5 and 3.0 T, where low R_2*

were measured (generally $<100\,s^{-1}$), consistent with previous studies.[13,14] Observations in patients were also consistent with phantom experiments for R_2* smaller than $100\,s^{-1}$. The bias measured in patients was statistically significant at 3.0 T. However, such a small bias is clinically not significant. The agreement observed between HISTO and qDixon is also consistent with a previous phantom study performed with no R_2* modulator at 1.5 and 3.0 T[11].

R_2* in phantoms was representative of values that can be measured in liver. R_2* between 24 and $74\,s^{-1}$ have been reported for healthy liver at 1.5 T.[26-28] At 1.5 T, a threshold between 67 and $88\,s^{-1}$ has been reported to distinguish between healthy liver and mild iron overload.[28] Severe iron overload can produce R_2* larger than $900\,s^{-1}$ at 1.5 T.[27,29] Fewer studies have investigated R_2* in liver at 3.0 T, but one study reported a threshold of $117\,s^{-1}$ for the presence of mild iron overload.[30]

qDixon provided unreliable R_2* measurements in phantoms with large [$MnCl_2$] and/or fat content at both 1.5 and 3.0 T. These inaccuracies were likely due to a TE selection that was too long, as the signal at longer TEs fell below the noise floor. Acquisitions with shorter TEs or custom data post-processing with compensation for longer TEs might address this issue. This was not investigated here in order to rely on commercial packages only.

qDixon PDFF quantification was robust to R_2* modulation. The PDFF error over the range of R_2* that could be studied was consistent with a previous phantom study performed with no R_2* modulations.[11] This suggests that qDixon PDFF accuracy might be independent of the presence of iron in the liver for an R_2* up to approximately $350\,s^{-1}$ at 1.5 T and $550\,s^{-1}$ at 3.0 T. The qDixon PDFF accuracy observed in this work is also consistent with a previous study performed in liver.[3] The qDixon post-processing algorithm assumes a liver fat spectrum at body temperature, which differs from the phantom fat spectrum at 20°C. However, given the PDFF accuracy observed in this work, the difference in the water chemical shift caused by the temperature difference,[31] as well as the differences between the phantom material and liver,[20] were considered negligible. The phantom's temperature was stable and consistent within a range of 0.3°C across scan sessions.

HISTO provided unreliable results in vials without $MnCl_2$. This was likely due to the large water T_1 (>3 s).[32] The TR used was not long enough to enable the complete longitudinal relaxation of water in the absence of $MnCl_2$. The HISTO TR was not increased to keep the scan time within a manageable breathhold duration. As previously suggested,[3] a correction based on the measured water T_1 would be required to remove T_1-weighting. Such correction was not implemented since no T_1 measurement was performed. This would not be a problem in vivo since human tissues have much shorter T_1.

The phantom experiments have certain limitations. The emulsion was liquid, while this is not the case for liver. Liquid solutions have longer T_1 and T_2 values. A gel-based phantom could be used in future work to address this limitation. In addition, the R_2* modulator used was not iron-based, while iron is the agent responsible for R_2* variations in liver. $MnCl_2$ has been used to mimic liver R_2* modulation in multiple studies,[17-19] but it is an imperfect surrogate.[33] However, iron is stored in liver in proteins called ferritin,[34] and $MnCl_2$ has been shown to have similar relaxation properties as aqueous ferritin solutions,[17] thus supporting its use in the context of this work. Future work should be performed in the presence of iron to confirm these findings.

The in vivo study also has some limitations. First, the HISTO PDFF was compared to the qDixon whole liver mean PDFF. Liver has an inhomogeneous fat spatial distribution.[2] This can lead to a significant difference between a single voxel measurement and the whole liver measurement. Second, only the HISTO sequence with the default TEs was used for data acquisition. The phantom experiments demonstrated that HISTO Iron Overload should be preferred because of its slightly higher PDFF accuracy. The use of the HISTO sequence with default TEs might explain the significant difference that was observed at 3.0 T between HISTO and qDixon. Third, low R_2* values were measured, a regime in which HISTO showed consistent results with qDixon in phantoms. Larger R_2* values remain to be tested. Most patients also had low fat content, with a median under 6%, which also contributed to low R_2*. Future in vivo validation should be performed through a prospective study that includes patients with moderate to severe liver iron overload. Fourth, no biopsy samples were available for the patient cohort. As a result, no independent ground truth was available. However, previous studies have demonstrated high correlation between Dixon PDFF and biopsy,[2,3] thus supporting the use of Dixon as a ground truth.

Conclusion

In conclusion, this study demonstrated that Dixon is more accurate than HISTO for PDFF quantification as R_2* increases. Dixon should therefore be preferred for hepatic fat quantification. The phantoms results suggest that HISTO should only be used when R_2* is below $100\,s^{-1}$, which correspond to patients without or with mild liver iron overload. The iron content cannot be known prior to the examination, which suggests a limited applicability of HISTO clinically, although it can be an alternative in the case of a fat-water swap with Dixon. The retrospective patient study confirmed that HISTO and Dixon provide consistent results in vivo in the absence of liver iron overload.

Acknowledgments

The authors acknowledge the Medical Physics Unit (McGill University) for access to lab space, Norma Ybarra (McGill University) for support with the phantom preparation, and the MRI Methods Research Group (McGill University) for useful discussion.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was supported by the Research Institute of the McGill University Health Center General Research Funds.

ORCID iDs

Véronique Fortier https://orcid.org/0000-0003-1859-003X
Jérémy Dana https://orcid.org/0000-0003-0352-4128
Caroline Reinhold https://orcid.org/0000-0002-8852-3273

Supplemental Material

Supplemental material for this article is available online.

