Rovigo 30 maggio 2024

Artificial Intelligence and Liver Diseases

Prof. Fabio Piscaglia

Professore Ordinario di Medicina Interna,
Direttore Scuola di Specialità in Medicina Interna,
Università di Bologna
Direttore UOC. Medicina Interna, Malattie Epatobiliari ed Immunoallergologiche
IRCCS, Azienda Ospedaliero-Universitaria di Bologna, Policlinico S.Orsola

EFSUMB

Conflicts of Interest of Fabio Piscaglia

Term 2021-2024

Astrazeneca, Bayer, Bracco, ESAOTE, EISAI, Exact Sciences, GE, Gilead, IPSEN, MSD, Nerviano, Roche, Samsung, Siemens Healthineers

Current main unmet **clinical** needs in hepatology to which AI could theoretically contribute for the daily practice:

- Non invasive precise fibrotic staging of chronic liver disease
- Non invasive distinction between MASLD and MASH
- Improved non invasive diagnosis of HCC and of its molecular characterization
- Individual precise prediction of the risk of HCC development
- Improved histological diagnosis
- Prediction of the response of HCC to treatment and risk of recurrence
- Identification of complex / rare etiologies of liver lab exam abnormalities
- Identification of subphenotypes within current single etiology and corresponding natural course of disease and response to treatment

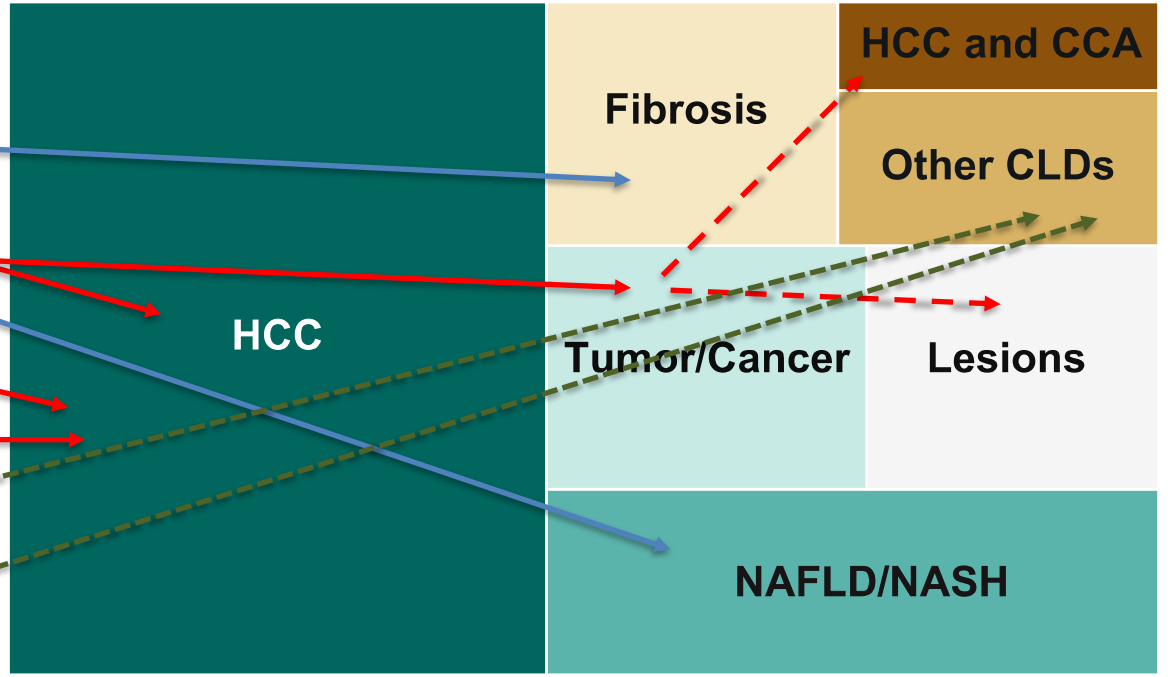
Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction

David Nam,[1] Julius Chapiro,[1] Valerie Paradis,[2,3] Tobias Paul Seraphin,[4,5] Jakob Nikolas Kather[5,6,7,*]

Current main unmet **clinical** needs in hepatology to which AI could theoretically contribute for the daily practice:

- Non invasive precise fibrotic staging of chronic liver disease
- Non invasive distinction between NAFLD and NASH
- Improved non invasive diagnosis of HCC and of its molecular characterization
- Individual precise prediction of the risk of HCC development
- Improved histological diagnosis ALL fields
- Prediction of the response of HCC to treatment and risk of recurrence
- Identification of complex / rare etiologies of liver lab exam abnormalities
- Identification of subphenotypes within current single etiology and corresponding natural course of disease and response to treatment

Number of studies by liver disease



JHEP Reports **2022** vol. 4 | 100443

Indeed, published research about applications of AI in liver disease has been largely addressed to the unmet clinical needs in hepatology so far

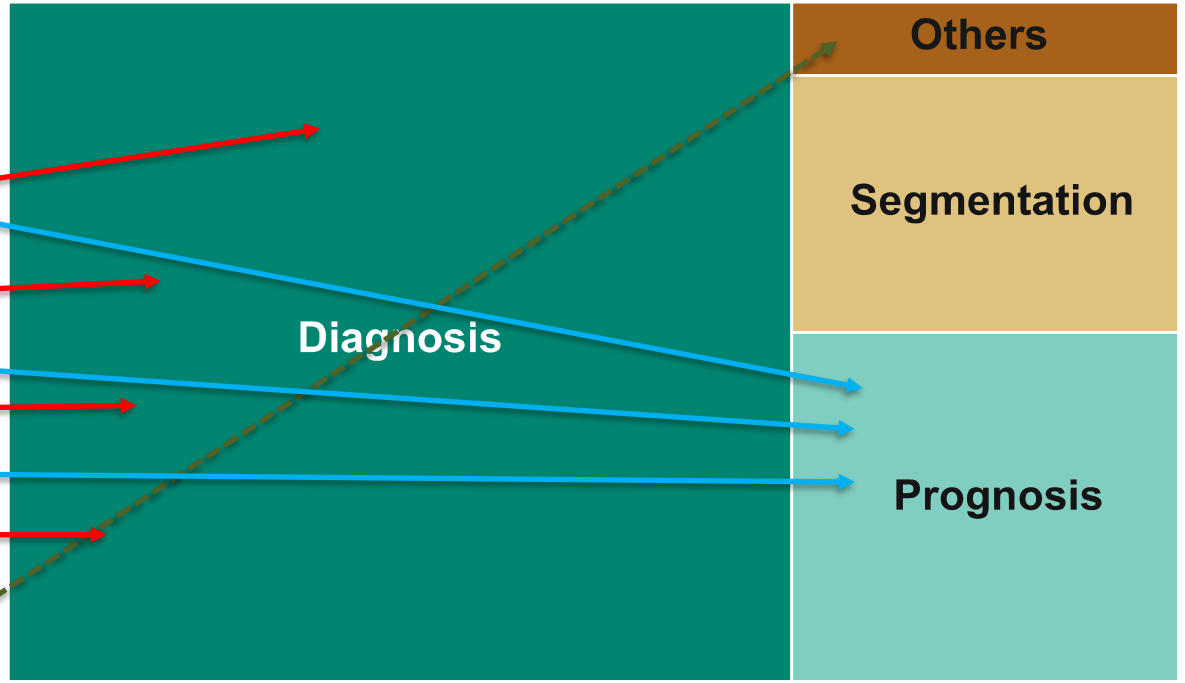
Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction

David Nam,[1] Julius Chapiro,[1] Valerie Paradis,[2,3] Tobias Paul Seraphin,[4,5] Jakob Nikolas Kather[5,6,7,*]

Current main unmet **clinical** needs in hepatology to which AI could theoretically contribute for the daily practice:

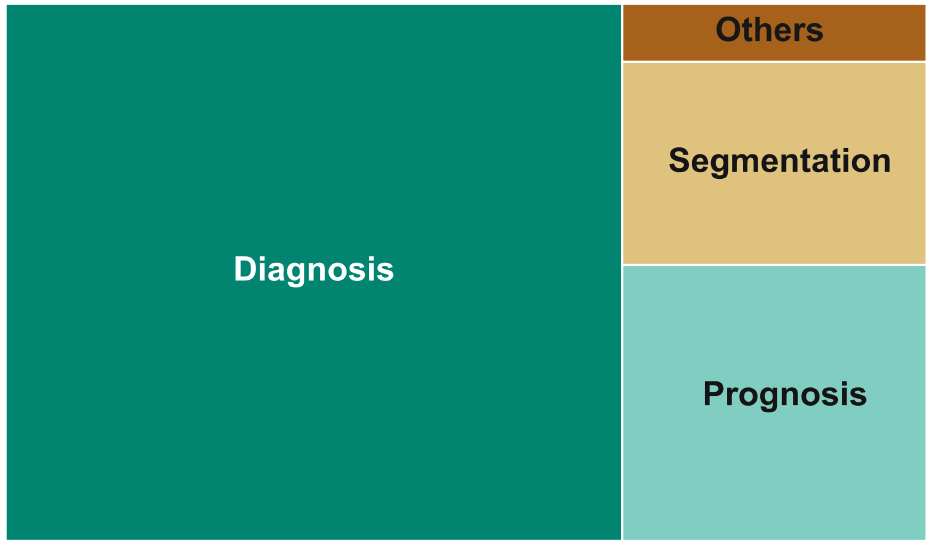
- Non invasive precise fibrotic staging of chronic liver disease
- Non invasive distinction between NAFLD and NASH
- Improved non invasive diagnosis of HCC and of its molecular characterization
- Individual precise prediction of the risk of HCC development
- Improved histological diagnosis
- Prediction of the response of HCC to treatment and risk of recurrence
- Identification of complex / rare etiologies of liver lab exam abnormalities
- Identification of subphenotypes within current single etiology and corresponding natural course of disease and response to treatment

Number of studies by prediction of the models

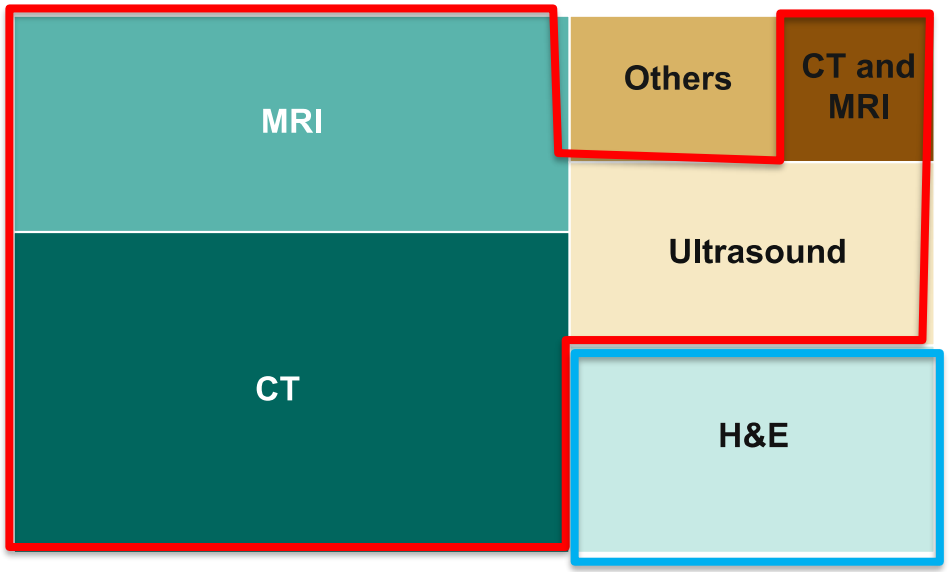


JHEP Reports **2022** vol. 4 | 100443

Number of studies by prediction of the models



Number of studies by input data used



Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction

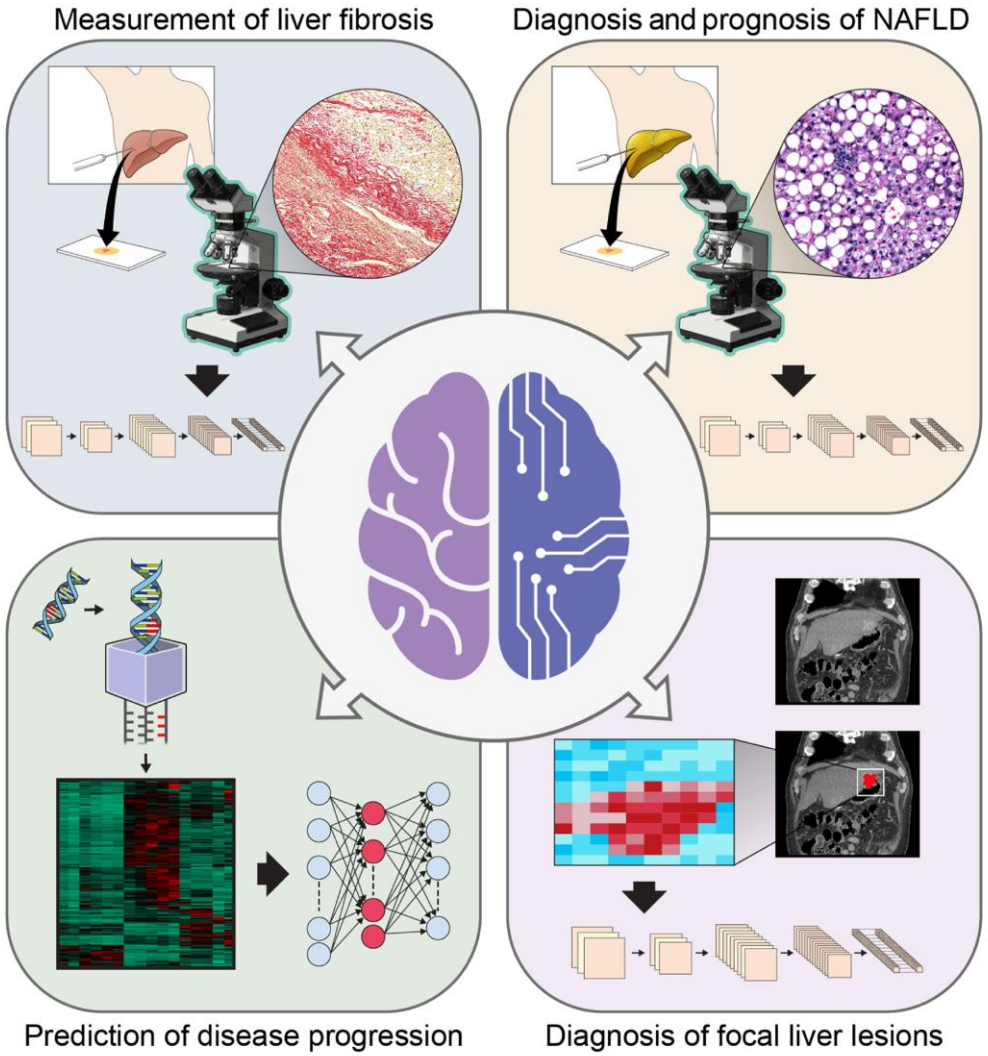
David Nam,[1] Julius Chapiro,[1] Valerie Paradis,[2,3] Tobias Paul Seraphin,[4,5] Jakob Nikolas Kather[5,6,7,*]

JHEP Reports **2022** vol. 4 | 100443

Imaging is very rich of information and data are easier to be extracted / collected and heterogeneity easier to be recognized and managed than laboratory / clinical data. Convolutionary Neural Networks models mainly apply.

Radiological Images

Histological images



Hye Won Lee et al Artificial intelligence in liver disease
Journal of Gastroenterology and Hepatology 36 (2021) 539–542

In front of all such efforts which are the main issues to be carefully taken into consideration before arriving to clinical applications of AI in liver disease?

Common issue to be considered before using AI in clinical practice (also in hepatology.

- Ground truth (i.e. reference standards)
- Clinicallly meaningful thresholds for diagnostic accuracy
- Responsibility
- External accessibility
- Applicability

Ground truth issue.

In order to train the AI outcomes must be “labeled” (=correctly and definitively diagnosed)

However, in the instance of focal liver lesions at risk for HCC or in the instance of NAFLD the ground truth (“reference standard”) is problematic.

The optimal ”ground truth” would be histology. However, this is not systematically available in the instance of non-HCC lesions (e.g. LI RADS L2, LR3 and several LR4 lesions) and similarly also in the instance of MASLD (to ascertain the presence of MASH and degree of fibrosis severity). *Therefore, the available case series based on which AI models are developed often contain selected cases, which might not adequately represent the general population to which the AI model is expected to be applied in the clinical practice*.

In the clinical practice many more non malignant lesions are visualized by radiology than those for which histology is available.
Should potentially AI models work at at a “patient” base rather than “lesion” base in this setting?

AMIA Annual Symposium Proceedings Archive

AMIA Annu Symp Proc. 2018; 2018: 430–439.
Published online 2018 Dec 5.
PMCID: PMC6371264
PMID: 30815083

Application of Machine Learning Methods to Predict Non-Alcoholic Steatohepatitis (NASH) in Non-Alcoholic Fatty Liver (NAFL) Patients

Suruchi Fialoke, Ph.D,[1] Anders Malarstig, Ph.D,[1] Melissa R. Miller, PhD,[1] and Alexandra Dumitriu, Ph.D[1]

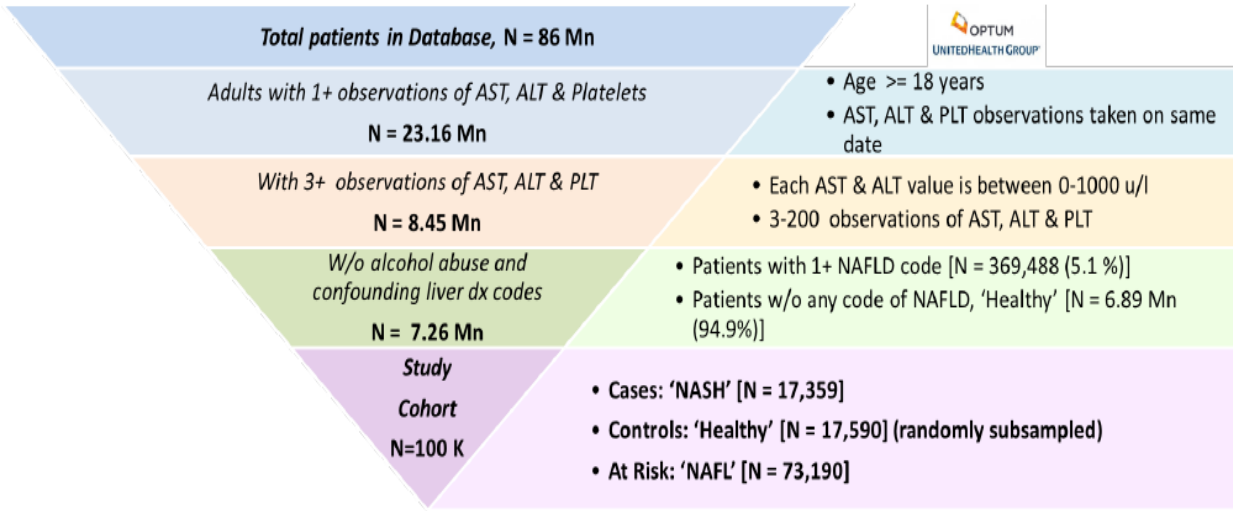


Figure 1: Cohort preparation from United States-based electronic health records (EHRs) provided by Optum

They used electronic health records from the Optum Analytics to:

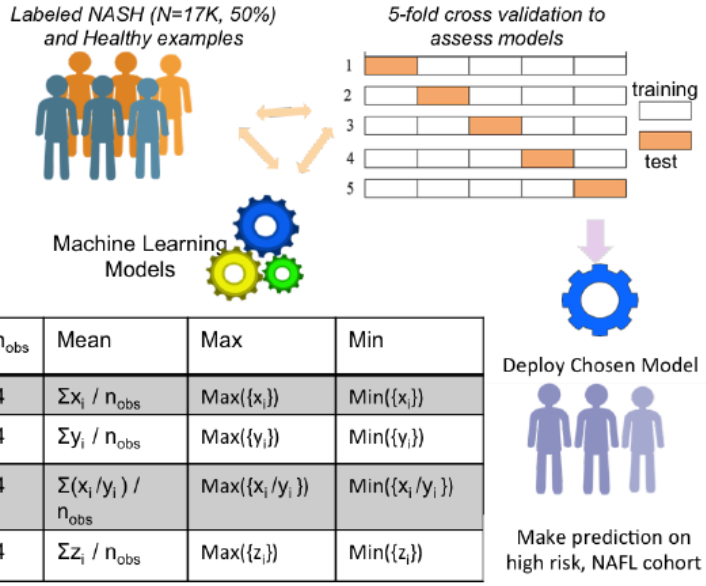
(1) identify patients diagnosed with either benign steatosis (NAFL) or NASH based on ICD codes

(2) train machine learning classifiers for NASH and healthy (non-NASH) populations

(3) predict NASH disease status on patients diagnosed with NAFL.