References

1. Rinella ME, Lazarus JV, Ratziu V, et al. A multisociety Delphi consensus statement on new fatty liver disease nomenclature. *J Hepatol*. 2023;79(6):1542-1556. doi:10.1016/j.jhep.2023.06.003
2. Starekova J, Reeder SB. Liver fat quantification: where do we stand? *Abdom Radiol*. 2020;45(11):1-14. doi:10.1007/s00261-020-02783-1
3. Kukuk GM, Hittatiya K, Sprinkart AM, et al. Comparison between modified Dixon MRI techniques, MR spectroscopic relaxometry, and different histologic quantification methods in the assessment of hepatic steatosis. *Eur Radiol*. 2015;25(10):2869-2879. doi:10.1007/s00330-015-3703-6
4. Dana J, Venkatasamy A, Saviano A, et al. Conventional and artificial intelligence-based imaging for biomarker discovery in chronic liver disease. *Hepatol Int*. 2022;16(3):509-522. doi:10.1007/s12072-022-10303-0
5. Caussy C, Reeder SB, Sirlin CB, Loomba R. Non-invasive, quantitative assessment of liver fat by MRI-PDFF as an endpoint in NASH trials. *Hepatology*. 2018;68(2):763-772. doi:10.1002/hep.29797
6. Henninger B, Zoller H, Kannengiesser S, Zhong X, Jaschke W, Kremser C. 3D multiecho Dixon for the evaluation of hepatic iron and fat in a clinical setting. *J Magn Reson Imaging*. 2017;46(3):793-800. doi:10.1002/jmri.25630
7. Ladefoged CN, Hansen AE, Keller SH, et al. Impact of incorrect tissue classification in Dixon-based MR-AC: fat-water tissue inversion. *EJNMMI Phys*. 2014;1(1):101. doi:10.1186/s40658-014-0101-0
8. Sharma P, Martin DR, Pineda N, et al. Quantitative analysis of T2-correction in single-voxel magnetic resonance spectroscopy of hepatic lipid fraction. *J Magn Reson Imaging*. 2009;29(3):629-635. doi:10.1002/jmri.21682
9. Pineda N, Sharma P, Xu Q, Hu X, Vos M, Martin DR. Measurement of hepatic lipid: high-speed T2-corrected multiecho acquisition at ^{1}H MR spectroscopy - a rapid and accurate technique. *Radiology*. 2009;252(2):568-576. doi:10.1148/radiol.2523082084
10. Sharma P, Altbach M, Galons JP, Kalb B, Martin DR. Measurement of liver fat fraction and iron with MRI and MR spectroscopy techniques. *Diagn Interv Radiol*. 2014;20(1):17-26. doi:10.5152/dir.2013.13124
11. Jang JK, Lee SS, Kim B, et al. Agreement and reproducibility of proton density fat fraction measurements using commercial MR sequences across different platforms: a multivendor, multi-institutional phantom experiment. *Invest Radiol*. 2019;54(8):517-523. doi:10.1097/RLI.0000000000000561
12. Li Z, Zeng H, Han C, et al. Effectiveness of high-speed T2-corrected multiecho MR spectroscopic method for quantifying thigh muscle fat content in boys with Duchenne muscular dystrophy. *AJR Am J Roentgenol*. 2019;212(6):1354-1360. doi:10.2214/AJR.18.20354
13. Zhao YZ, Zhou JL, Liu JQ, et al. Accuracy of multi-echo Dixon sequence in quantification of hepatic steatosis in Chinese children and adolescents. *World J Gastroenterol*. 2019;25(12):1513-1523. doi:10.3748/wjg.v25.i12.1513
14. Horng DE, Hernando D, Reeder SB. Quantification of liver fat in the presence of iron overload. *J Magn Reson Imaging*. 2017;45(2):428-439. doi:10.1002/jmri.25382
15. Aigner E, Weiss G, Datz C. Dysregulation of iron and copper homeostasis in nonalcoholic fatty liver. *World J Hepatol*. 2015;7(2):177-188. doi:10.4254/wjh.v7.i2.177
16. Dongiovanni P, Fracanzani AL, Fargion S, Valenti L. Iron in fatty liver and in the metabolic syndrome: a promising therapeutic target. *J Hepatol*. 2011;55(4):920-932.
17. Tao R, Zhang J, Dai Y, et al. An in vitro and in vivo analysis of the correlation between susceptibility-weighted imaging phase values and R2* in cirrhotic livers. *PLoS One*. 2012;7(9):1-7. doi:10.1371/journal.pone.0045477
18. St. Pierre TG, Clark PR, Chua-Anusorn W, et al. Noninvasive measurement and imaging of liver iron concentrations using proton magnetic resonance. *Blood*. 2005;105(2):855-861. doi:10.1182/blood-2004-01-0177
19. Yokoo T, Yuan Q, Sénégas J, Wiethoff AJ, Pedrosa I. Quantitative R_2* MRI of the liver with rician noise models for evaluation of hepatic iron overload: simulation, phantom, and early clinical experience. *J Magn Reson Imaging*. 2015;42(6):1544-1559. doi:10.1002/jmri.24948
20. Hamilton G, Yokoo T, Bydder M, et al. In vivo characterization of the liver fat ^{1}H MR spectrum. *NMR Biomed*. 2011;24(7):784-790. doi:10.1002/nbm.1622.In
21. Fortier V, Levesque IR. MR-oximetry with fat DESPOT. *Magn Reson Imaging*. 2023;97:112-121. doi:10.1016/j.mri.2022.12.023
22. Mulkern RV, Hung YP, Ababneh Z, et al. On the strong field dependence and nonlinear response to gadolinium contrast agent of proton transverse relaxation rates in dairy cream. *Magn Reson Imaging*. 2005;23(6):757-764. doi:10.1016/j.mri.2005.07.001
23. Gross AM. Confidence intervals for bisquare regression estimates. *J Am Stat Assoc*. 1977;72(358):341-354. doi:10.1080/01621459.1977.10481001
24. Vymazal J, Zak O, Bulte JWM, Aisen P, Brooks RA. T_1 and T_2 of ferritin solutions: effect of loading factor. *Magn Reson Med*. 1996;36(1):61-65.
25. Thangavel K, Saritaş EÜ. Aqueous paramagnetic solutions for MRI phantoms at 3 T: a detailed study on relaxivities. *Turk J Electr Eng Comput Sci*. 2017;25(3):2108-2121. doi:10.3906/elk-1602-123

26. Obrzut M, Atamaniuk V, Glaser KJ, et al. Value of liver iron concentration in healthy volunteers assessed by MRI. *Sci Rep*. 2020;10(1):1-8. doi:10.1038/s41598-020-74968-z
27. Wood JC, Enriquez C, Ghugre N, et al. MRI R2 and R2* mapping accurately estimates hepatic iron concentration in transfusion-dependent thalassemia and sickle cell disease patients. *Blood*. 2005;106(4):1460-1465. doi:10.1182/blood-2004-10-3982
28. Henninger B, Alustiza J, Garbowski M, Gandon Y. Practical guide to quantification of hepatic iron with MRI. *Eur Radiol*. 2020;30(1):383-393. doi:10.1007/s00330-019-06380-9
29. Yokoo T, Serai SD, Pirasteh A, et al. Linearity, bias, and precision of hepatic proton density fat fraction measurements by using MR imaging: a meta-analysis. *Radiology*. 2018;286(2):486-498. doi:10.1148/radiol.2017170550
30. Dehnad H, Nederveen AJ, van der Heide UA, van Moorselaar RJA, Hofman P, Lagendijk JJW. Clinical feasibility study for the use of implanted gold seeds in the prostate as reliable positioning markers during megavoltage irradiation. *Radiother Oncol*. 2003;67(3):295-302. doi:10.1016/S0167-8140(03)00078-1
31. Hernando D, Sharma SD, Kramer H, Reeder SB. On the confounding effect of temperature on chemical shift-encoded fat quantification. *Magn Reson Med*. 2014;72(2):464-470. doi:10.1002/mrm.24951
32. Chiarotti G, Cristiani G, Giulotto L. Proton relaxation in pure liquids and in liquids containing paramagnetic gases in solution. *Il Nuovo Cimento*. 1955;1(5):863-873. doi:10.1007/BF02731333
33. Zhao R, Hamilton G, Brittain JH, Reeder SB, Hernando D. Design and evaluation of quantitative MRI phantoms to mimic the simultaneous presence of fat, iron, and fibrosis in the liver. *Magn Reson Med*. 2020;85(2):734-747. doi:10.1002/mrm.28452
34. Jensen JH, Tang H, Tosti CL, et al. Separate MRI quantification of dispersed (ferritin-like) and aggregated (hemosiderin-like) storage iron. *Magn Reson Med*. 2010;63(5):1201-1209. doi:10.1002/mrm.22273

References

1 Rumgay, H. *et al.* Global burden of primary liver cancer in 2020 and predictions to 2040. *Journal of Hepatology* **77**, 1598-1606 (2022). https://doi.org:https://doi.org/10.1016/j.jhep.2022.08.021

2 Reig, M. *et al.* BCLC strategy for prognosis prediction and treatment recommendation: The 2022 update. *Journal of Hepatology* **76**, 681-693 (2022). https://doi.org:10.1016/j.jhep.2021.11.018

3 Mazzaferro, V. *et al.* Liver Transplantation for the Treatment of Small Hepatocellular Carcinomas in Patients with Cirrhosis. *New England Journal of Medicine* **334**, 693-700 (1996). https://doi.org:doi:10.1056/NEJM199603143341104

4 Sarasin, F. P., Giostra, E. & Hadengue, A. Cost-effectiveness of screening for detection of small hepatocellular carcinoma in western patients with Child-Pugh class A cirrhosis. *The American Journal of Medicine* **101**, 422-434 (1996). https://doi.org:https://doi.org/10.1016/S0002-9343(96)00197-0

5 Cucchetti, A., Cescon, M., Erroi, V. & Pinna, A. D. Cost-effectiveness of liver cancer screening. *Best Practice & Research Clinical Gastroenterology* **27**, 961-972 (2013). https://doi.org:https://doi.org/10.1016/j.bpg.2013.08.021

6 Nahon, P., Vo Quang, E. & Ganne-Carrié, N. Stratification of Hepatocellular Carcinoma Risk Following HCV Eradication or HBV Control. *Journal of Clinical Medicine* **10**, 353 (2021).

7 Papatheodoridis, G. V., Chan, H. L.-Y., Hansen, B. E., Janssen, H. L. A. & Lampertico, P. Risk of hepatocellular carcinoma in chronic hepatitis B: Assessment and modification with current antiviral therapy. *Journal of Hepatology* **62**, 956-967 (2015). https://doi.org:https://doi.org/10.1016/j.jhep.2015.01.002

8 Nahon, P. *et al.* Eradication of Hepatitis C Virus Infection in Patients With Cirrhosis Reduces Risk of Liver and Non-Liver Complications. *Gastroenterology* **152**, 142-156.e142 (2017). https://doi.org:10.1053/j.gastro.2016.09.009

9 Nahon, P. *et al.* Incidence of Hepatocellular Carcinoma After Direct Antiviral Therapy for HCV in Patients With Cirrhosis Included in Surveillance Programs. *Gastroenterology* **155**, 1436-1450.e1436 (2018). https://doi.org:10.1053/j.gastro.2018.07.015