Features (From EHR)

- ❑ *Gender, Age, Race*
- ❑ n_{obs} *(number of ALT, AST & PLT available)*
- ❑ *Most recent, max, min and mean of ALTs*
- ❑ *Most recent, max, min and mean of ASTs*
- ❑ *Most recent, max, min and mean of AST/ALT*
- ❑ *Most recent, max, min and mean of PLT*
- ❑ *Age in most recent lab, longitudinal history*
- ❑ *Diabetes (Y/N)*



Time →	t_1	t_2	t_3	t_4	…	t_n	n_{obs}	Mean	Max	Min
AST	x_1	x_2	x_3	x_4			4	$\Sigma x_i / n_{obs}$	$Max(\{x_i\})$	$Min(\{x_i\})$
ALT	y_1	y_2	y_3	y_4			4	$\Sigma y_i / n_{obs}$	$Max(\{y_i\})$	$Min(\{y_i\})$
AST/ALT	x_1/y_1	x_2/y_2	x_3/y_3	x_4/y_4			4	$\Sigma(x_i/y_i) / n_{obs}$	$Max(\{x_i/y_i\})$	$Min(\{x_i/y_i\})$
PLT	z_1	z_2	z_3	z_4			4	$\Sigma z_i / n_{obs}$	$Max(\{z_i\})$	$Min(\{z_i\})$

Figure 2: Features employed in supervised machine learning models for NASH

	Threshold independent		Threshold dependent (threshold = 0.50)				
Model	AUROC	Log loss	Accuracy	Precision	Recall	F-Measure	MCC
LR	0.835 (0.011)	0.520 (0.012)	0.762 (0.013)	0.770 (0.020)	0.743 (0.013)	0.756 (0.01)	0.52 (0.026)
DT	0.842 (0.013)	0.492 (0.021)	0.772 (0.012)	0.786 (0.023)	0.745 (0.046)	0.764 (0.02)	0.54 (0.024)
RF	0.870 (0.010)	0.451 (0.014)	0.792 (0.010)	0.804 (0.014)	0.768 (0.010)	0.786 (0.01)	0.58 (0.021)
XGB	**0.876 (0.010)**	**0.440 (0.016)**	**0.797 (0.010)**	**0.808 (0.015)**	**0.774 (0.008)**	**0.791 (0.01)**	**0.594 (0.02)**

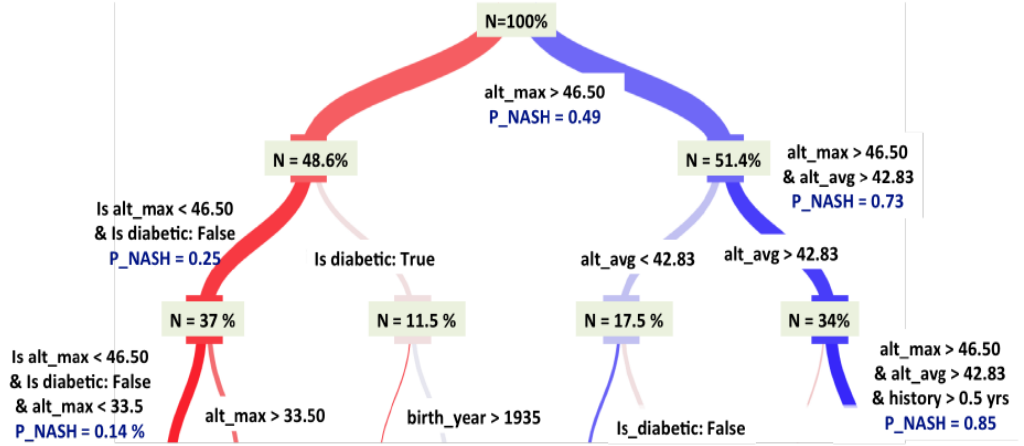


Figure 3: The top three branches of the Decision Tree model (with tree-depth = 5) highlights the decision making process and the relative importance of the various variables.

They usedXGBoost to make prediction of NASH on a third cohort of benign fatty liver (NAFL) patients. Consistent with recorded prevalence of diabetes, the NAFL cohort classified as NASH using the model had significantly higher prevalence of diabetes (48%) when compared to those classified as Healthy (18%).

The performance metrics from 4 popular machine learning classifiers, Logistic Regression20 , Decision Tree21 , Random Forest22 , and XG-Boost23.

It is clear that while all classifiers perform well at classifying positive and negative examples of NASH, we gain performance boost at the cost of interpretability with the XGBoost model, which shows an AUROC of 88%.

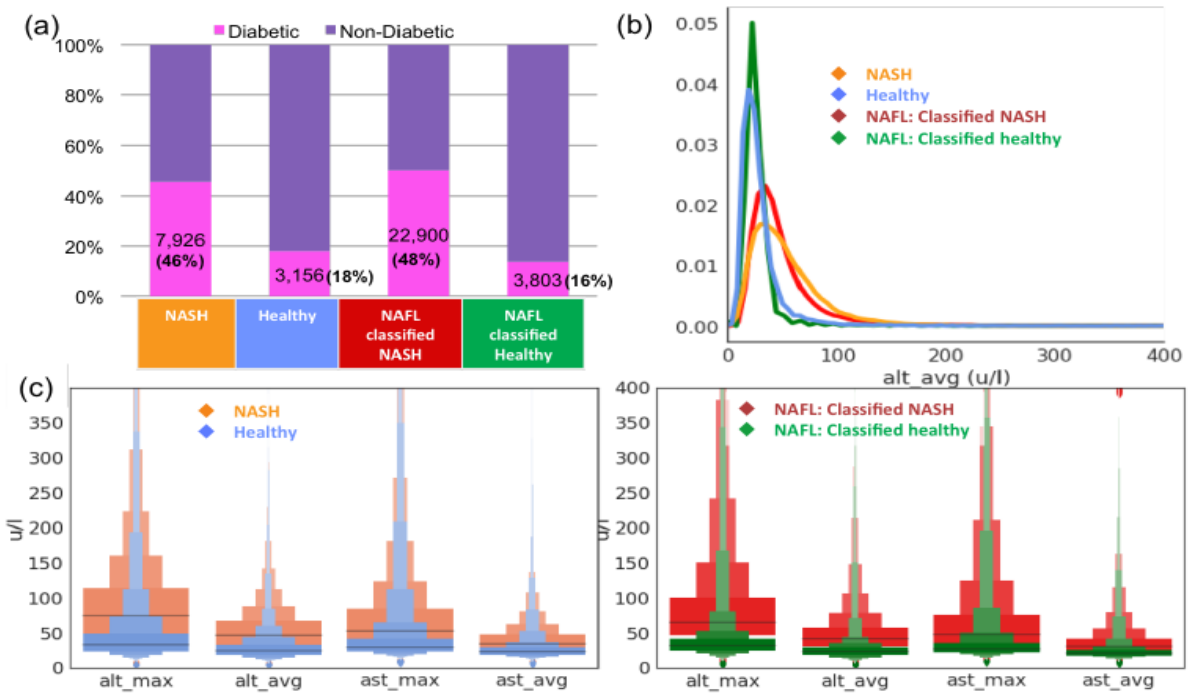


Figure 6: XGBoost classifier's predictions on the Optum EHR NAFL cohort (a) Prevalence of diabetes in the NAFL cohort classified as NASH is significantly higher than in the NAFL cohort classified as Healthy. (b) The distributions of longitudinal average of ALT values (most important lab-based feature learnt from classifier), 'alt_avg', for the NASH patients is similar to the distribution in NAFL patients classified as NASH. (c) Distributions of relevant lab-based longitudinal parameters are summarized in the Letter Value box-plots[26] indicating the various quartiles of the distribution (grey lines correspond to the middle-quartiles or the medians).

	Threshold independent		Threshold dependent (threshold = 0.50)				
Model	AUROC	Log loss	Accuracy	Precision	Recall	F-Measure	MCC
LR	0.835 (0.011)	0.520 (0.012)	0.762 (0.013)	0.770 (0.020)	0.743 (0.013)	0.756 (0.01)	0.52 (0.026)
DT	0.842 (0.013)	0.492 (0.021)	0.772 (0.012)	0.786 (0.023)	0.745 (0.046)	0.764 (0.02)	0.54 (0.024)
RF	0.870 (0.010)	0.451 (0.014)	0.792 (0.010)	0.804 (0.014)	0.768 (0.010)	0.786 (0.01)	0.58 (0.021)
XGB	**0.876 (0.010)**	**0.440 (0.016)**	**0.797 (0.010)**	**0.808 (0.015)**	**0.774 (0.008)**	**0.791 (0.01)**	**0.594 (0.02)**

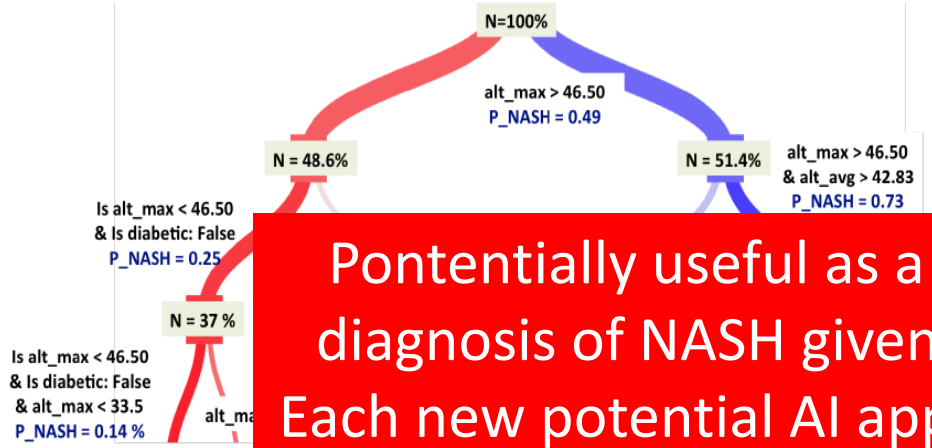


Figure 3: The top three branch
process and the relative importa

The performance metrics from 4 popular machine learning classifiers, Logistic Regression20 , Decision Tree21 , Random Forest22 , and XG-Boost23.

It is clear that while all classifiers perform well at classifying positive and negative examples of NASH, we gain performance boost at the cost of interpretability with the XGBoost model, which shows an AUROC of 88%.

Pontentially useful as a case finding method, but not to estalish any diagnosis of NASH given the relatively weak unverified ground truth. Each new potential AI application must be adopted strictly according to the ground truth utilized to set it up.

They usedXGBoost to make prediction of NASH on a third cohort of benign fatty liver (NAFL) patients. Consistent with recorded prevalence of diabetes, the NAFL cohort classified as NASH using the model had significantly higher prevalence of diabetes (48%) when compared to those classified as Healthy (18%).

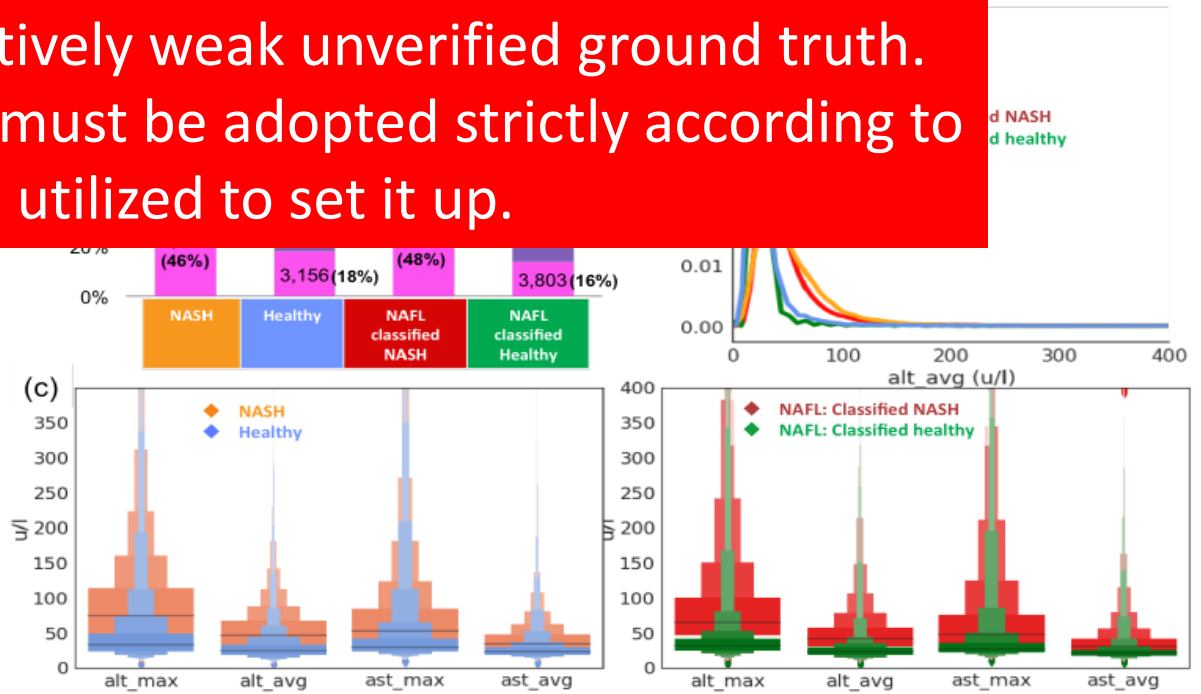


Figure 6: XGBoost classifier's predictions on the Optum EHR NAFL cohort (a) Prevalence of diabetes in the NAFL cohort classified as NASH is significantly higher than in the NAFL cohort classified as Healthy. (b) The distributions of longitudinal average of ALT values (most important lab-based feature learnt from classifier), 'alt_avg', for the NASH patients is similar to the distribution in NAFL patients classified as NASH. (c) Distributions of relevant lab-based longitudinal parameters are summarized in the Letter Value box-plots[26] indicating the various quartiles of the distribution (grey lines correspond to the middle-quartiles or the medians).

Thresholds for clinical impact of AI models

AI models are expected to improve the current diagnostic / prognostic capacities.

However, not any improvement translate into a clinically meaningful application and in order to replace a current ground truth (despite this may be imperfect or impracticle to achieve) the scientific community has to establish **thresholds for clinically significant improved performance**.

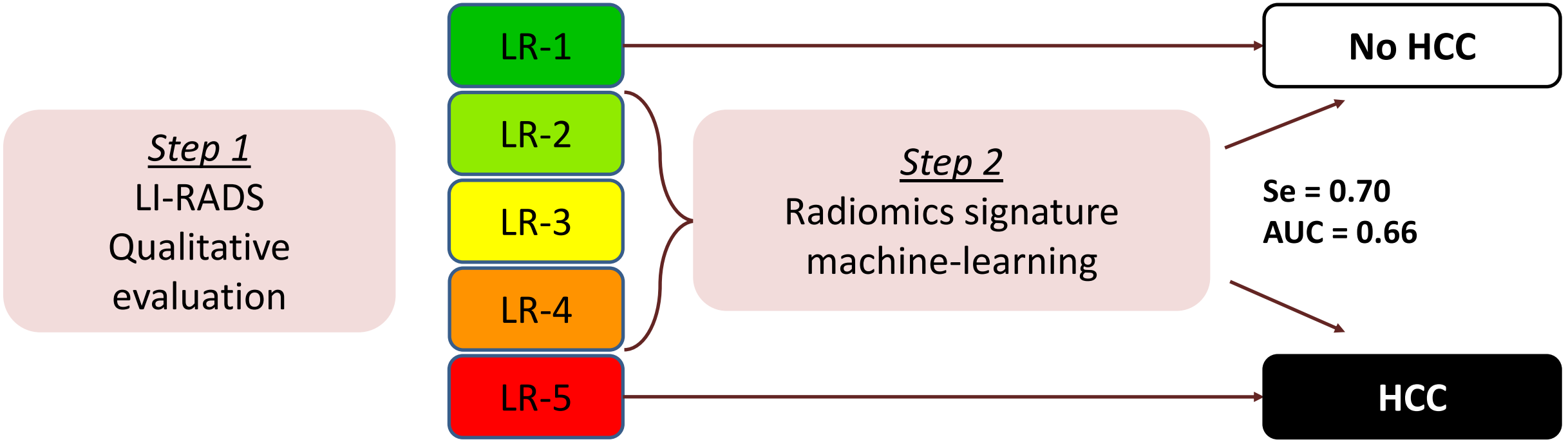
E.g.: current standard models to distinguish simple NAFLD from NASH are largely suboptimal, with diagnostic accuracies in the range of 0.60-0.65. If an AI model increases this rate to 0.70 it would be a huge step forward in statistical terms, but would it make any clinically relevant impact? Would a positive predictive values for NASH raised from 65% to 75% change the clinical practice?.

Increasing the positive predictive value of the LR4 class for FLL at risk of HCC from 85% to 89% or for LR3 from 35% to 50% would be scientifically relevant, but would it bring any clinical impact? Which is the desired threshold for which an AI model would be accepted? To include a patient in a trial for NASH with fibrosis ≥F2, without obtaining histology, which would be the minimal accepted diagnostic likelyhood ratio of an AI model (>10, >15, >20 or PPV >95% or 98%, etc)?

Radiomics in indeterminate nodules at risk for HCC. Radiomics signature performance may improve existing clinical classification.