10 Tzartzeva, K. *et al.* Surveillance Imaging and Alpha Fetoprotein for Early Detection of Hepatocellular Carcinoma in Patients With Cirrhosis: A Meta-analysis. *Gastroenterology* **154**, 1706-1718 e1701 (2018). https://doi.org:10.1053/j.gastro.2018.01.064

11 Park, H. J. *et al.* Non-enhanced magnetic resonance imaging as a surveillance tool for hepatocellular carcinoma: Comparison with ultrasound. *J Hepatol* **72**, 718-724 (2020). https://doi.org:10.1016/j.jhep.2019.12.001

12 Goutté, N. *et al.* Geographical variations in incidence, management and survival of hepatocellular carcinoma in a Western country. *Journal of Hepatology* **66**, 537-544 (2017). https://doi.org:https://doi.org/10.1016/j.jhep.2016.10.015

13 Simmons, O. *et al.* Predictors of adequate ultrasound quality for hepatocellular carcinoma surveillance in patients with cirrhosis. *Aliment Pharmacol Ther* **45**, 169-177 (2017). https://doi.org:10.1111/apt.13841

14 Del Poggio, P. *et al.* Factors that affect efficacy of ultrasound surveillance for early stage hepatocellular carcinoma in patients with cirrhosis. *Clin Gastroenterol Hepatol* **12**, 1927-1933 e1922 (2014). https://doi.org:10.1016/j.cgh.2014.02.025

15 Singal, A. G. *et al.* Detection of hepatocellular carcinoma at advanced stages among patients in the HALT-C trial: where did surveillance fail? *Am J Gastroenterol* **108**, 425-432 (2013). https://doi.org:10.1038/ajg.2012.449

16 Trinchet, J. C. *et al.* Ultrasonographic surveillance of hepatocellular carcinoma in cirrhosis: a randomized trial comparing 3- and 6-month periodicities. *Hepatology* **54**, 1987-1997 (2011). https://doi.org:10.1002/hep.24545

17 Nahon, P. *et al.* Early hepatocellular carcinoma detection using magnetic resonance imaging is cost-effective in high-risk patients with cirrhosis. *JHEP Reports* **4**, 100390 (2022). https://doi.org:10.1016/j.jhepr.2021.100390

18 Gupta, P. *et al.* Abbreviated MRI for hepatocellular carcinoma screening: A systematic review and meta-analysis. *J Hepatol* **75**, 108-119 (2021). https://doi.org:10.1016/j.jhep.2021.01.041

19 Whang, S. *et al.* Comparison of diagnostic performance of non-contrast MRI and abbreviated MRI using gadoxetic acid in initially diagnosed hepatocellular carcinoma patients: a simulation study of surveillance for hepatocellular carcinomas. *European Radiology* **30**, 4150-4163 (2020). https://doi.org:10.1007/s00330-020-06754-4

20 An, J. Y. *et al.* Abbreviated MRI for Hepatocellular Carcinoma Screening and Surveillance. *Radiographics* **40**, 1916-1931 (2020). https://doi.org:10.1148/rg.2020200104

21 Vietti Violi, N. *et al.* Gadoxetate-enhanced abbreviated MRI is highly accurate for hepatocellular carcinoma screening. *European Radiology* **30**, 6003-6013 (2020). https://doi.org:10.1007/s00330-020-07014-1

22 Khatri, G. *et al.* Abbreviated-protocol screening MRI vs. complete-protocol diagnostic MRI for detection of hepatocellular carcinoma in patients with cirrhosis: An equivalence study using LI-RADS v2018. *J Magn Reson Imaging* **51**, 415-425 (2020). https://doi.org:10.1002/jmri.26835

23 Chan, M. V. *et al.* Noncontrast MRI for Hepatocellular Carcinoma Detection: A Systematic Review and Meta-analysis - A Potential Surveillance Tool? *Clin Gastroenterol Hepatol* **20**, 44-56 e42 (2022). https://doi.org:10.1016/j.cgh.2021.02.036

24 Kim, D. H. *et al.* Meta-Analysis of the Accuracy of Abbreviated Magnetic Resonance Imaging for Hepatocellular Carcinoma Surveillance: Non-Contrast versus Hepatobiliary Phase-Abbreviated Magnetic Resonance Imaging. *Cancers* **13**, 2975 (2021).

25 Besa, C. *et al.* Hepatocellular carcinoma detection: diagnostic performance of a simulated abbreviated MRI protocol combining diffusion-weighted and T1-weighted imaging at the delayed phase post gadoxetic acid. *Abdom Radiol (NY)* **42**, 179-190 (2017). https://doi.org:10.1007/s00261-016-0841-5

26 Park, M. S. *et al.* Hepatocellular carcinoma: detection with diffusion-weighted versus contrast-enhanced magnetic resonance imaging in pretransplant patients. *Hepatology* **56**, 140-148 (2012). https://doi.org:10.1002/hep.25681

27 Ronot, M., Nahon, P. & Rimola, J. Screening of liver cancer with abbreviated MRI. *Hepatology* **78**, 670-686 (2023). https://doi.org:10.1097/HEP.0000000000000339

28 Lee, Y. T., Fujiwara, N., Yang, J. D. & Hoshida, Y. Risk stratification and early detection biomarkers for precision HCC screening. *Hepatology* **78**, 319-362 (2023). https://doi.org:10.1002/hep.32779

29 Semmler, G. *et al.* HCC risk stratification after cure of hepatitis C in patients with compensated advanced chronic liver disease. *Journal of Hepatology* **76**, 812-821 (2022). https://doi.org:https://doi.org/10.1016/j.jhep.2021.11.025