Radiomics after LI RADS

178 cirrhotic patients

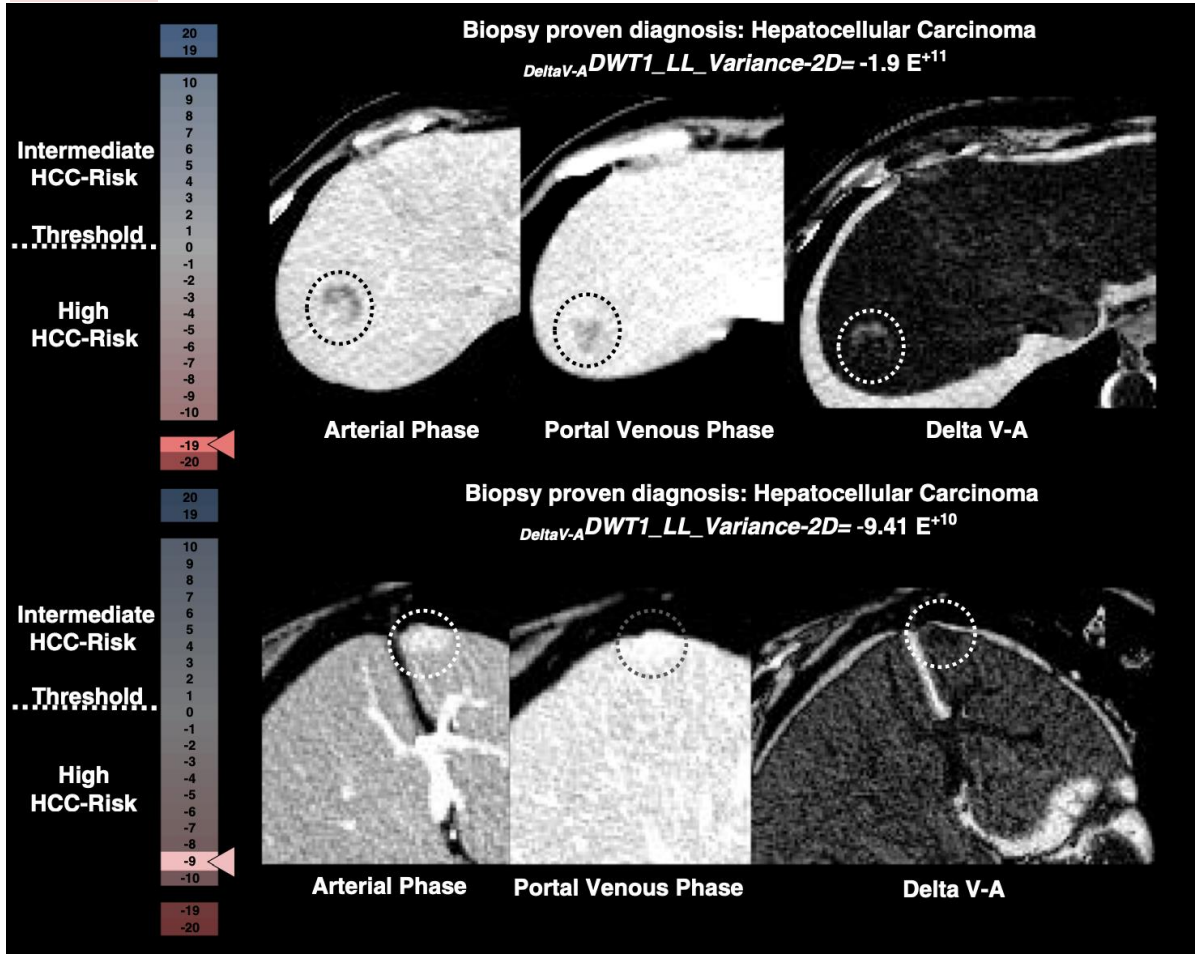


Mokrane Eur Radiol 2020

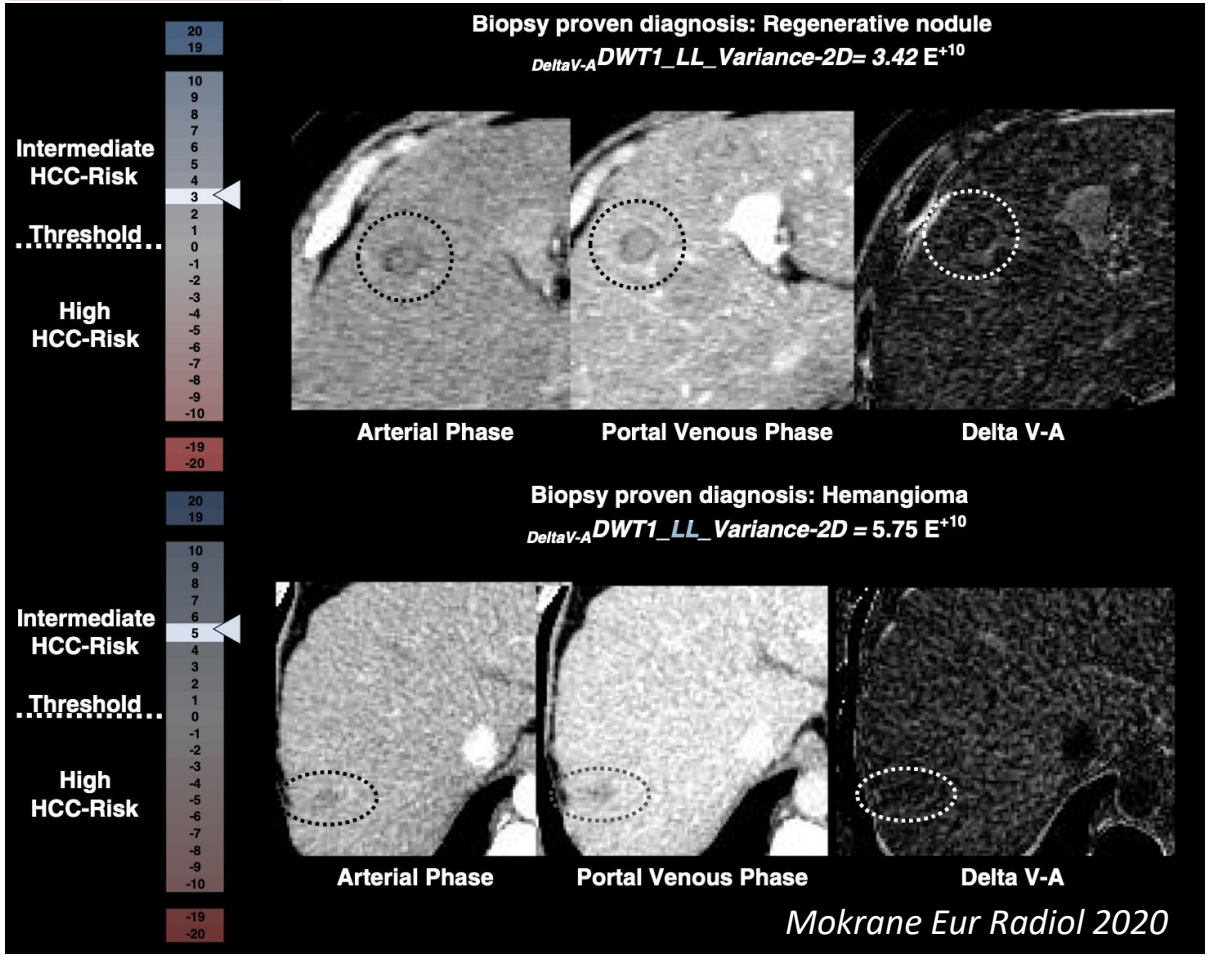
Radiomics in indeterminate nodules at risk for HCC. Radiomics signature performance may improve existing clinical classification.

Radiomics after LI RADS

HCC



Non HCC



Mokrane Eur Radiol 2020

Responsibility and accessibility

Who is responsible for the final diagnosis?

Should patients have access to the information provided by AI tools or is it only a part of the diagnostic process (and not the conclusion)?

Example (the same may apply to liver tumors in the future):

Deep Learning for Pediatric Posterior Fossa Tumor Detection and Classification: A Multi-Institutional Study

J.L. Quon, W. Bala, L.C. Chen, J. Wright, L.H. Kim, M. Han, K. Shpanskaya, E.H. Lee, E. Tong, M. Iv, J. Seekins, M.P. Lungren, K.R.M. Braun, T.Y. Poussaint, S. Laughlin, M.D. Taylor, R.M. Lober, H. Vogel, P.G. Fisher, G.A. Grant, V. Ramaswamy, N.A. Vitanza, C.Y. Ho, M.S.B. Edwards, S.H. Cheshier, and K.W. Yeom

AJNR Am J Neuroradiol 41:1718–25 Sep 2020

Ground Truth Labels

Pathology from surgical specimens served as ground truth (MB, EP, PA) except for most patients with DMG who were diagnosed primarily by MR imaging. An attending pediatric neuroradiologist (K.W.Y.) manually classified each axial slice as having tumor versus no tumor: A slice was considered positive if any tumor was visible.

Deep Learning Model Architecture

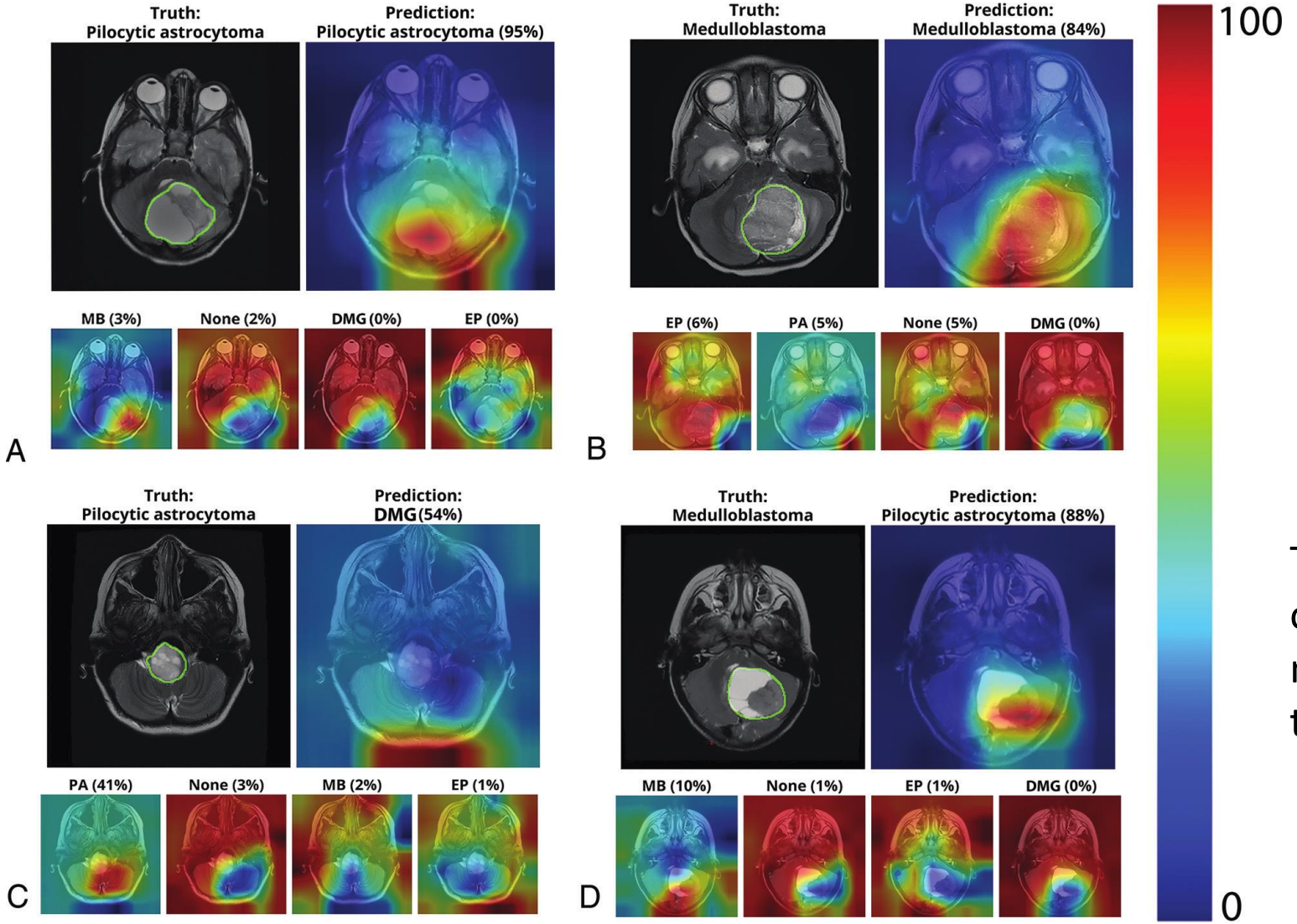
We chose a 2D ResNeXt-50-32x4d deep learning architecture (https://github.com/titu1994/Keras-ResNeXt) rather than a 3D architecture, given the wide variation in slice thickness across scans. Transfer learning was implemented using weights from a model pretrained on ImageNet (http://image-net.org/),[12] a con-

Table 2: A total of 739 scans were distributed into a training set, a validation set, and a held-out test set

	Training	Validation	Test	Total
MB	242	34	55	331
DMG	88	10	24	122
EP	83	13	15	111
PA	114	20	41	175
Total	527	77	135	739

Radiologist Interpretation

Board-certified attending radiologists with Certificates of Added Qualification in either Pediatric Radiology (J.S. with >10 years' experience; M.P.L. with >5 years' experience) or Neuroradiology (M.I. with >5 years' experience; E.T. with >2 years' experience) were given all T2 scans from the held-out test set and asked to detect tumors and select pathology among the 4 subtypes (MB, EP, PA, DMG). Radiologists were blinded to the ground truth labels and other clinical information and allowed to interpret at their own pace. They were permitted to window the scans and view in all orientations (axial, sagittal, or coronal).



MB = Medulloblastoma

PA = Pilocytic astrocytoma

DMA = Diffuse midline glyoma

EP = Ependymoma

CAMs = class actrivation maps

The model suggest a degree of probability of specific tumor type, which may o may not coincide with the reference truth or the expert radiologist opinion

FIG 2. CAMs depicting the areas of the input slice that the model preferentially emphasizes when predicting tumor subtype on individual scan slices. The *upper row* of each subpanel shows the T2 slice with tumor areas manually denoted (*upper left*) and CAM overlay of the most confident prediction of the model (*upper right*). The *lower row* of each panel shows less confident predictions. Examples of correct predictions of PA (*A*) and MB (*B*) and incorrect predictions of PA (*C*) and MB (*D*) are shown.

AJNR Am J Neuroradiol 41:1718–25 Sep 2020

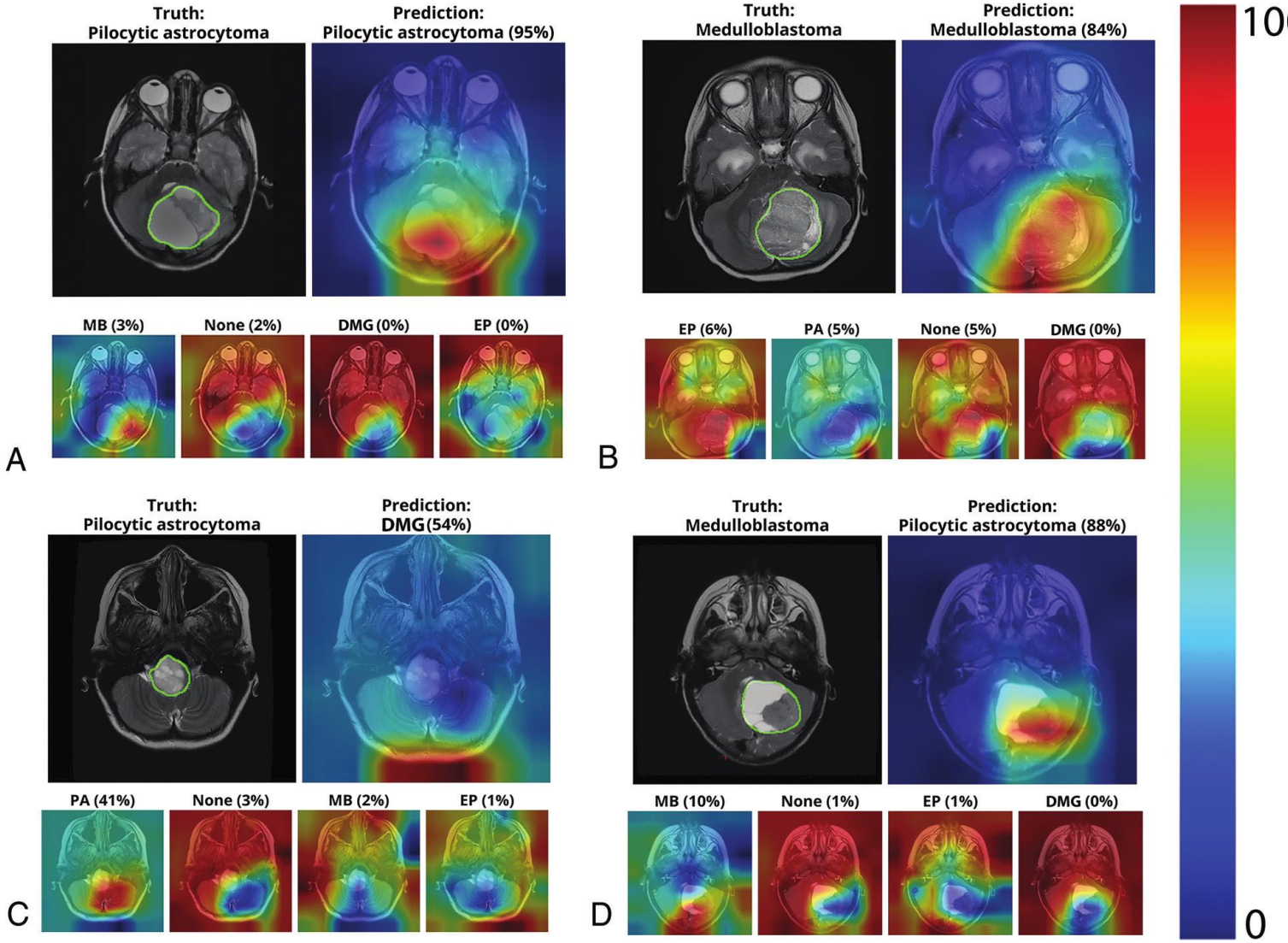


FIG 2. CAMs depicting the areas of the input slice that the model preferentially emphasizes when predicting tumor subtype on individual scan slices. The *upper row* of each subpanel shows the T2 slice with tumor areas manually denoted (*upper left*) and CAM overlay of the most confident prediction of the model (*upper right*). The *lower row* of each panel shows less confident predictions. Examples of correct predictions of PA (*A*) and MB (*B*) and incorrect predictions of PA (*C*) and MB (*D*) are shown.

CAMs = class activation maps

Class Activation Maps for Discriminative Localization of Tumor Type

Internal operations of deep learning algorithms often appear opaque and have been referred to as a "black box." Post hoc approaches for interpreting results have been described, such as using class activation maps (CAMs) to improve transparency and understanding of the model.[14] CAMs can serve as a quality assurance tool such that they highlight image regions relevant to the model's prediction and denote the model's confidence in the prediction but are not intended to precisely segment tumor voxels.[15] We implemented CAMs to visualize which regions of the image were most contributory to model prediction (Fig 2).[16] Qualitatively, pixels in close vicinity to the tumor appeared to strongly influence correct predictions, whereas incorrect predictions showed scattered CAMs that prioritized pixels in non-tumor regions. Because CAMs are not intended to provide perfect segmentations of tumor boundaries, we performed additional analyses to evaluate whether CAM mismatch correlated with the softmax score. The CAM for each slice was thresholded so that

Important and potentially useful strategy also in liver tumors to help EXPLANABILITY

AJNR Am J Neuroradiol 41:1718–25 Sep 2020

In the setting of liver cancer the different tumor types could be replaced by the different LI RADS classes

Table 3: Comparison of tumor detection and classification results between the deep learning model and radiologists[a]

	Tumor Detection			Tumor Classification		
	Sensitivity	Specificity	*P*	Accuracy	F_1 Score	*P*
Model	0.96	1.00	–	0.92	0.80	–
Radiologist average	0.99	0.98	–	0.87	0.75	–
Radiologist A	1.00	1.00	.06	0.95	0.89	.09
Radiologist B	0.99	0.97	1.00	0.89	0.79	.24
Radiologist C	0.99	0.98	1.00	0.79	0.61	<.01
Radiologist D	0.98	0.98	.73	0.84	0.70	<.01

Note:— –indicates n/a.

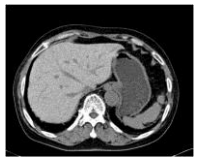
[a] *P* value calculated using the McNemar test comparing the model with individual radiologists.

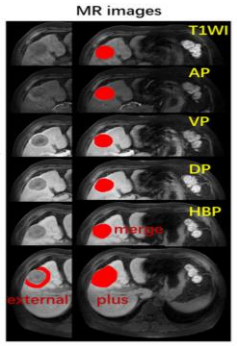
AJNR Am J Neuroradiol 41:1718–25 Sep 2020

Interobserver variability is known and accepted to a certain degree, but each observer is set responsible for his own tasks. But with AI models one «operator» servers the second. Explanability of AI models are needed in order to understand discrepancy

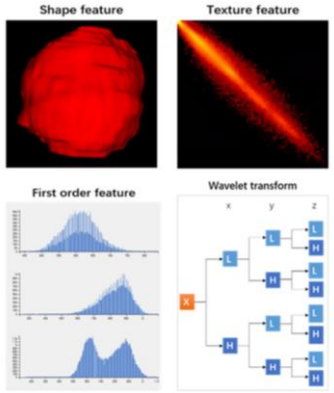
Radiomics VS Deep Learning

Radiomics



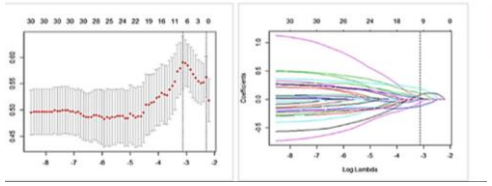


ROI Extraction

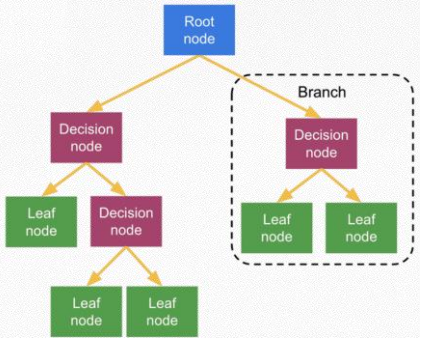


Feature Extraction

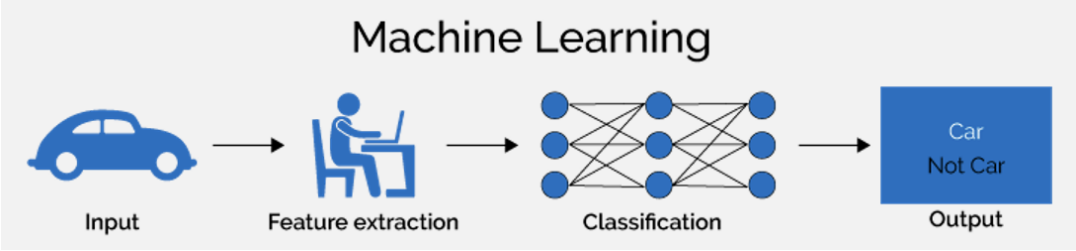
- Intraclass correlation coefficients
- mRMR
- LASSO
- Stepwise regression selection with AIC



Feature Selection



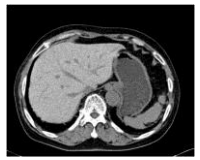
Model Building

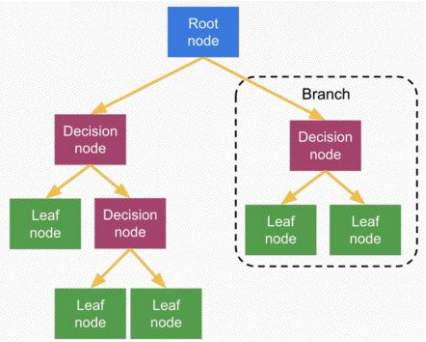


Courtesy of Prof. Greco

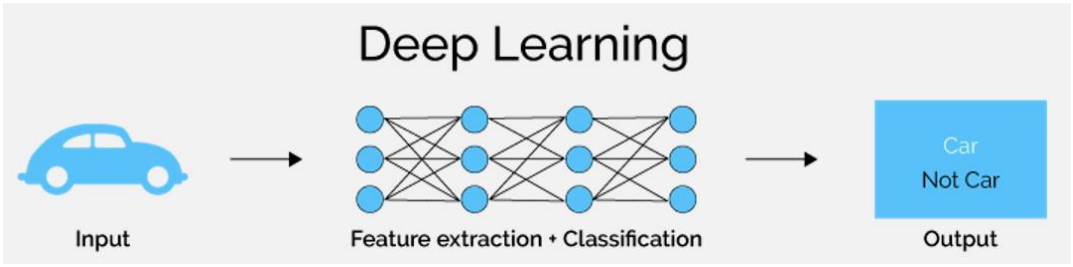
Radiomics VS Deep Learning

Deep learning: better performance (?) (lower explanability?)





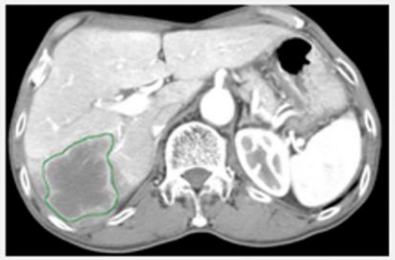
Model Building



Courtesy of Prof. Greco

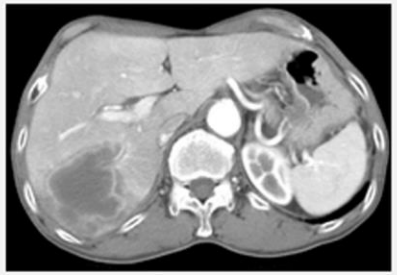
Possible interobserver variability impacting model development

Segmentation



Define regions-of-interest

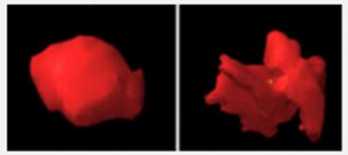
Preprocessing



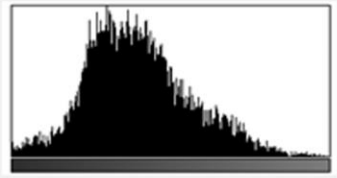
Interpolation
Normalization

Feature extraction

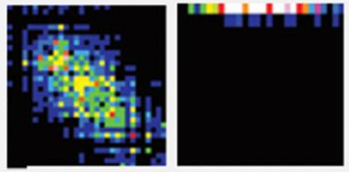
✓ **Morphologic features**



✓ **First-order features**

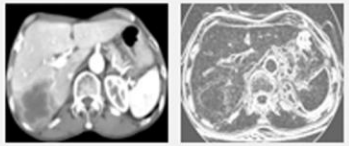


✓ **Texture features**



GLCM GLRLM

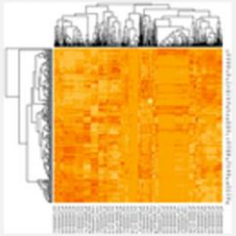
✓ **Higher-order features**



Texture features from filtered images

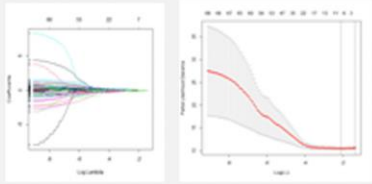
Feature selection

Excluding unreliable features
Selecting informative features from correlated features



Modeling

✓ **Penalized regression**

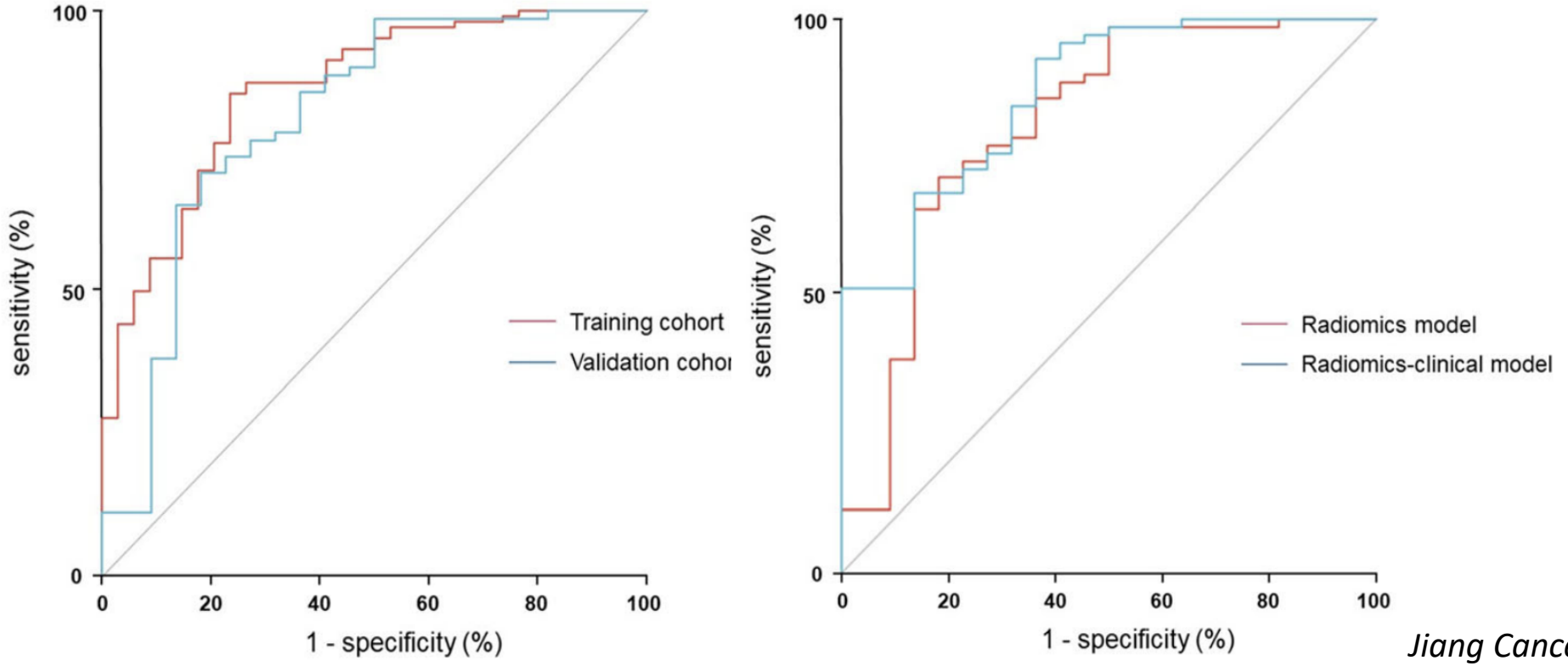


✓ **Machine learning**

Classification of nodules at risk for HCC. Not always radiomics signature alone perform better than other classifications.

Radiomics versus LI RADS

229 pathologically confirmed nodules (173 HCCs) in 211 patients
AUCs of the radiomics signature (0.810), LI-RADS (0.841) and EASL criteria (0.811)



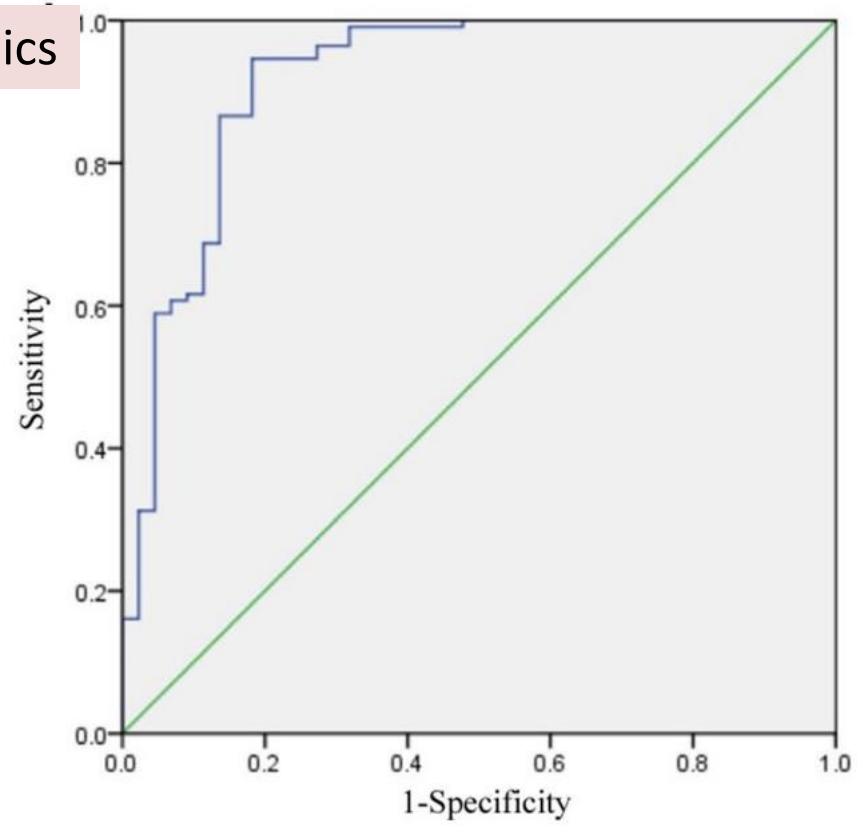
Jiang Cancer Imaging 2019

Classification of nodules at risk for HCC. Radiomics signature performance may be improved by its addition on existing clinical informations

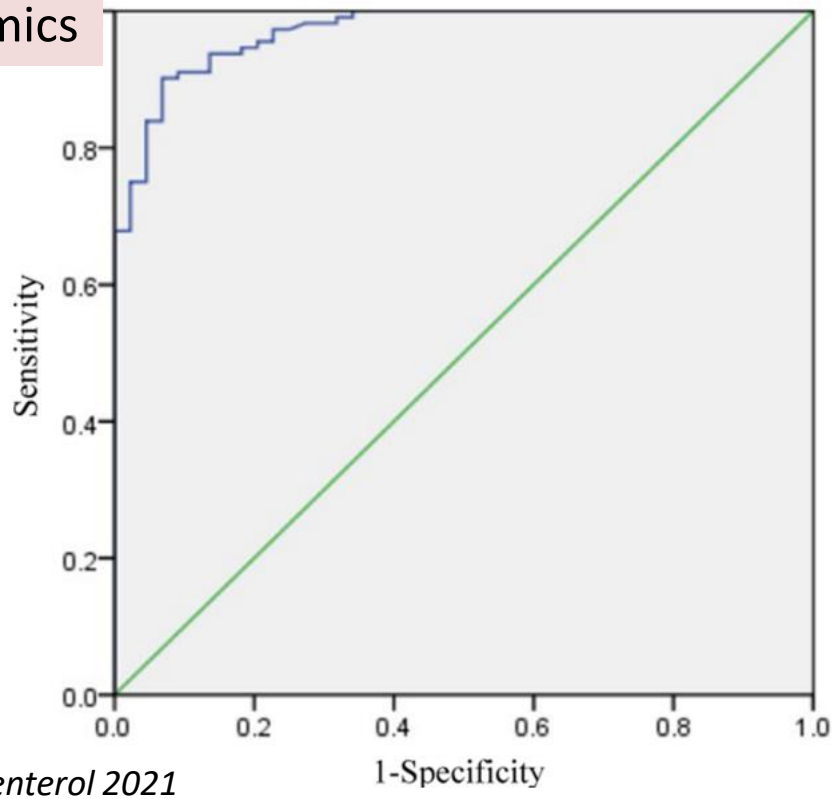
Radiomics with LI RADS

150 cirrhotic patients
Small liver nodules (HCC, 112; non malignant nodules, 44)
8 features: radiomics signature extracted from T1-W, T2W, and ADC

Radiomics



LI-RADS + radiomics



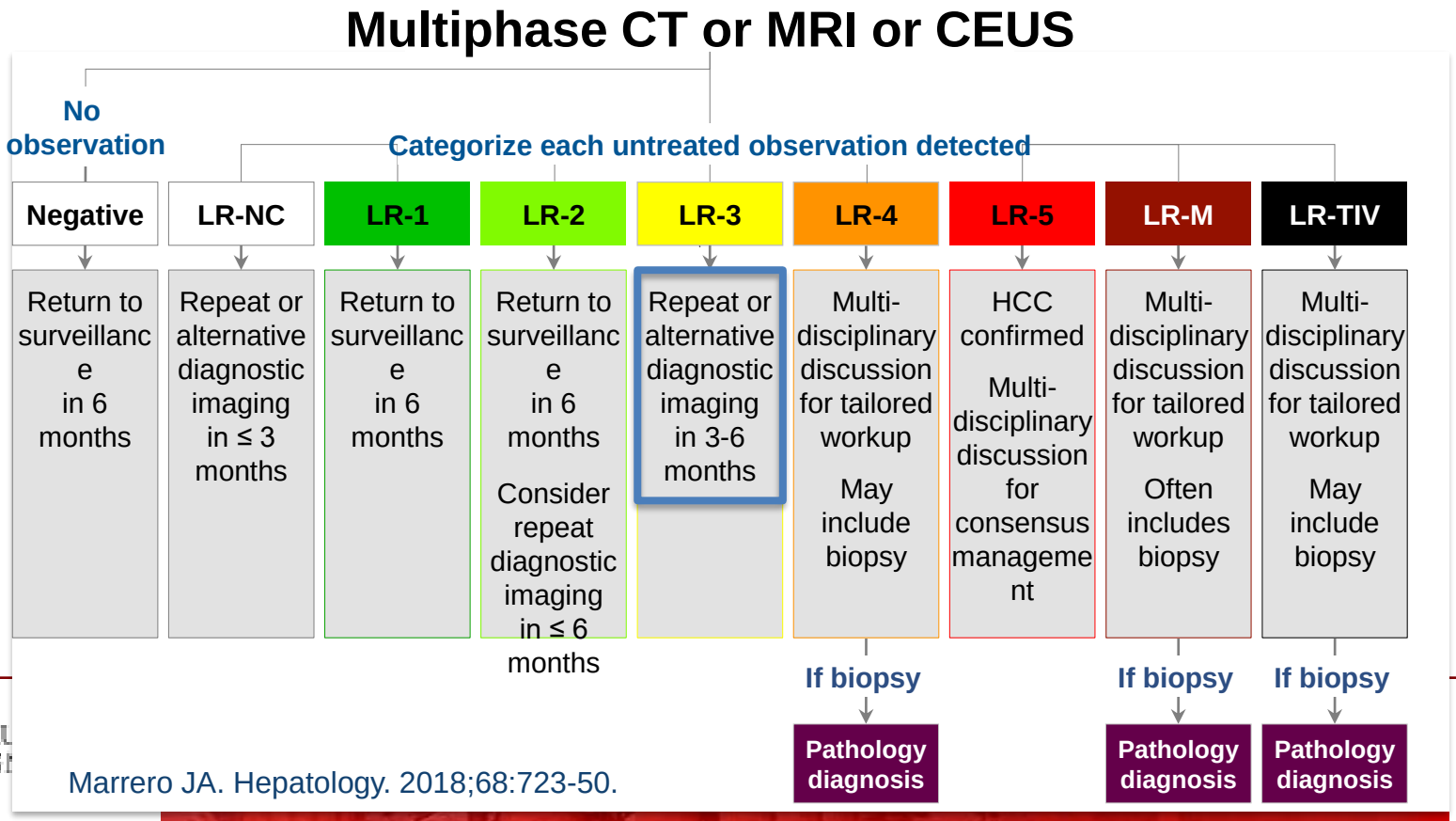
Zong BMC Gastroenterol 2021

Possible exemplary hurdle in the implementation of AI models into a liver tumor setting. Responsibility and accessibility

E.g. a 22 mm FLL detected close to the right portal vein branch. Not fully meeting LR5 class criteria (i.e. HCC)

The new hypothetical model says HCC 80%, HGDN 12%, LDGN 8%

Radiologist concludes: L3, likely HGDN. Location makes biopsy not easy. Consequently, no treatment is planned at the moment. Patient changes, however, follow up interval from 6 (surveillance) to 4 mos (enhanced surveillance)



Marrero JA. Hepatology. 2018;68:723-50.

At the next follow up in 4 months: nodule enlarged to 3 cm, now with features of HCC but with onset of portal vein infiltration.

Should the patient have access to the model conclusion (already at report collection or only accessible upon request)?

Is the radiologist held more responsible for having taken a discordant (potentially wrong) recommendation compared with the past when no AI tools were available?

APPLICABILITY

In order to work efficiently the AI model should be applied to a population similar to that utilized to train and develop the AI.

E.g. (example found in internet) AI detection of skin cancer

Skin cancer is most common in Caucasian and most AI are trained accordingly.
How does this apply in Asian or African or other. The user must be adequately be informed?

AJNR Am J Neuroradiol 41:1718–25 Sep 2020

APPLICABILITY

In order to work efficiently the AI model should be applied to a population similar to that utilized to train and develop the AI.

Which population was utilized to train a possible HCC identification AI model or to build a prognostic model? Are the information provided to the operator?

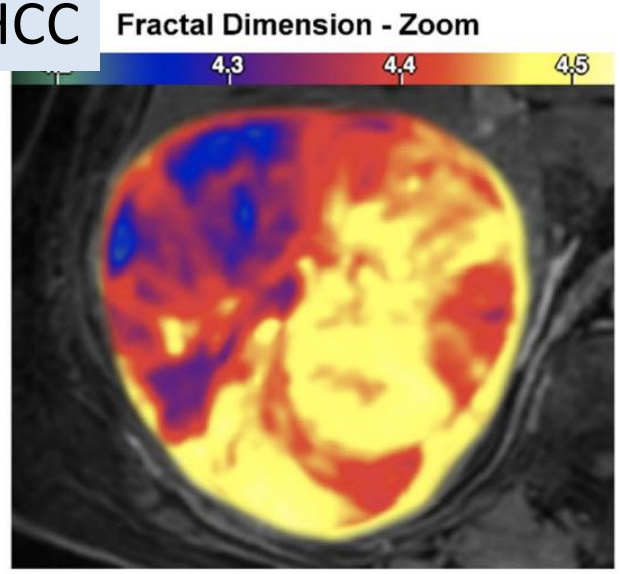
- Only cirrhosis or AASLD criteria?
- HBV predominant? NASH predominant? Aflatoxin exposure?
- Patients from which geographycal region?
- If model trained on pathology confirmed (usually resected) specimens the patient population may favor patient with preserved liver function
- Open question: better to develop on targeted population for training and apply only to the same population or have a much larger heterogeneous population
- Should the degree of certainty expressed based on Applicability?
- Can (or must) the applicability be coded?