30 Ioannou, G. N., Green, P., Kerr, K. F. & Berry, K. Models estimating risk of hepatocellular carcinoma in patients with alcohol or NAFLD-related cirrhosis for risk stratification. *J Hepatol* **71**, 523-533 (2019). https://doi.org:10.1016/j.jhep.2019.05.008

31 Innes, H. *et al.* Performance of models to predict hepatocellular carcinoma risk among UK patients with cirrhosis and cured HCV infection. *JHEP Reports* **3**, 100384 (2021). https://doi.org:https://doi.org/10.1016/j.jhepr.2021.100384

32 Audureau, E. *et al.* Personalized surveillance for hepatocellular carcinoma in cirrhosis - using machine learning adapted to HCV status. *J Hepatol* **73**, 1434-1445 (2020). https://doi.org:10.1016/j.jhep.2020.05.052

33 Singal, A. G. *et al.* International Liver Cancer Association (ILCA) white paper on hepatocellular carcinoma risk stratification and surveillance. *J Hepatol* **79**, 226-239 (2023). https://doi.org:10.1016/j.jhep.2023.02.022

34 Fan, R. *et al.* aMAP risk score predicts hepatocellular carcinoma development in patients with chronic hepatitis. *Journal of Hepatology* **73**, 1368-1378 (2020). https://doi.org:10.1016/j.jhep.2020.07.025

35 Alonso López, S. *et al.* A Model Based on Noninvasive Markers Predicts Very Low Hepatocellular Carcinoma Risk After Viral Response in Hepatitis C Virus-Advanced Fibrosis. *Hepatology (Baltimore, Md.)* **72**, 1924-1934 (2020). https://doi.org:10.1002/hep.31588

36 Fujiwara, N. *et al.* A blood-based prognostic liver secretome signature and long-term hepatocellular carcinoma risk in advanced liver fibrosis. *Med* **2**, 836-850.e810 (2021). https://doi.org:10.1016/j.medj.2021.03.017

37 Hoshida, Y. *et al.* Prognostic Gene Expression Signature for Patients With Hepatitis C–Related Early-Stage Cirrhosis. *Gastroenterology* **144**, 1024-1030 (2013). https://doi.org:https://doi.org/10.1053/j.gastro.2013.01.021

38 Fujiwara, N. *et al.* Molecular signatures of long-term hepatocellular carcinoma risk in nonalcoholic fatty liver disease. *Science Translational Medicine* **14**, eabo4474 (2022). https://doi.org:doi:10.1126/scitranslmed.abo4474

39 Nahon, P. *et al.* Integrating genetic variants into clinical models for hepatocellular carcinoma risk stratification in cirrhosis. *Journal of Hepatology* **78**, 584-595 (2023). https://doi.org:https://doi.org/10.1016/j.jhep.2022.11.003

40 Kitamura, S. *et al.* Liver with hypoechoic nodular pattern as a risk factor for hepatocellular carcinoma. *Gastroenterology* **108**, 1778-1784 (1995). https://doi.org:10.1016/0016-5085(95)90140-x

41 Tarao, K. *et al.* Patients with ultrasonic coarse-nodular cirrhosis who are anti-hepatitis C virus-positive are at high risk for hepatocellular carcinoma. *Cancer* **75**, 1255-1262 (1995). https://doi.org:10.1002/1097-0142(19950315)75:6<1255::aid-cncr2820750607>3.0.co;2-q

42 Caturelli, E. *et al.* Coarse nodular US pattern in hepatic cirrhosis: risk for hepatocellular carcinoma. *Radiology* **226**, 691-697 (2003). https://doi.org:10.1148/radiol.2263011737

43 Savadjiev, P. *et al.* Demystification of AI-driven medical image interpretation: past, present and future. *European radiology* **29**, 1616-1624 (2019). https://doi.org:10.1007/s00330-018-5674-x

44 Dana, J., Agnus, V., Ouhmich, F. & Gallix, B. Multimodality Imaging and Artificial Intelligence for Tumor Characterization: Current Status and Future Perspective. *Semin Nucl Med* **50**, 541-548 (2020). https://doi.org:10.1053/j.semnuclmed.2020.07.003

45 Dana, J. *et al.* Conventional and artificial intelligence-based imaging for biomarker discovery in chronic liver disease. *Hepatology International* (2022). https://doi.org:10.1007/s12072-022-10303-0

46 Hasegawa, K. *et al.* Comparison of resection and ablation for hepatocellular carcinoma: a cohort study based on a Japanese nationwide survey. *J Hepatol* **58**, 724-729 (2013). https://doi.org:10.1016/j.jhep.2012.11.009

47 Roayaie, S. *et al.* Resection of hepatocellular cancer </=2 cm: results from two Western centers. *Hepatology* **57**, 1426-1435 (2013). https://doi.org:10.1002/hep.25832

48 Bradshaw, T. J., Huemann, Z., Hu, J. & Rahmim, A. A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging. *Radiology: Artificial Intelligence* **5**, e220232 (2023). https://doi.org:10.1148/ryai.220232

49 Schoenberger, H. *et al.* Dynamic Changes in Ultrasound Quality for Hepatocellular Carcinoma Screening in Patients With Cirrhosis. *Clinical Gastroenterology and Hepatology* **20**, 1561-1569.e1564 (2022). https://doi.org:https://doi.org/10.1016/j.cgh.2021.06.012

50 Chernyak, V. *et al.* Liver Imaging Reporting and Data System (LI-RADS) Version 2018: Imaging of Hepatocellular Carcinoma in At-Risk Patients. *Radiology* **289**, 816-830 (2018). https://doi.org:10.1148/radiol.2018181494

51 Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging* **29**, 1310-1320 (2010).