APPLICABILITY

63 patients with pathologically characterized HCA and HCC in non-cirrhotic livers

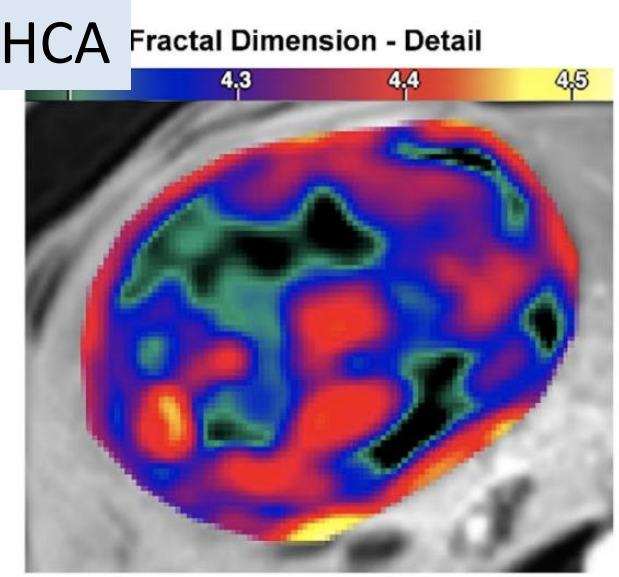
Fractal analysis showed differences between lesion subtypes (multi-class AUC=0.90, p<001)
Sensitivity and specifcity 43% and 47% for qualitative MRI features
96% and 68% when adding fractal analysis.

HCC

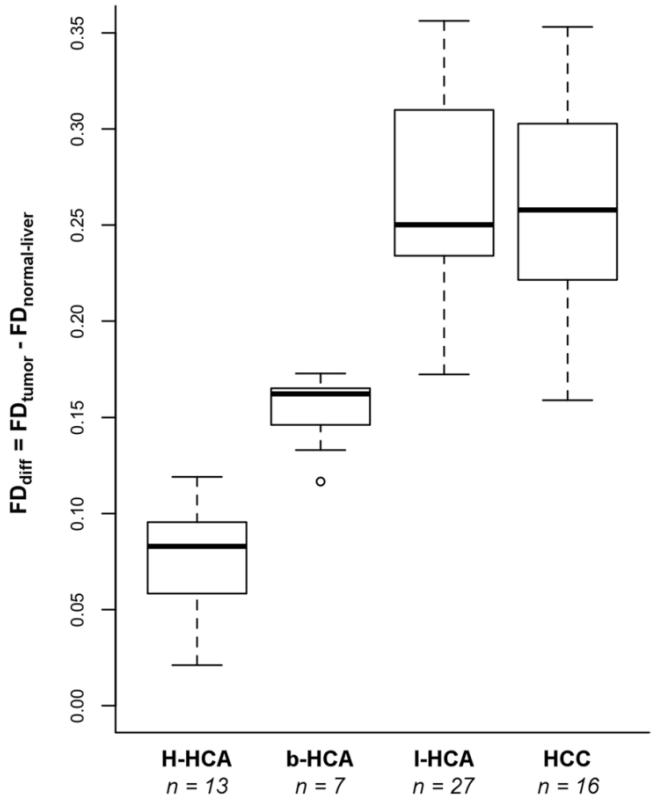


global FD_{diff} = 0.222

B-HCA



global FD_{diff} = 0.133



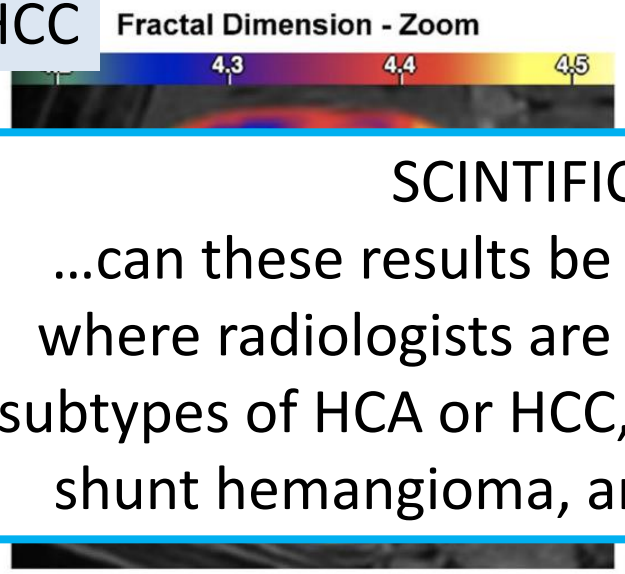
Michallek Isight Imaging 2022

APPLICABILITY

63 patients with pathologically characterized HCA and HCC in non-cirrhotic livers

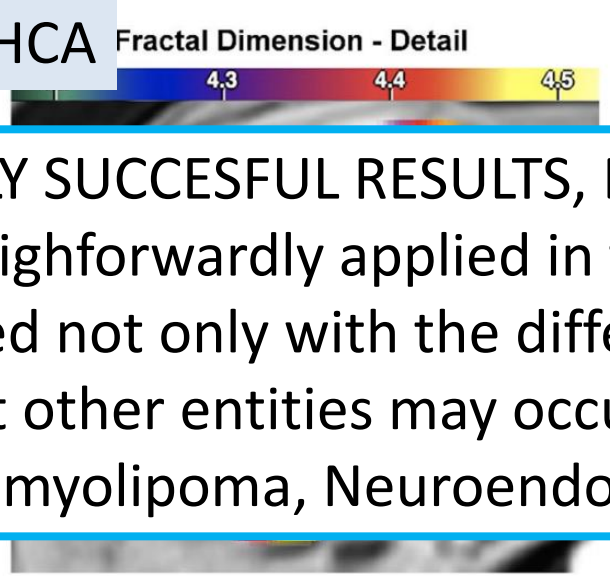
Fractal analysis showed differences between lesion subtypes (multi-class AUC=0.90, p<001)
Sensitivity and specifcity 43% and 47% for qualitative MRI features
96% and 68% when adding fractal analysis.

HCC



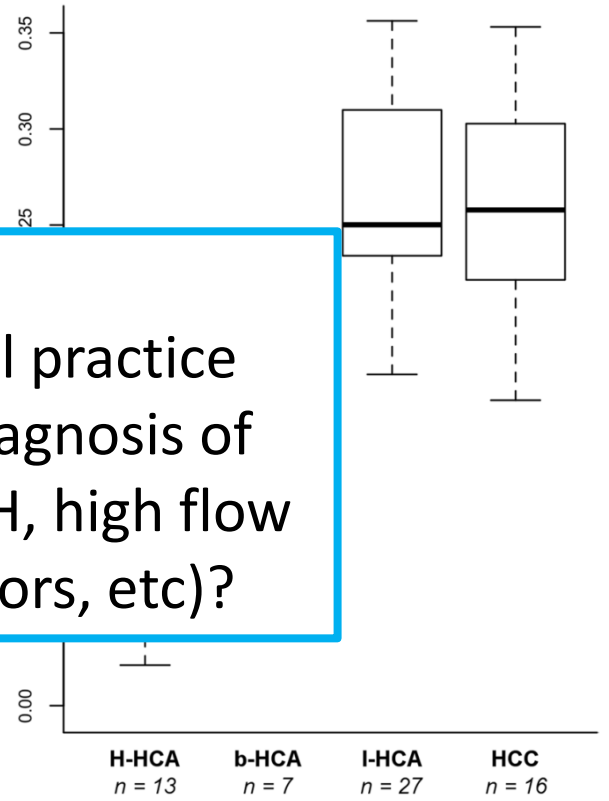
global FD_{diff} = 0.222

B-HCA



global FD_{diff} = 0.133

SCINTIFICALLY SUCCESFUL RESULTS, BUT….
…can these results be straighforwardly applied in the clinical practice where radiologists are faced not only with the differential diagnosis of subtypes of HCA or HCC, but other entities may occur (e.g FNH, high flow shunt hemangioma, angiomyolipoma, Neuroendocrine tumors, etc)?



Michallek Isight Imaging 2022

Gastroenterology 2022;162:1948–1961

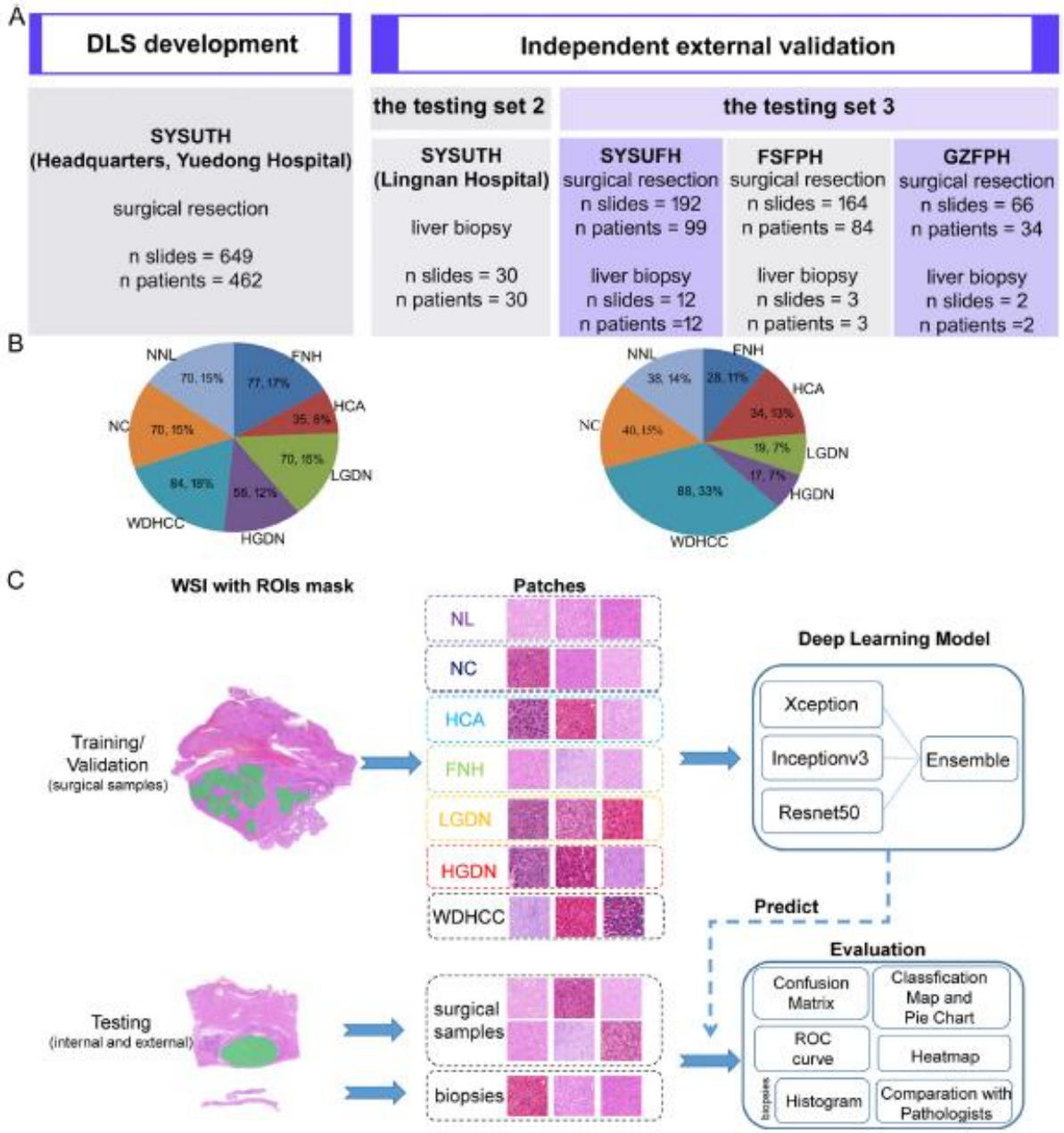
CLINICAL—LIVER

Deep Learning-Based Classification of Hepatocellular Nodular Lesions on Whole-Slide Histopathologic Images

Na Cheng,[1] Yong Ren,[2,3] Jing Zhou,[1] Yiwang Zhang,[1] Deyu Wang,[1] Xiaofang Zhang,[1] Bing Chen,[4] Fang Liu,[5] Jin Lv,[5] Qinghua Cao,[6] Sijin Chen,[1] Hong Du,[7] Dayang Hui,[1] Zijin Weng,[1] Qiong Liang,[1] Bojin Su,[1] Luying Tang,[8] Lanqing Han,[3] Jianning Chen,[1] and Chunkui Shao[1]

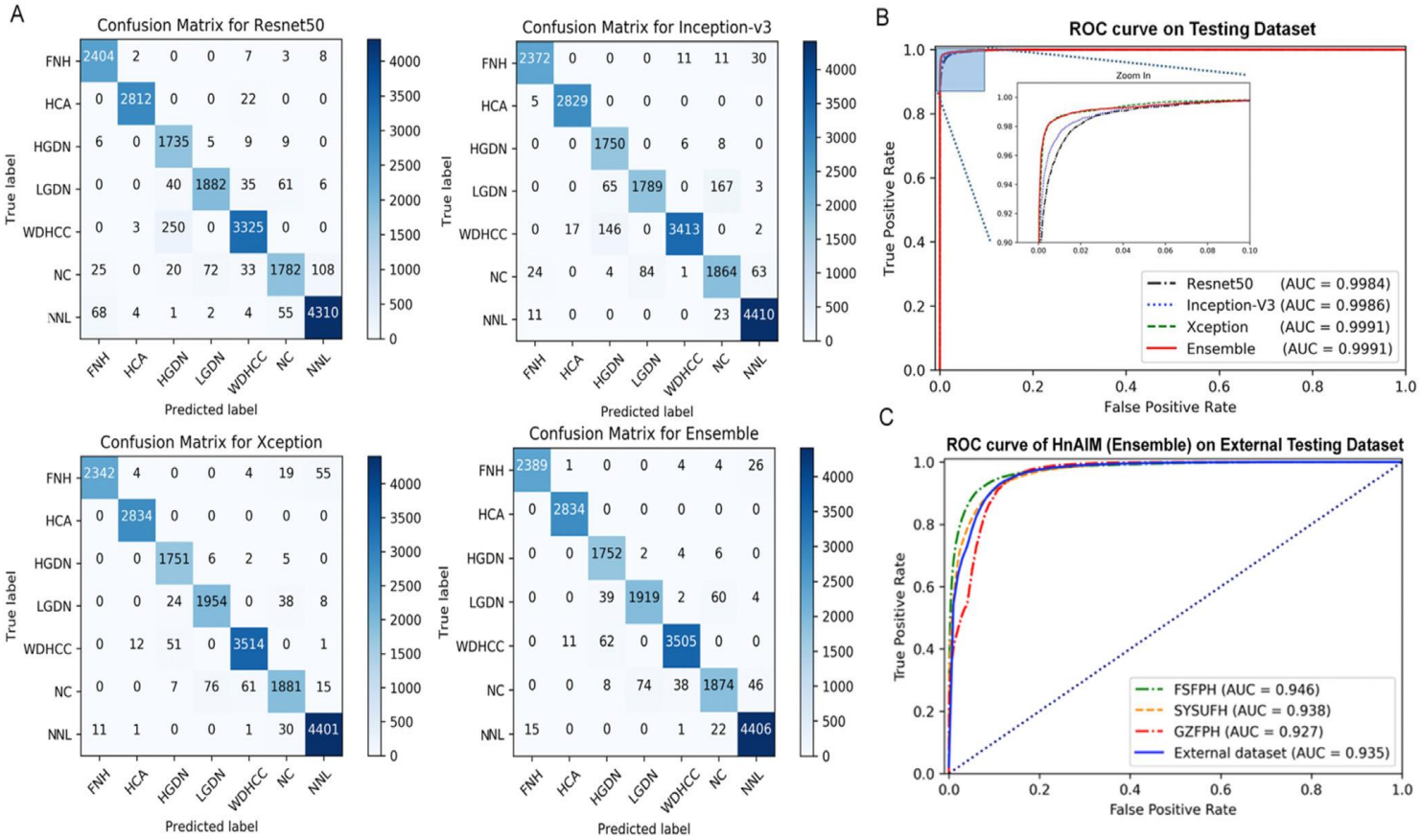
AIMS: develop a deep learning system to improve the histopathologic diagnosis of HNLs (WD-HCC, HGDN, low-grade DN, focal nodular hyperplasia, hepatocellular adenoma), and background tissues (nodular cirrhosis, normal liver tissue).

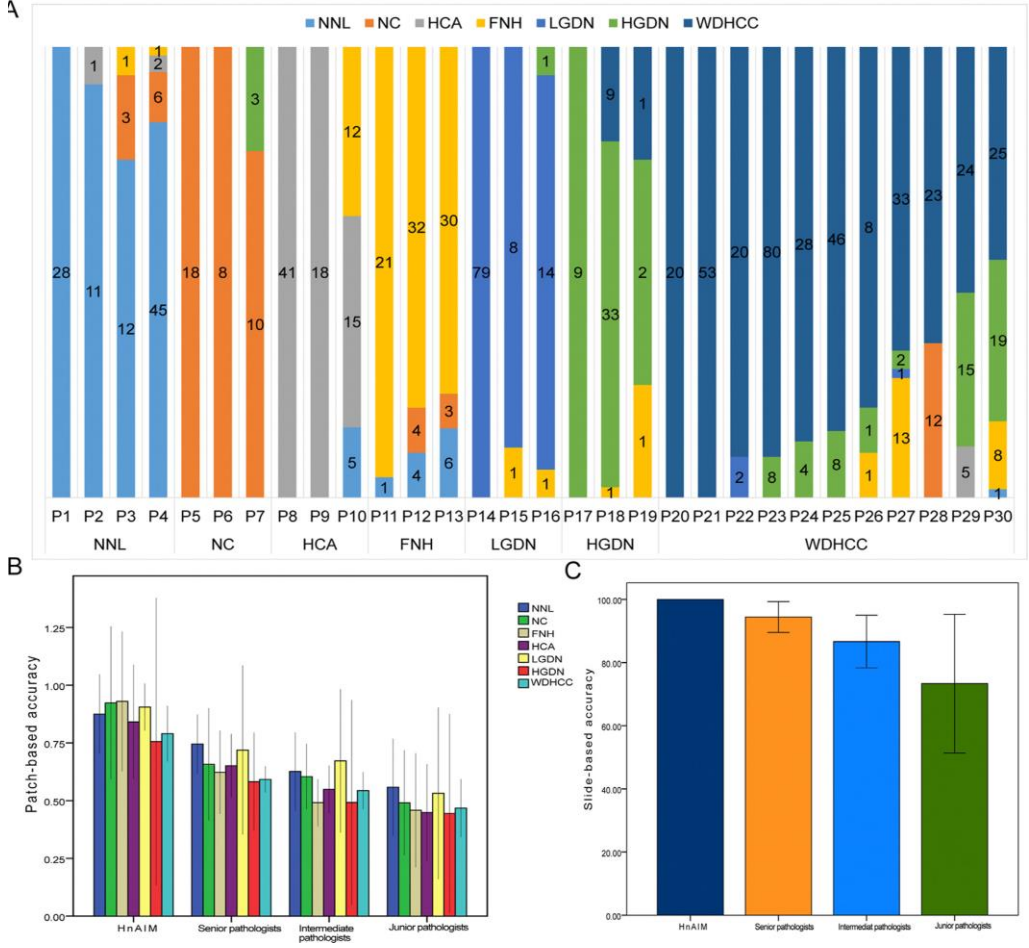
METHODS: Four deep neural networks were used. Their performances were evaluated by confusion matrix, receiver operating characteristic curve, classification map, and heat map. The predictive efficiency of the optimal model was further verified by comparing with that of pathologists.