52 Chow, S.-C., Shao, J., Wang, H. & Lokhnygina, Y. *Sample size calculations in clinical research*. (chapman and hall/CRC, 2017).

53 Nahon, P. *et al.* Study protocol for FASTRAK: a randomised controlled trial evaluating the cost impact and effectiveness of FAST-MRI for HCC suRveillance in pAtients with high risK of liver cancer. *BMJ Open* **14**, e083701 (2024). https://doi.org:10.1136/bmjopen-2023-083701

54 Roberts, L. R. *et al.* Imaging for the diagnosis of hepatocellular carcinoma: A systematic review and meta-analysis. *Hepatology* **67**, 401-421 (2018). https://doi.org:10.1002/hep.29487

55 Mazellier, J.-P. *et al.* MOSaiC: a Web-based Platform for Collaborative Medical Video Assessment and Annotation. *arXiv preprint arXiv:2312.08593* (2023).

56 Fleiss, J. L., Levin, B. & Paik, M. C. Statistical Methods for Rates and Proportions. *Wiley Series in Probability and Statistics* (2004).

57 Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174 (1977).

58 Dana, J. & Venkatasamy, A. High-resolution (7-T) Liver MRI for Pathologic Examination. *Radiology*, 220410 (2022). https://doi.org:10.1148/radiol.220410

59 Wang, K. *et al.* Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study. *Gut* **68**, 729-741 (2019). https://doi.org:10.1136/gutjnl-2018-316204

60 Lee, J. H. *et al.* Deep learning with ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network. *European Radiology* **30**, 1264-1273 (2020). https://doi.org:10.1007/s00330-019-06407-1

61 Xue, L.-Y. *et al.* Transfer learning radiomics based on multimodal ultrasound imaging for staging liver fibrosis. *European Radiology* **30**, 2973-2983 (2020). https://doi.org:10.1007/s00330-019-06595-w

62 Hectors, S. J. *et al.* Fully automated prediction of liver fibrosis using deep learning analysis of gadoxetic acid–enhanced MRI. *European Radiology* **31**, 3805-3814 (2021). https://doi.org:10.1007/s00330-020-07475-4

63 Yasaka, K., Akai, H., Kunimatsu, A., Abe, O. & Kiryu, S. Liver Fibrosis: Deep Convolutional Neural Network for Staging by Using Gadoxetic Acid–enhanced Hepatobiliary Phase MR Images. *Radiology* **287**, 146-155 (2018). https://doi.org:10.1148/radiol.2017171928

64 Choi, K. J. *et al.* Development and Validation of a Deep Learning System for Staging Liver Fibrosis by Using Contrast Agent-enhanced CT Images in the Liver. *Radiology* **289**, 688-697 (2018). https://doi.org:10.1148/radiol.2018180763

65 Chlebus, G. *et al.* Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections. *PLOS ONE* **14**, e0217228 (2019). https://doi.org:10.1371/journal.pone.0217228

66 Chlebus, G. *et al.* Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. *Scientific Reports* **8**, 15497 (2018). https://doi.org:10.1038/s41598-018-33860-7

67 Zabron, A. *et al.* Clinical and prognostic associations of liver volume determined by computed tomography in acute liver failure. *Liver International* **38**, 1592-1601 (2018). https://doi.org:https://doi.org/10.1111/liv.13725

68 Lecointre, L., Dana, J., Lodi, M., Akladios, C. & Gallix, B. Artificial intelligence-based radiomics models in endometrial cancer: A systematic review. *Eur J Surg Oncol* **47**, 2734-2741 (2021). https://doi.org:10.1016/j.ejso.2021.06.023

69 Meyer, A., Mazellier, J.-P., Dana, J. & Padoy, N. On-the-fly point annotation for fast medical video labeling. *International Journal of Computer Assisted Radiology and Surgery* (2024). https://doi.org:10.1007/s11548-024-03098-y

70 Dashevsky, B. Z. *et al.* The Potential of High Resolution Magnetic Resonance Microscopy in the Pathologic Analysis of Resected Breast and Lymph Tissue. *Sci Rep* **5**, 17435 (2015). https://doi.org:10.1038/srep17435

71 Faria, S. C. *et al*. MR Imaging of Liver Fibrosis: Current State of the Art. *RadioGraphics* **29**, 1615-1635 (2009). https://doi.org:10.1148/rg.296095512

72 Serai, S. D. *et al*. Putting it all together: established and emerging MRI techniques for detecting and measuring liver fibrosis. *Pediatr Radiol* **48**, 1256-1272 (2018). https://doi.org:10.1007/s00247-018-4083-2

73 Minamikawa, T. *et al*. Assessment of Ultra-Early-Stage Liver Fibrosis in Human Non-Alcoholic Fatty Liver Disease by Second-Harmonic Generation Microscopy. *Int J Mol Sci* **23** (2022). https://doi.org:10.3390/ijms23063357

74 Rinella, M. E. *et al*. A multi-society Delphi consensus statement on new fatty liver disease nomenclature. *Journal of Hepatology* (2023). https://doi.org:https://doi.org/10.1016/j.jhep.2023.06.003

RESUME DE LA THESE DE DOCTORAT

Introduction

Au cours des dernières décennies, la prévalence et la morbi-mortalité des maladies hépatiques chroniques ont considérablement augmenté, en particulier avec la progression de la maladie hépatique stéatosique associées aux dysfonctions du métabolisme. Une proportion importante de ces patients développera une fibrose hépatique et éventuellement une cirrhose. La cirrhose est une impasse clinique avec des complications potentiellement mortelles (insuffisance hépatique, hypertension portale, carcinome hépatocellulaire (CHC), etc.), qui est responsable d'environ 1,8 % des décès chaque année (OMS). Lorsque les lésions hépatiques chroniques progressent, une décompensation de la maladie (ascite, ictère, saignement digestif ou encéphalopathie hépatique) peut se produire, entraînant une diminution dramatique du taux de survie global. Actuellement, nous sommes dans l'incapacité d'évaluer séquentiellement la progression des maladies hépatiques sur la seule base de l'imagerie. La caractérisation des maladies hépatiques chroniques repose sur des méthodes invasives telles que la biopsie, pour évaluer la fibrose, la stéatose et l'"activité" (c'est-à-dire l'inflammation), et le cathétérisme trans-jugulaire pour l'hypertension portale (mesure du gradient de pression veineuse hépatique). Pour des raisons évidentes, ces examens invasifs et coûteux ne sont pas adaptés au dépistage et à la surveillance séquentielle. En outre, la biopsie hépatique est également sujette à des risques de sous-échantillonnage et/ou de variabilité entre les lecteurs, et ne permet pas de stratifier le risque de progression de la maladie, y compris l'hépatocarcinogenèse et les complications liées à l'hypertension portale. Ceci doit mener à une transition vers une évaluation non invasive de la progression des maladies hépatiques chroniques et de leur pronostic. Les biomarqueurs basés sur l'imagerie peuvent fournir une représentation quantitative et reproductible du parenchyme hépatique. Ils peuvent être utilisés lors du diagnostic initial ou à tout moment au cours de l'évolution de la maladie, ce qui permet d'influer sur la prise en charge clinique.

Objectif 1 : Stratification du risque d'hépatocarcinogenèse chez les patients à haut risque grâce à une approche par apprentissage profond utilisant l'échographie et l'IRM.

Le dépistage du CHC chez des patients atteints d'une maladie hépatique chronique avancée repose sur une échographie semestrielle. Cependant, la sensibilité de l'échographie est insuffisante, en particulier pour la détection des CHC de moins de 2 cm, où la sensibilité chute à 22 %. La surveillance par IRM a donc été proposée pour améliorer le dépistage, car nettement plus performante que l'échographie, avec un taux de détection cinq fois supérieur à celui de l'échographie pour un CHC à un stade très précoce. Compte tenu du coût élevé et de la disponibilité réduite de l'IRM, l'IRM abrégée (IRMa) sans contraste intraveineux a été récemment proposée car elle offre un gain de temps considérable, limité à moins de 10 minutes, par rapport au protocole d'IRM conventionnel, qui dure de 25 à 40 minutes. Des analyses récentes de cohortes européennes prospectives, associées à une évaluation modélisée de la détection du CHC à un stade très précoce, ont confirmé que la surveillance par IRM est rentable pour une incidence annuelle de base de 3 % chez les patients atteints de cirrhose et ne présentant pas de réplication virale active. Par conséquent, le dépistage par IRMa ne peut être envisagé que pour une sous-population présentant un risque très élevé d'hépatocarcinogenèse, qui serait sélectionnée dans la population faisant actuellement l'objet d'un dépistage échographique standard. L'identification d'une telle population implique le développement d'outils de stratification du risque d'hépatocarcinogenèse. Des modèles préliminaires ont été développés, intégrant des paramètres cliniques (âge, sexe, indice de masse corporelle et diabète) et biologiques (AST/ALT, plaquettes, albumine). Cependant, ces modèles ne prennent pas en compte l'analyse de la micro et macrostructure du foie pourtant accessible par imagerie, qui reflète les mécanismes physiopathologiques responsables de l'hépatocarcinogenèse et qui a le potentiel d'améliorer de manière significative la stratification du risque.

Nous avons émis l'hypothèse que le parenchyme hépatique cirrhotique non tumoral est riche en informations structurelles reflétant la sévérité de la maladie hépatique, son risque carcinogène ainsi que le processus d'hépatocarcinogenèse. L'objectif principal de l'étude était de développer un modèle par apprentissage profond de stratification du risque

d'hépatocarcinogenèse basé sur l'imagerie, distinctement sur l'échographie et l'IRM, afin de proposer différentes modalités de dépistage aux patients les plus à risque.

Pour développer un modèle de stratification du risque basé sur l'échographie, nous avons mené une étude prospective multicentrique qui a inclus 402 patients atteints d'une maladie hépatique chronique avancée. En utilisant un résultat à deux classes (faible risque défini par l'absence de CHC avec 209 patients ; haut risque défini par la présence d'un CHC à un stade précoce avec 193 patients), nous avons développé un modèle de classification (C3D) qui a atteint une précision de 0,72 avec un rapport de cotes de 6,6 pour des patients prédits à haut risque dans un ensemble de test de 50 patients équilibrés entre les deux classes.

Quant au modèle de stratification du risque basé sur l'IRM, nous avons réalisé une étude rétrospective monocentrique avec une méthodologie similaire qui a inclus 333 patients avec une IRM sans contraste abrégée simulée (230 patients à faible risque et 103 patients à haut risque). Le développement d'un modèle par apprentissage machine de type 3D ResNet à partir d'un protocole d'IRM abrégée sans contraste est en cours.

Objectif 2 : Détection du CHC au stade précoce chez les patients à haut risque grâce à une approche par apprentissage profond utilisant l'échographie et l'IRM.

Dans la même étude STARHE, nous avons également développé un modèle de détection basé sur l'apprentissage profond pour le carcinome hépatocellulaire au stade précoce. Le modèle de détection d'objets mis au point a obtenu d'excellentes performances dans la détection des carcinomes hépatocellulaires très précoces (< 2 cm) et précoces (taux global de lésions détectées = 68 % et mAP10 = 0,67) sur les vidéos échographiques. Le niveau de confiance de 70% dans la boîte de prédiction devrait être testé dans des études longitudinales prospectives, en aide à la lecture des images échographiques par les radiologues. Ce modèle pourrait devenir un outil essentiel pour les radiologues et les échographistes afin d'améliorer les performances de l'échographie pour le dépistage du carcinome hépatocellulaire au stade précoce. Le développement d'un modèle similaire en IRM abrégée sans contraste est en cours.