Performance of deep learning models

The Xception and the Ensemble models both performed the best, with an AUC value of 0.9991, indicating models were trained with high prediction accuracy. Due to the higher sensitivity and F1 score, the Ensemble model was our optimal model, named HnAIM finally.

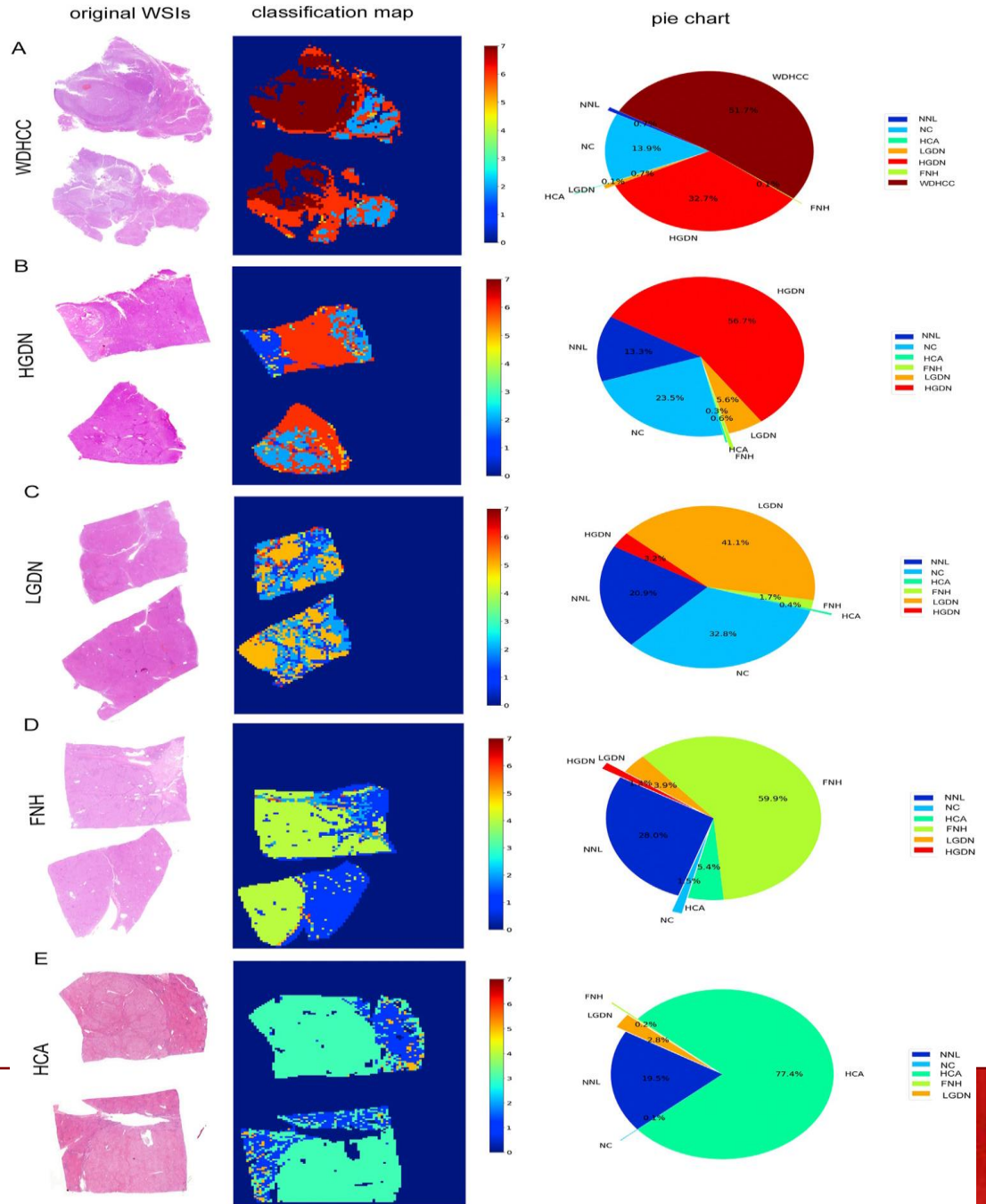




Performance of HnAIM in biopsy specimens and comparison with pathologists.

For validation purposes, an additional external data setwas used to evaluate the diagnostic performance of our HnAIM.

WSI-level panoramic classification map of surgical sample



Classification maps were constructed from model's predictions of corresponding patches.
Colors from blue to red meant different liver lesions.
For NC, LGDN, HGDN, and WDHCC, gradually deepening color even indicated increased degree of malignancy

Should the patient become aware of the suggestions of the AI models?
Should the classification maps be part of the report?

Issues recommended to be clarified before clinical use of AI tools in diagnosis of lesions at risk for HCC

- **Mandatory or optional use?** (I would prefer optional, at least until it can be verified that all conditions allowing their use are met and if the responsbility remains in the radiologist hands (he/she has to decide)
- **Applicability** is highly relevant (to be clarified how to to verify it).
  E.g. current patient part of a population similar to the one based on which the tool was developed? AI tools should clearly state their limits of applicability (full, partial, not applicable? Or express the degree of uncertainty).
- **Responsibility** of the radiologist? In case of discordance who is responsible. A way to refer the case to a third point of view or MDT is to be entitled?

Current main unmet **clinical** needs in hepatology to which AI could theoretically contribute for the daily practice:

- Non invasive precise fibrotic staging of chronic liver disease
- Non invasive distinction between NAFLD and NASH
- Improved non invasive diagnosis of HCC and of its molecular characterization
- Individual precise prediction of the risk of HCC development
- Improved histological diagnosis
- Prediction of the response of HCC to treatment and risk of recurrence
- Identification of complex / rare etiologies of liver lab exam abnormalities
- Identification of subphenotypes within current single etiology and corresponding natural course of disease and response to treatment

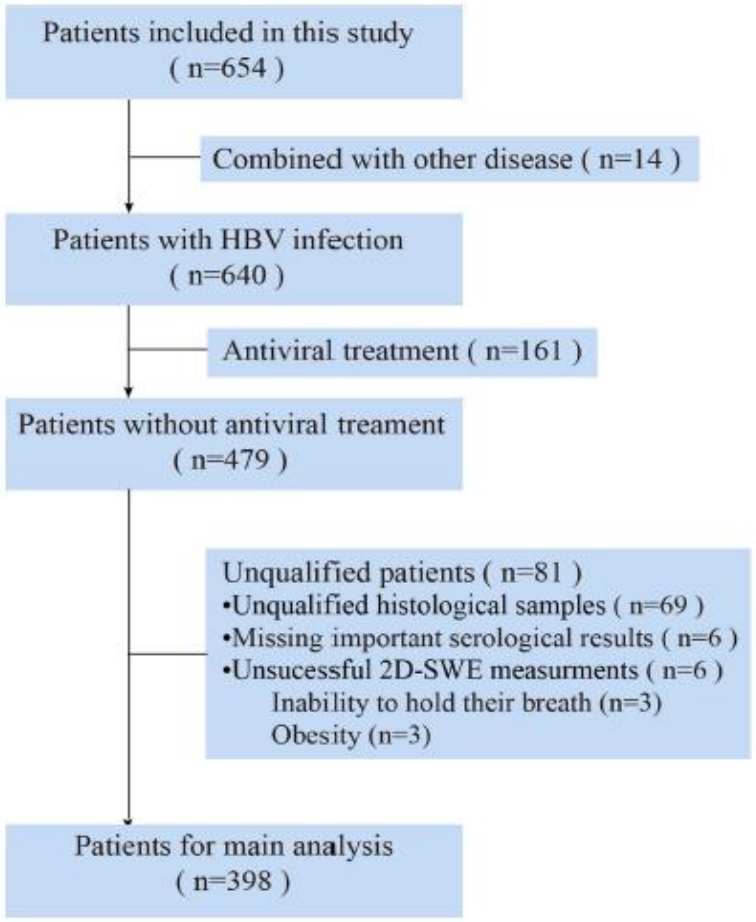
Multicenter Study > Gut. 2019 Apr;68(4):729-741. doi: 10.1136/gutjnl-2018-316204.
Epub 2018 May 5.

Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study

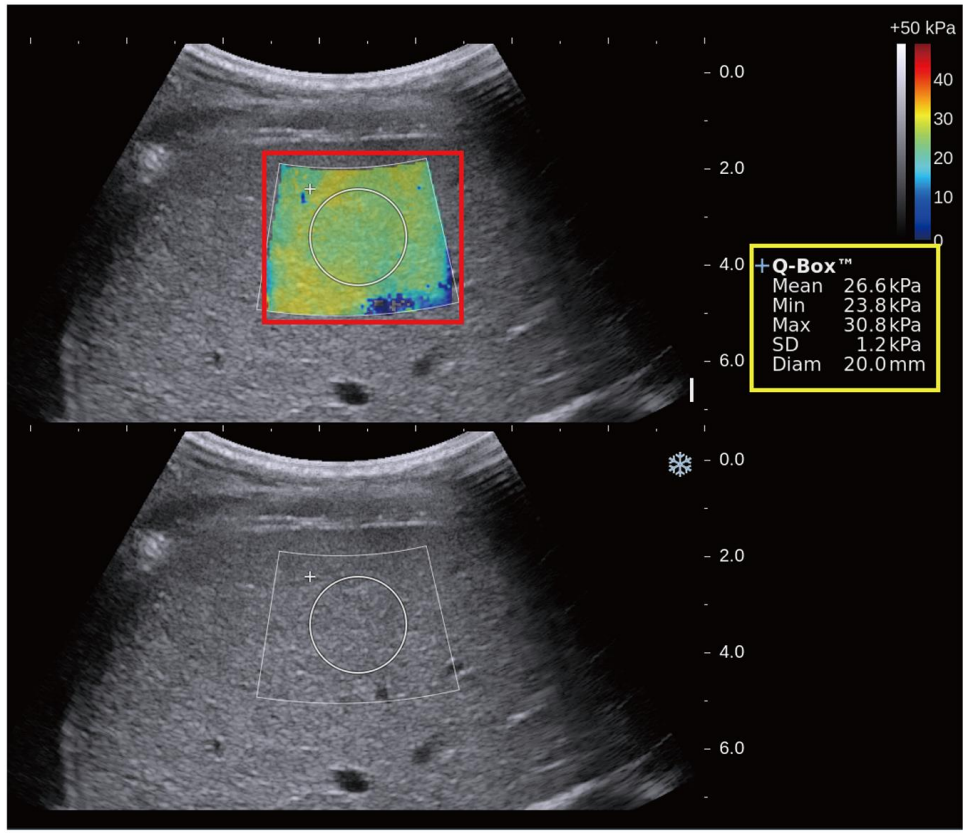
Kun Wang [#1,2], Xue Lu [#1], Hui Zhou [#2,3], Yongyan Gao [#4], Jian Zheng [1,5], Minghui Tong [6], Changjun Wu [7], Changzhu Liu [8], Liping Huang [9], Tian'an Jiang [10], Fankun Meng [11], Yongping Lu [12], Hong Ai [13], Xiao-Yan Xie [14], Li-Ping Yin [15], Ping Liang [3], Jie Tian [2,3], Rongqin Zheng [1]

AIMS: evaluate the performance of the deep learning Radiomics of elastography (DLRE) for assessing liver fibrosis stages.

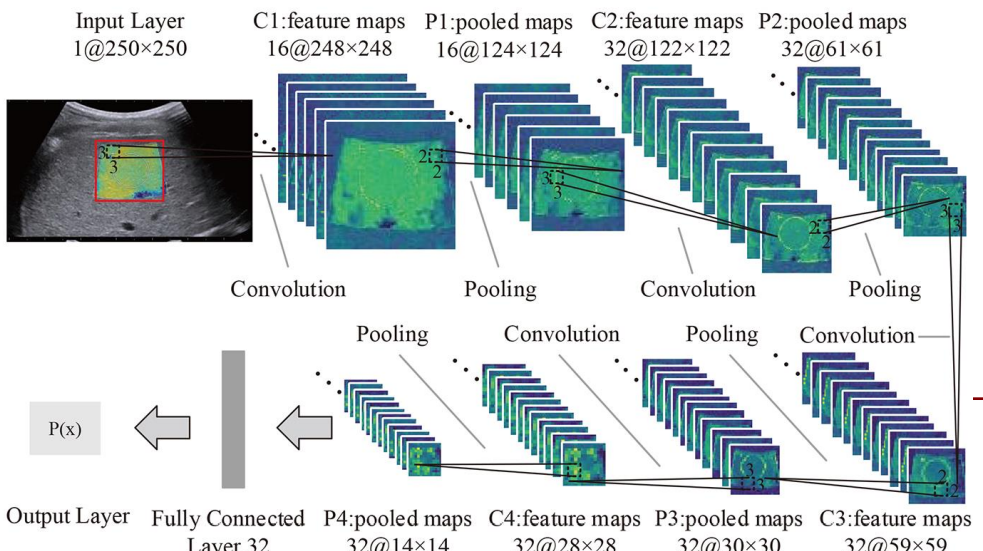
DLRE adopts the radiomic strategy for quantitative analysis of the heterogeneity in two-dimensional shear wave elastography (2D-SWE) images.



A

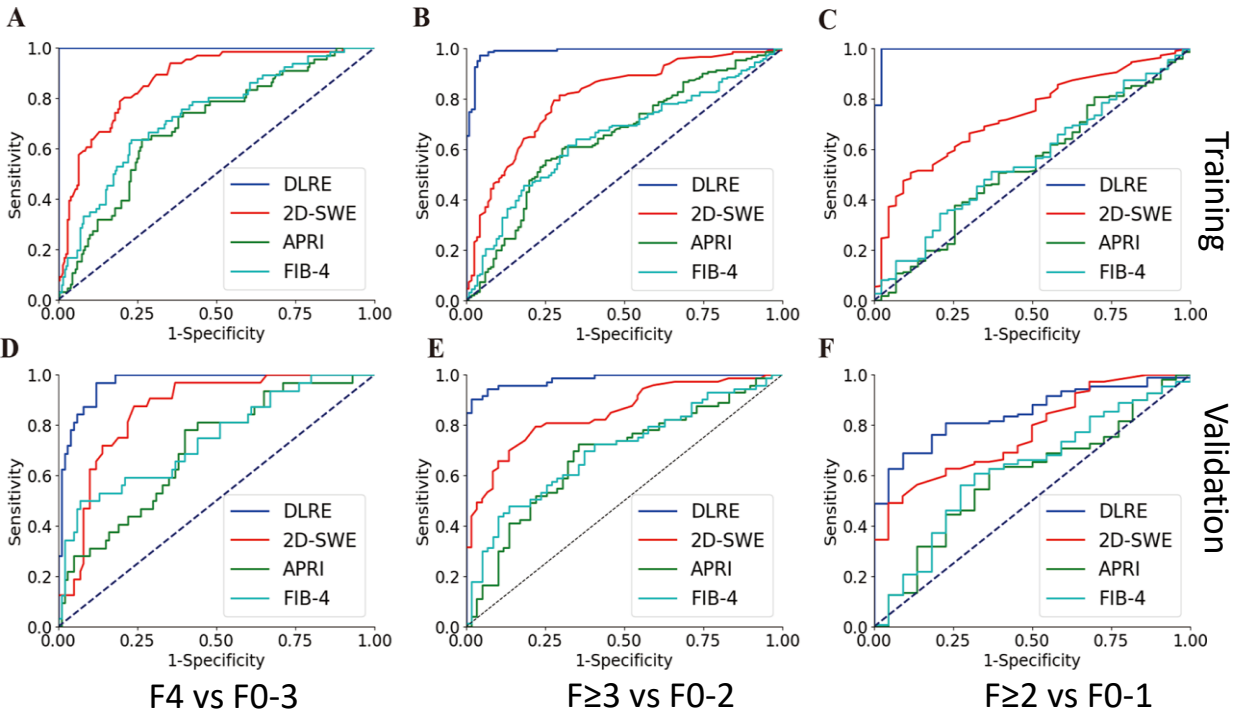


B



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

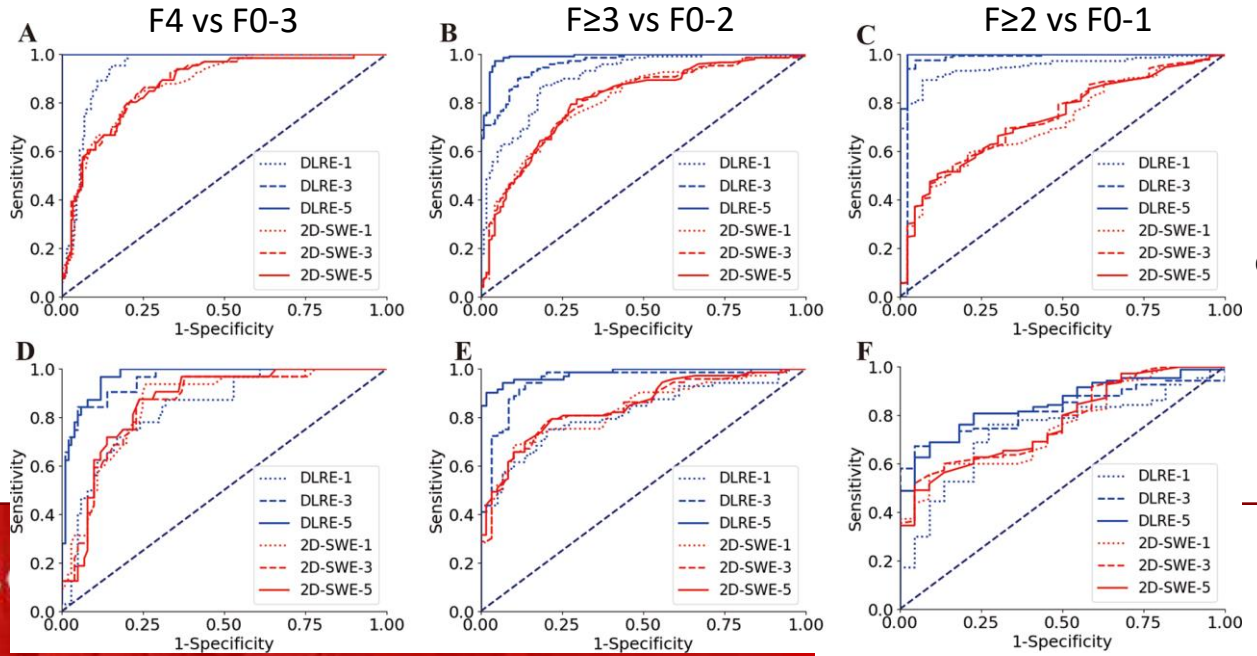
Overall diagnostic accuracy of DLRE in comparison with 2D-SWE, APRI and FIB-4



In the training cohort, AUCs of DLRE were significantly better than those of 2D-SWE in all stratifications. However, in the validation cohort, if more than 1 image was adopted, DLRE outperformed 2D-SWE in the stratification of F4 and ≥F3 (all p<0.01), but it did not offer significantly higher AUC for ≥F2.

In the training cohort, DLRE had the highest diagnostic accuracy compared with all other methods for classifying of F4, ≥F3 and ≥F2 (figure 3A–C), and differences of AUCs were all statistically significant (p<0.00).
In the validation cohort, AUCs of DLRE dropped slightly for the diagnosis of F4 and ≥F3 (figure 3D,E). However, the performance of DLRE for ≥F2 became much poorer than it was in the training cohort (figure 3F).

Diagnostic accuracy versus number of acquisitions: intrastrategy and interstrategy comparison of DLRE and 2D-SWE *(i.e. a higher number of elastograms improves accuracy)*



Current main unmet **clinical** needs in hepatology to which AI could theoretically contribute for the daily practice:

- Non invasive precise fibrotic staging of chronic liver disease
- Non invasive distinction between MASLD and MASH
- Improved non invasive diagnosis of HCC and of its molecular characterization
- Individual precise prediction of the risk of HCC development
- Improved histological diagnosis
- Prediction of the response of HCC to treatment and risk of recurrence
- Identification of complex / rare etiologies of liver lab exam abnormalities
- Identification of subphenotypes within current single etiology and corresponding natural course of disease and response to treatment

One example among the many.

Chang D et al. Machine learning models are superior to noninvasive tests in identifying clinically significant stages of NAFLD and NAFLD-related cirrhosis. Hepatology. 2022 Jul 9. doi: 10.1002/hep.32655

Histological stages of fibrosis (≥F2, ≥F3, and F4) were predicted using ML (Machine Learning) based on 17 clinical variables, FibroScan liver stiffness measurements, and Fibrosis-4 index (FIB-4).
All ML models had primarily higher accuracy and AUC con
versus FibroScan and FIB-4 for ≥F2, ≥F3, and F4 were 0.89
performed better in sens, spec, PPV, NPV than traditional

These models will have a clinical role if:

Cheaper
More widely applicable
More reproducible among operators and over time
More acceptable by the patients
Preserving the privacy of the patients
Largely more precise in distinguishing different stages th

At present AI models appear already well suited for a case finding approach for hepatology referral

AI models are not ready be used to assign a precise fibrosis stage

AI models increase accuracy to distinguish NAFLD from NASH but not to a level sufficient for individual diagnosis

Current main unmet **clinical** needs in hepatology to which AI could theoretically contribute for the daily practice:

- Non invasive precise fibrotic staging of chronic liver disease
- Non invasive distinction between MASLD and MASH
- Improved non invasive diagnosis of HCC and of its molecular characterization
- Individual precise prediction of the risk of HCC development
- Improved histological diagnosis
- Prediction of the response of HCC to treatment and risk of recurrence
- Identification of complex / rare etiologies of liver lab exam abnormalities
- Identification of subphenotypes within current single etiology and corresponding natural course of disease and response to treatment

Research Article
Hepatic and Biliary Cancer

JOURNAL OF HEPATOLOGY

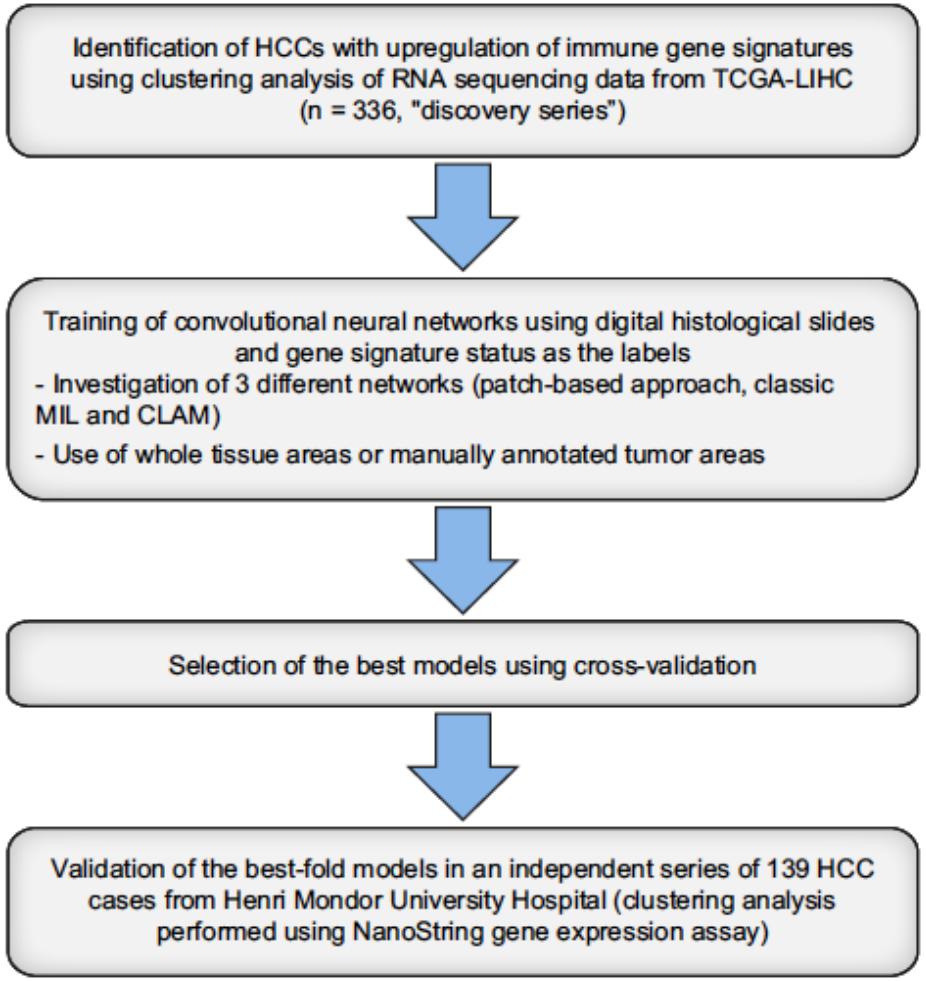
Artificial intelligence predicts immune and inflammatory gene signatures directly from hepatocellular carcinoma histology

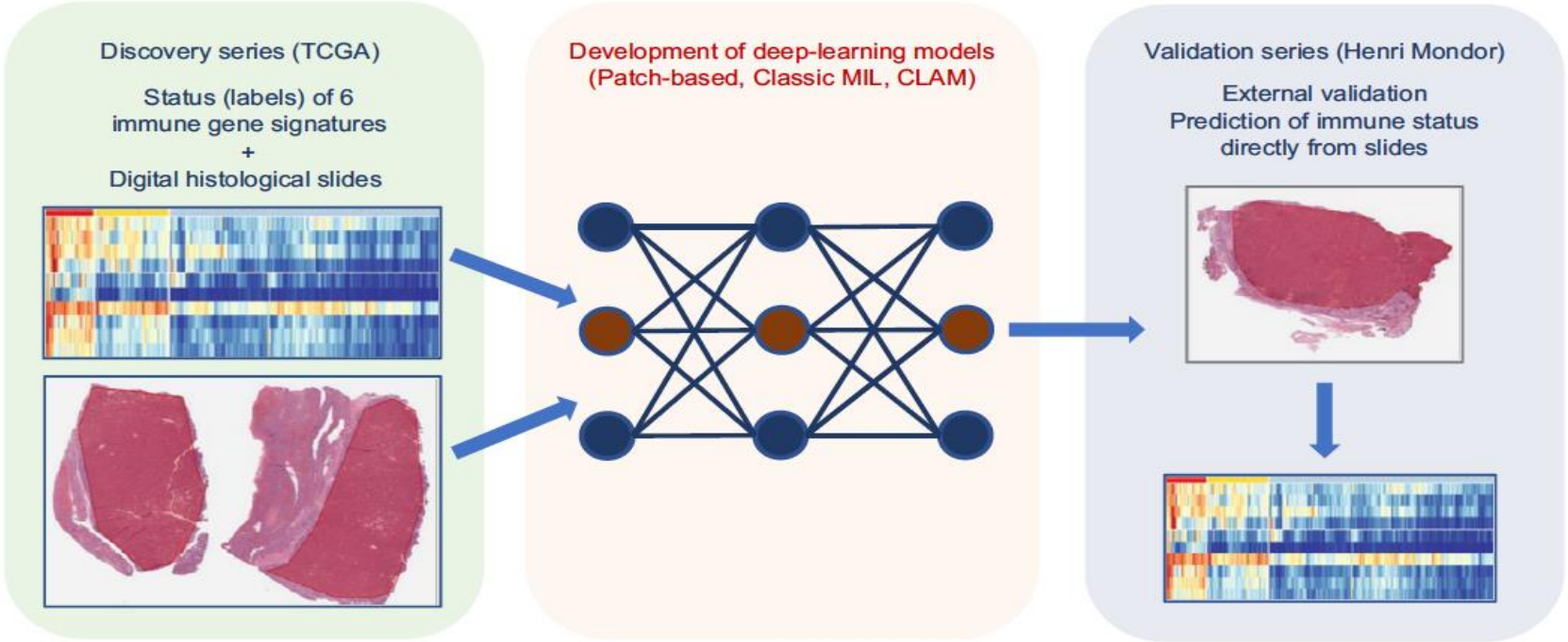
Qinghe Zeng[1,2,†], Christophe Klein[1,†], Stefano Caruso[3], Pascale Maille[4,5,6], Narmin Ghaffari Laleh[7,8], Daniele Sommacale[9], Alexis Laurent[9], Giuliana Amaddeo[10], David Gentien[11], Audrey Rapinat[11], Hélène Regnault[10], Cécile Charpy[4], Cong Trung Nguyen[5,6], Christophe Tournigand[12], Raffaele Brustia[9], Jean Michel Pawlotsky[5,6], Jakob Nikolas Kather[7,8], Maria Chiara Maiuri[1], Nicolas Loménie[2,#], Julien Calderaro[4,5,6,*,#]

Aims: Patients with HCC displaying overexpression of immune gene signatures are likely to be more sensitive to immunotherapy. The aim of the study was, using artificial intelligence (AI) on whole-slide digital histological images, to develop models able to predict the activation of 6 immune gene signatures.

Methods: AI models were trained in patients with HCC treated by surgical resection. Three deep learning approaches were investigated: patch-based, classic MIL and CLAM. Pathological reviewing of the most predictive tissue areas was performed for all gene signatures.

Flowchart of the study





- The CLAM model showed the best overall performance in the discovery series. Its best-fold areas under the receiver operating characteristic curves **(AUCs)** for the prediction of tumors with upregulation of the immune gene signatures ranged from **0.78 to 0.91**.
- The different models generalized well in the validation dataset with AUCs ranging from 0.81 to 0.92.
- Pathological analysis of highly predictive tissue areas showed enrichment in lymphocytes, plasma cells, and neutrophils.

Would this data be enough to choose a specific drug therapy?

Current main unmet **clinical** needs in hepatology to which AI could theoretically contribute for the daily practice:

- Non invasive precise fibrotic staging of chronic liver disease
- Non invasive distinction between MASLD and MASH
- Improved non invasive diagnosis of HCC and of its molecular characterization
- Individual precise prediction of the r
- Improved histological diagnosis
- Prediction of the response of HCC t
- Identification of complex / rare etiol
- Identification of subphenotypes within current single etiology and corresponding natural course of disease and response to treatment

Molecular characterization has still limited impact. The use is not foreseen in the short term but appears very appealing for the future (pending identification of molecular targeted effective drugs).

Current main unmet **clinical** needs in hepatology to which AI could theoretically contribute for the daily practice:

- Non invasive precise fibrotic staging of chronic liver disease
- Non invasive distinction between MASLD and MASH
- Improved non invasive diagnosis of HCC and of its molecular characterization
- Individual precise prediction of the risk of HCC development
- Improved histological diagnosis
- Prediction of the response of HCC to treatment and risk of recurrence
- Identification of complex / rare etiologies of liver lab exam abnormalities
- Identification of subphenotypes within current single etiology and corresponding natural course of disease and response to treatment

Research Article
Viral Hepatitis

JOURNAL OF HEPATOLOGY

An artificial intelligence model to predict hepatocellular carcinoma risk in Korean and Caucasian patients with chronic hepatitis B

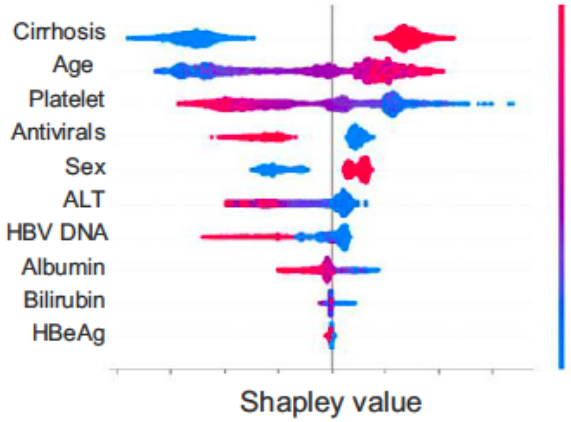
Hwi Young Kim[1,†], Pietro Lampertico[2,3,†], Joon Yeul Nam[4,†], Hyung-Chul Lee[5,†], Seung Up Kim[6], Dong Hyun Sinn[7], Yeon Seok Seo[8], Han Ah Lee[8,9], Soo Young Park[10], Young-Suk Lim[11], Eun Sun Jang[12], Eileen L. Yoon[9,13], Hyoung Su Kim[14], Sung Eun Kim[15], Sang Bong Ahn[16], Jae-Jun Shim[17], Soung Won Jeong[18], Yong Jin Jung[19], Joo Hyun Sohn[20], Yong Kyun Cho[21], Dae Won Jun[13], George N. Dalekos[22], Ramazan Idilman[23], Vana Sypsa[24], Thomas Berg[25], Maria Buti[26], Jose Luis Calleja[27], John Goulis[28], Spilios Manolakopoulos[29], Harry L.A. Janssen[30], Myoung-jin Jang[31], Yun Bin Lee[4], Yoon Jun Kim[4], Jung-Hwan Yoon[4], George V. Papatheodoridis[32,*,‡], Jeong-Hoon Lee[4,‡,*]

HIGHLIGHTS

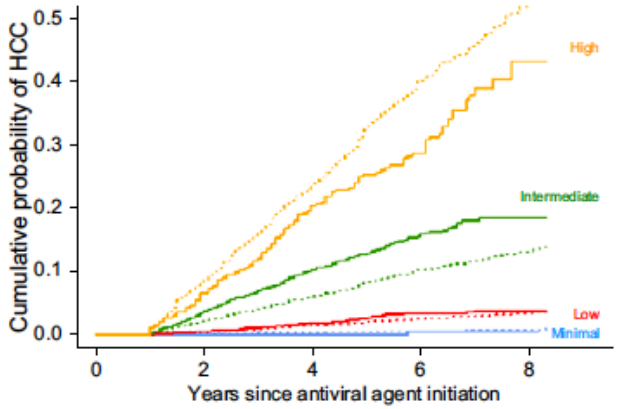
- A new HCC prediction model (PLAN-B) was developed using machine learning algorithms in antiviral-treated patients with chronic hepatitis B.
- The utility of the model was validated in independent Korean and Caucasian cohorts.
- PLAN-B comprises 10 baseline parameters: cirrhosis, age, platelet count, ETV/TDF, sex, serum ALT and HBV DNA, albumin and bilirubin levels, and HBeAg status.
- The PLAN-B model demonstrated satisfactory predictive performance for HCC development and outperformed other risk scores.

PLAN-B model for the prediction of HCC in patients with chronic hepatitis B

- Machine learning approaches (gradient-boosting machine algorithm)
- Entecavir or tenofovir-treated

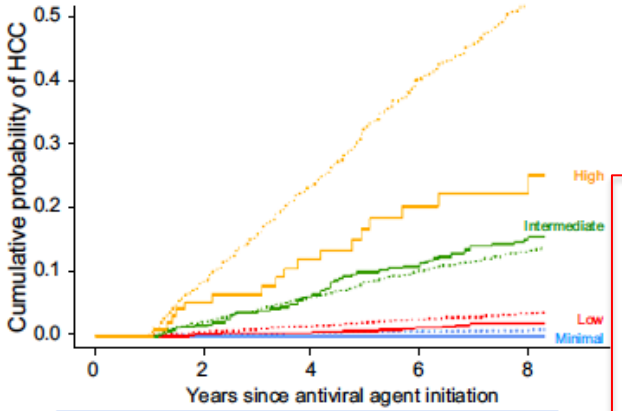


Derivation cohort
(Korea, n = 6,051)



Model	c-index	95% CI		p
		Lower	Upper	
PLAN-B	0.79	0.78	0.80	Ref.
PAGE-B	0.73	0.72	0.74	<0.001
mPAGE-B	0.75	0.74	0.76	0.004
REACH-B	0.63	0.61	0.64	<0.001
CU-HCC	0.72	0.71	0.73	<0.001

Korean validation cohort
(n = 5,817)



Model	c-index	95% CI		p
		Lower	Upper	
PLAN-B	0.81	0.79	0.83	Ref.
PAGE-B	0.75	0.73	0.77	<0.001
mPAGE-B	0.80	0.79	0.82	0.424
REACH-B	0.57	0.54	0.59	<0.001
CU-HCC	0.76	0.74	0.78	0.002

Caucasian validation cohort
(n = 1,640)

In this study, they developed an AI-based HCC risk prediction model for patients with CHB receiving a potent NA treatment using large-scale Korean and Caucasian cohort datasets.

The **PLAN-B model** showed satisfactory discriminant function (c-index, 0.82), which was significantly better than other models for both the Korean (PAGE-B, modified PAGE-B, REACH-B, and CU-HCC) and Caucasian validation cohorts (PAGE-B, REACH-B, and CU-HCC).

Current main unmet **clinical** needs in hepatology to which AI could theoretically contribute for the daily practice:

- Non invasive precise fibrotic staging of chronic liver disease
- Non invasive distinction between MASLD and MASH
- Improved non invasive diagnosis of HCC and of its molecular characterization
- Individual precise prediction of the risk of HCC development
- Improved histological diagnosis
- Prediction of the response of HCC to treatment and risk of recurrence
- Identification of complex / rare etiologies of liver lab exam abnormalities
- Identification of subphenotypes within current single etiology and corresponding natural course of disease and response to treatment