Objectif 3 : Étudier des techniques innovantes pour caractériser les maladies hépatiques

L'intelligence artificielle n'est pas la seule solution pour améliorer la caractérisation des maladies hépatiques chroniques. Dans cette thèse, nous avons étudié les capacités diagnostiques de l'IRM 7T à haute résolution sur des échantillons de foie ex-vivo et développé un score METAVIR-IRM en analogie avec les critères histologiques de stadification basés sur la présence et la distribution de la fibrose. L'imagerie haute résolution 7T a montré d'excellentes performances (précision de 0,93) dans la stadification précise de la fibrose hépatique par rapport à l'histopathologie, soulignant son potentiel en tant qu'outil de substitution innovant pour l'histologie à faible grossissement. Nous avons montré que les capacités physiques de l'IRM peuvent fournir un contraste suffisant entre les tissus pour permettre un examen de type histopathologique sans qu'il soit nécessaire de procéder à une coloration. Atteignant une très haute résolution spatiale (~75µm), proche de celle de l'histologie à faible grossissement, l'architecture du foie a été visualisée pour la première fois par IRM, ce qui n'aurait pas été possible avec la résolution spatiale d'une IRM clinique standard (~ 1mm). Les modifications fibreuses du parenchyme hépatique observées sur les lames d'histologie colorées au trichrome de Masson étaient facilement décelables sur l'IRM à haute résolution, apparaissant comme des septa hyperintenses en pondération T2 avec des distorsions micro-architecturales liées à la fibrose à des stades précoces. En outre, contrairement à l'histologie, qui est limitée aux images 2D, l'IRM fournit des images de l'ensemble de l'échantillon de tissu sous forme de volume. Cette analyse volumétrique est également un avantage par rapport à l'histologie traditionnelle, particulièrement pertinente pour l'étude des distorsions structurelles liées à la fibrose dans les tissus. En outre, l'IRM présente l'avantage indéniable par rapport à l'histologie de pouvoir acquérir des images en un temps relativement court avec une interprétation immédiate de l'image (comme pour toute IRM de routine). L'échantillon, qui reste intact, peut encore être traité pour un examen pathologique ultérieur.

REMERCIEMENTS

A Gabrielle, mon épouse et confidente, pour son soutien constant et amour.

A ma famille et mes amis, pour leur accompagnement depuis tant d'années.

A mes directeurs de thèse, Pr Benoit Gallix et Pr Thomas Baumert, pour leur encadrement et conseils.

A l'équipe de recherche clinique de l'IHU Strasbourg, Armelle, Elsa, Kahina, Laura, et Pierre.

A l'équipe de computer science du Pr Nicolas Padoy et particulièrement à Adrien Meyer.

A l'équipe du Work Package 2 du programme DELIVER et notamment les Pr Pierre Nahon et Pr Maxime Ronot pour leur expertise et encadrement.

Au Pr Valérie Vilgrain pour sa bienveillance et conseils.

Au Pr Caroline Reinhold pour son soutien.

A l'ensemble des équipes investigatrices de STARHE.

Université de Strasbourg

Jérémy Dana

Risk stratification of hepatocarcinogenesis using deep learning analysis of radiological images

École doctorale Sciences de la vie et de la santé | ED 414 Université de Strasbourg

Résumé

La cirrhose est une impasse clinique avec des complications potentiellement mortelles (insuffisance hépatique, hypertension portale, carcinome hépatocellulaire), responsable d'environ 1,8 % des décès chaque année. Lorsque les lésions hépatiques chroniques progressent, une décompensation de la maladie peut se produire, entraînant une diminution dramatique du taux de survie global. Une transition vers une évaluation non invasive de la progression des maladies hépatiques chroniques et de leur pronostic est crucial. Nous avons développé un modèle par apprentissage profond de stratification du risque de développer un carcinome hépatocellulaire à partir d'images échographiques du parenchyme hépatique non-tumoral. Le modèle de classification développé a atteint de bonnes performances diagnostiques avec un odds-ratio de 6,6 et une précision de 0,72 pour prédire les patients à haut risque d'hépatocarcinogenèse. Ce nouveau biomarqueur d'imagerie pourrait aider à stratifier le risque d'hépatocarcinogenèse parallèlement aux biomarqueurs cliniques ou biochimiques et permettre des stratégies de dépistage personnalisées basées sur le risque.

Mots clés : carcinome hépatocellulaire, apprentissage profond, échographie

Résumé en anglais

Cirrhosis is a clinical dead end with potentially fatal complications (liver failure, portal hypertension, hepatocellular carcinoma), responsible for around 1.8% of deaths each year. As chronic liver damage progresses, decompensation of the disease can occur, leading to a dramatic reduction in overall survival. A transition to non-invasive assessment of chronic liver disease progression and prognosis is crucial. We have developed a deep learning model for stratifying the risk of developing hepatocellular carcinoma based on ultrasound images of non-tumoral liver parenchyma. The classification model developed achieved good diagnostic performance with an odds ratio of 6.6 and an accuracy of 0.72 for predicting patients at high risk of hepatocarcinogenesis. This new imaging biomarker could help to stratify the risk of hepatocarcinogenesis alongside clinical or biochemical biomarkers and enable personalised risk-based screening strategies.

Keywords: hepatocellular carcinoma, deep learning, ultrasound