HEPATOLOGY

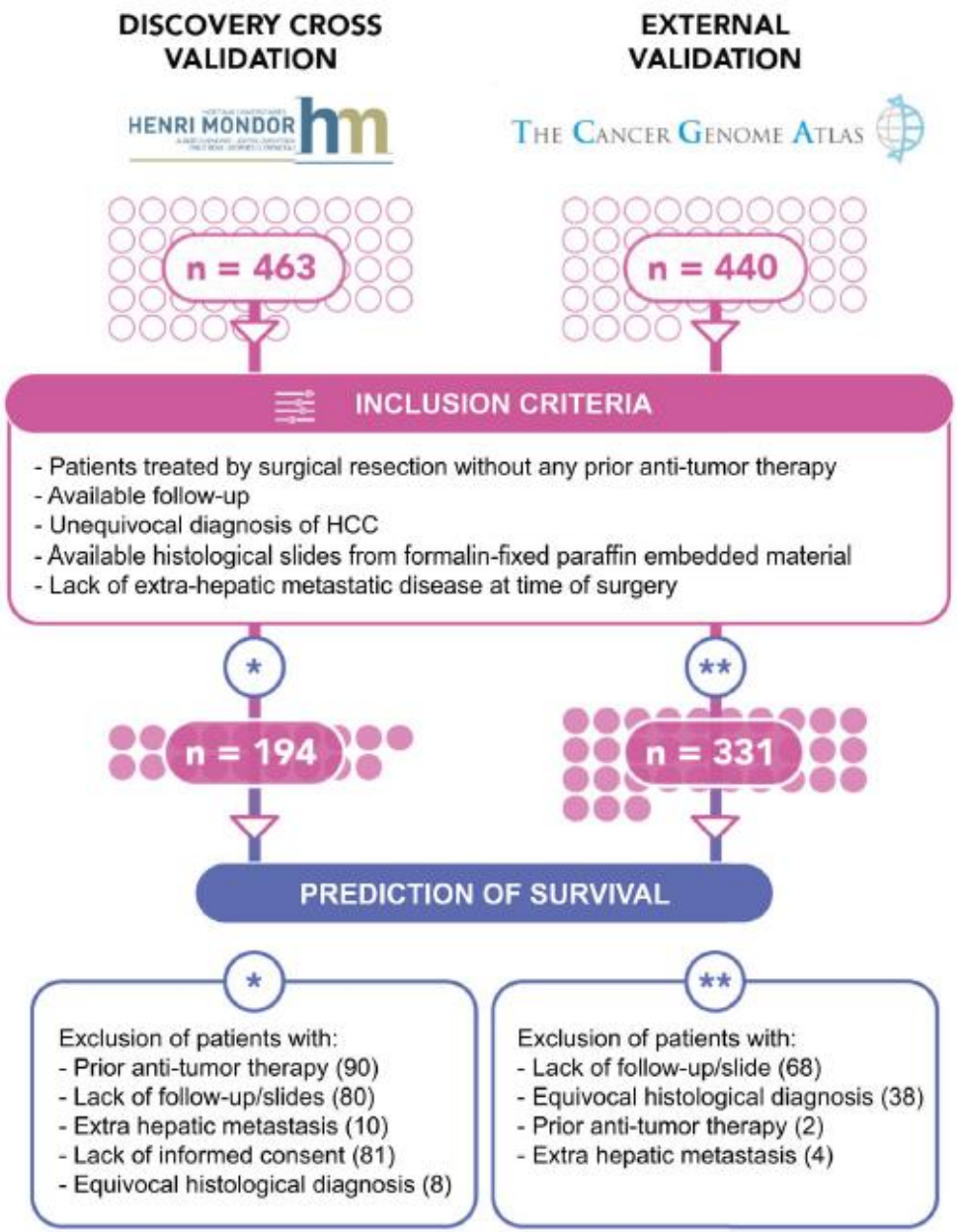
HEPATOLOGY, VOL. 72, NO. 6, 2020

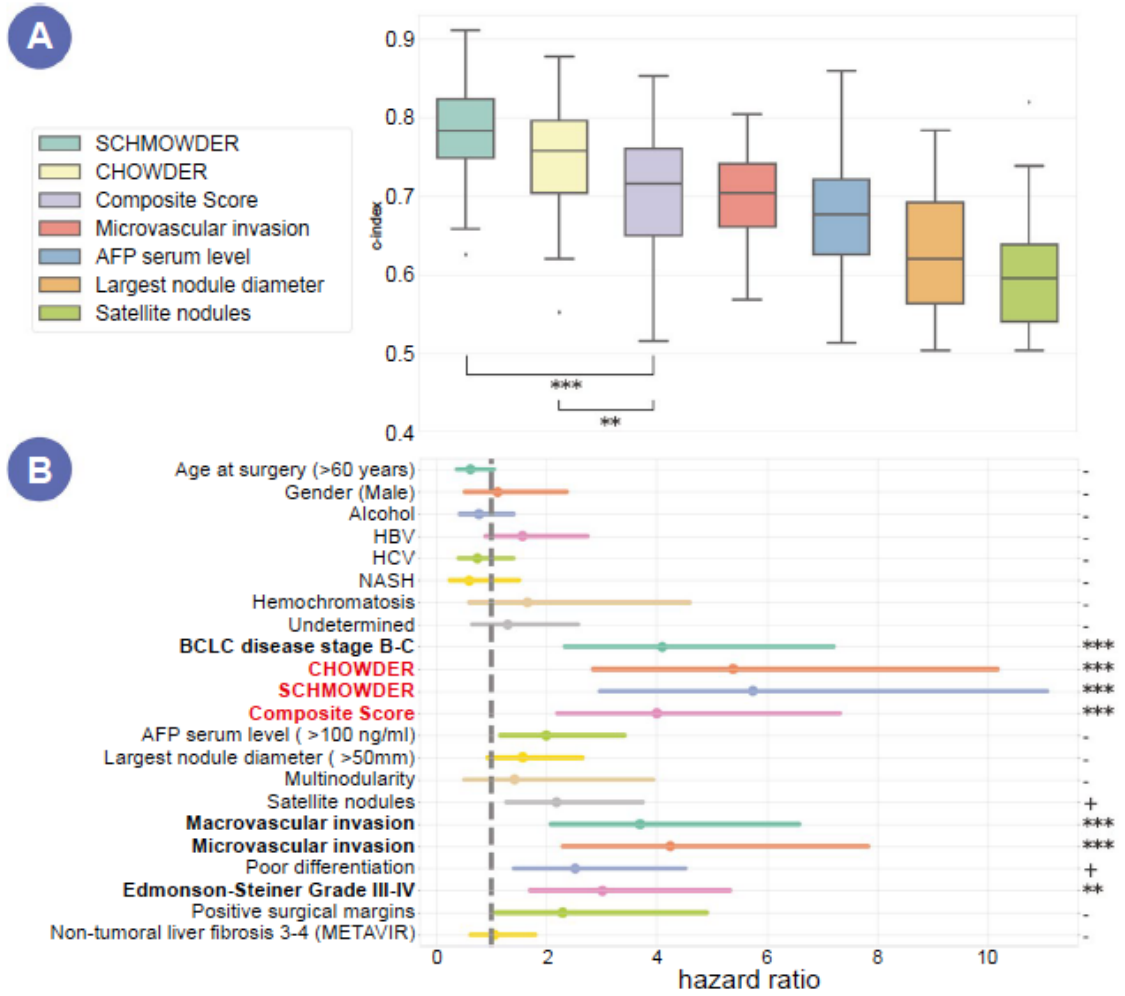
Predicting Survival After Hepatocellular Carcinoma Resection Using Deep Learning on Histological Slides

Charlie Saillard ,[1] Benoit Schmauch,[1] Oumeima Laifa,[1] Matahi Moarii,[1] Sylvain Toldo,[1] Mikhail Zaslavskiy,[1] Elodie Pronier, Alexis Laurent,[2,3] Giuliana Amaddeo,[3-5] Hélène Regnault,[5] Daniele Sommacale,[2-4] Marianne Ziol,[6,7] Jean-Michel Pawlotsky,[3,4,8] Sébastien Mulé ,[3,4,9] Alain Luciani,[3,4,9] Gilles Wainrib,[1] Thomas Clozel,[1] Pierre Courtiol,[1] and Julien Calderaro[3,4,10]

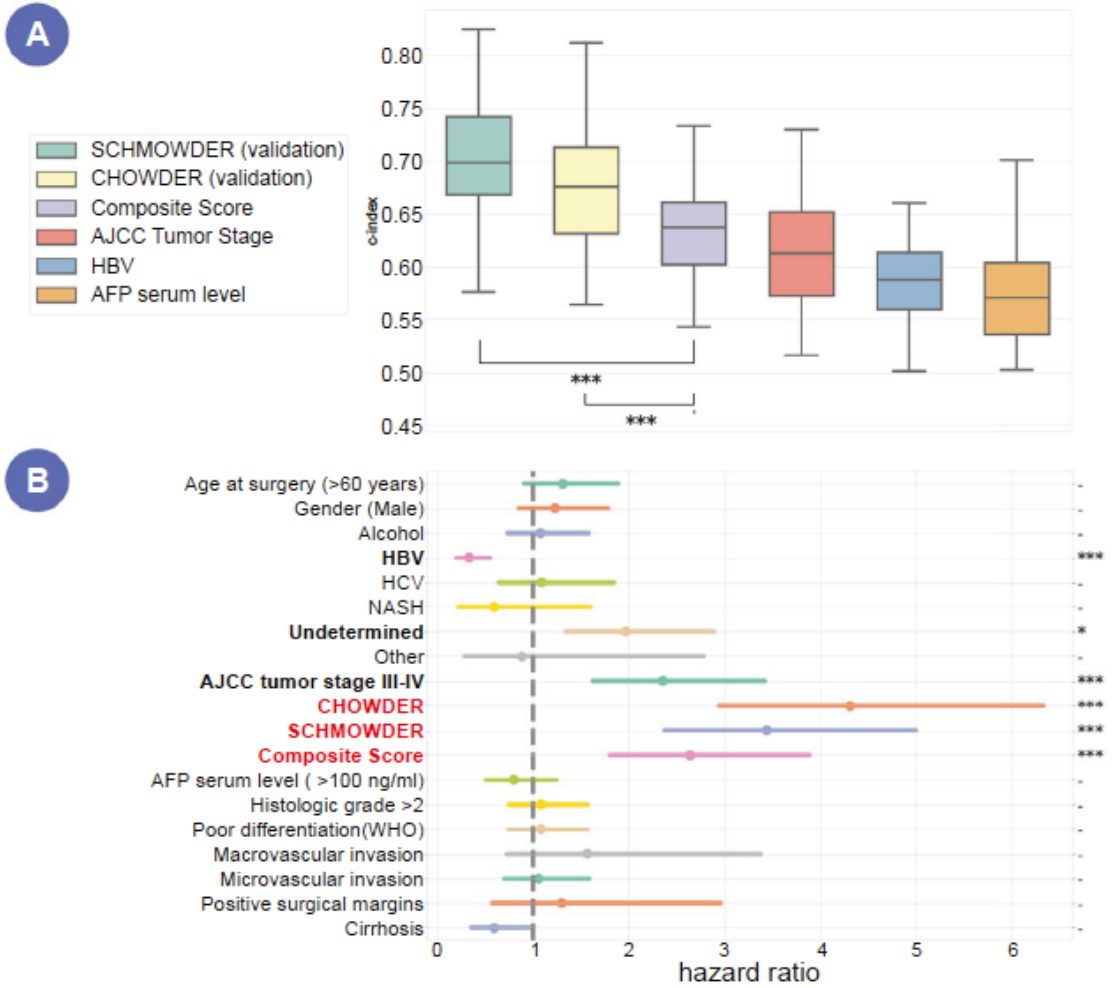
AIMS: Standardized and robust risk-stratification systems for patients with HCCare required to improve therapeutic strategies after curative resection/ablation.

METHODS: **two deep-learning algorithms based on whole-slide digitized histological slides (WSI) were used to build models for predicting survival of patients with HCC treated by resection**.. The first deep-learning–based algorithm (“SCHMOWDER”) uses an attention mechanism on tumoral areas annotated by a pathologist whereas the second (“CHOWDER”) does not require human expertise.





SCHMOWDER and CHOWDER models can predict survival after HCC resection more effectively than all other clinical, biological, or pathological variables.



SCHMOWDER and CHOWDER models can predict survival after HCC resection, outperforming all other clinical, biological, or pathological variables in the validation set (TCGA).

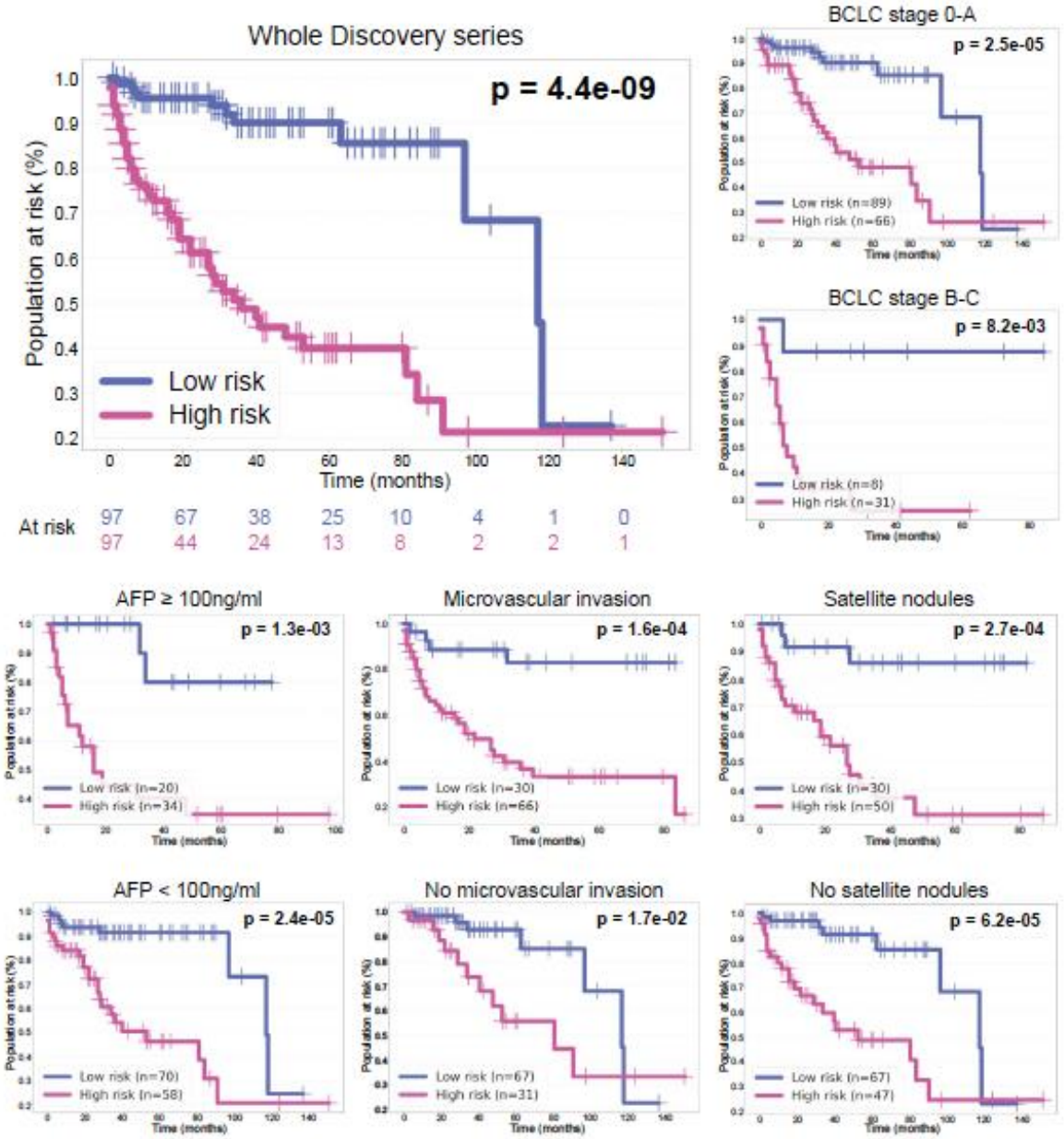


FIG. 4. Prognostic value of SCHMOWDER risk subclasses in the whole discovery series and after stratification for common baseline variables. The median SCHMOWDER risk score was used as a threshold to categorize patients into low- and high-risk subgroups. The prognostic value of SCHMOWDER was conserved, even after stratification, according to common clinical and pathological variables.

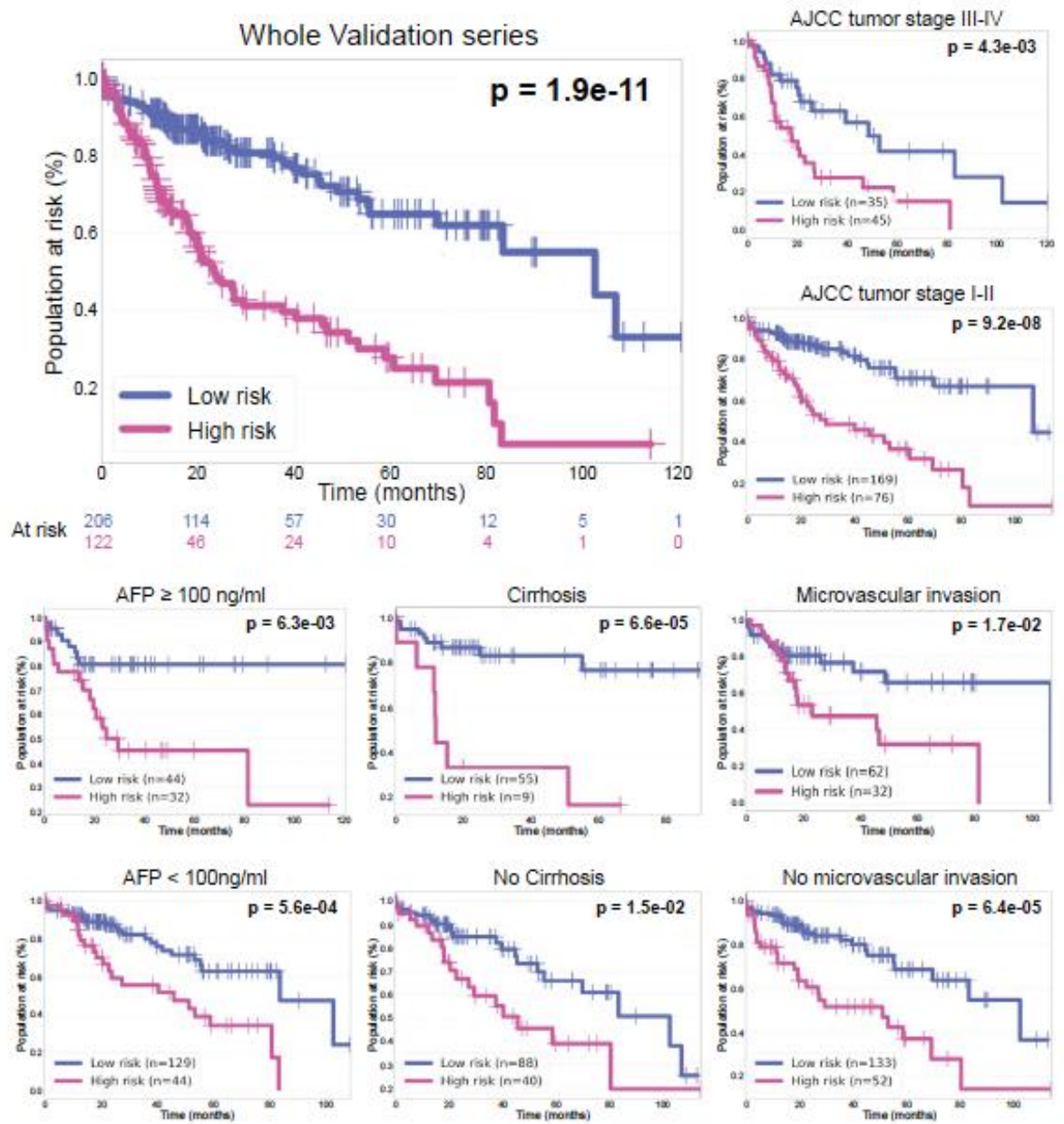


FIG. 6. Prognostic of SCHMOWDER risk subclasses for the whole validation set and after stratification for common baseline variables. The median SCHMOWDER risk score for the discovery set was used as a threshold to categorize patients into low- and high-risk subgroups. The risk scores obtained also predict survival after stratification for common baseline variables.

Research Article
Cancer

JOURNAL
OF HEPATOLOGY

Radiomic analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma

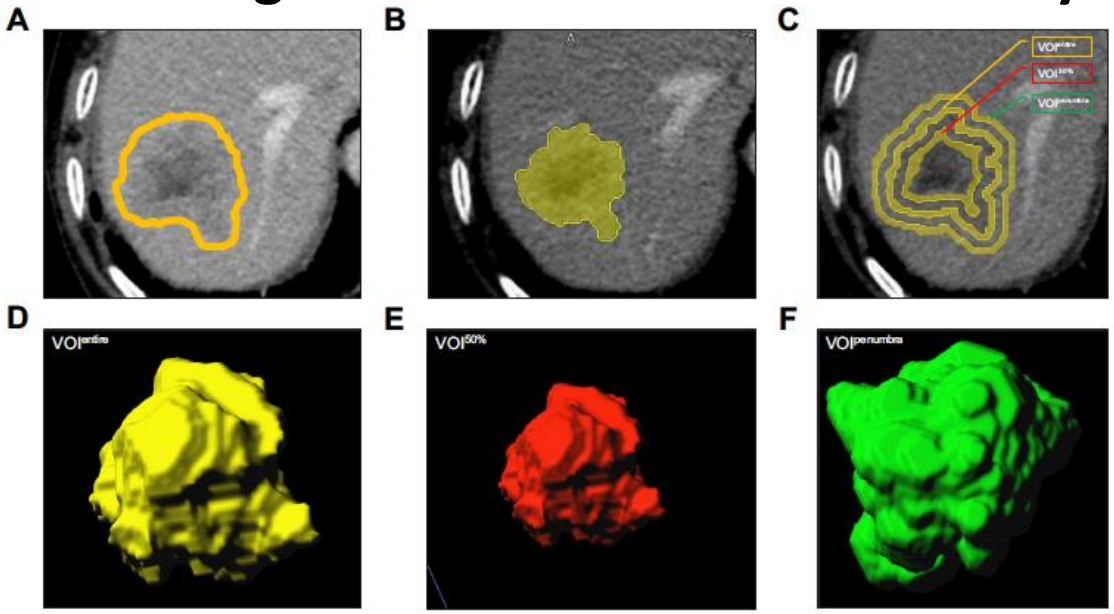
Xun Xu[1,4,5], Hai-Long Zhang[1,4,5], Qiu-Ping Liu[1], Shu-Wen Sun[1], Jing Zhang[2], Fei-Peng Zhu[1], Guang Yang[2], Xu Yan[3], Yu-Dong Zhang[1,*], Xi-Sheng Liu[1,*]

[1]*Department of Radiology, The First Affiliated Hospital with Nanjing Medical University, Nanjing, Jiangsu Province, China;* [2]*Shanghai Key Laboratory of Magnetic Resonance, East China Normal University, Shanghai, China;* [3]*MR Scientific Marketing, Siemens Healthcare, Shanghai, China*

Aims: As there is no single highly reliable factor to preoperatively predict microvascular invasion (MVI), they developed a computational approach integrating large-scale clinical and imaging modalities, especially radiomic features from contrast-enhanced CT, to predict MVI and clinical outcomes in patients with HCC.

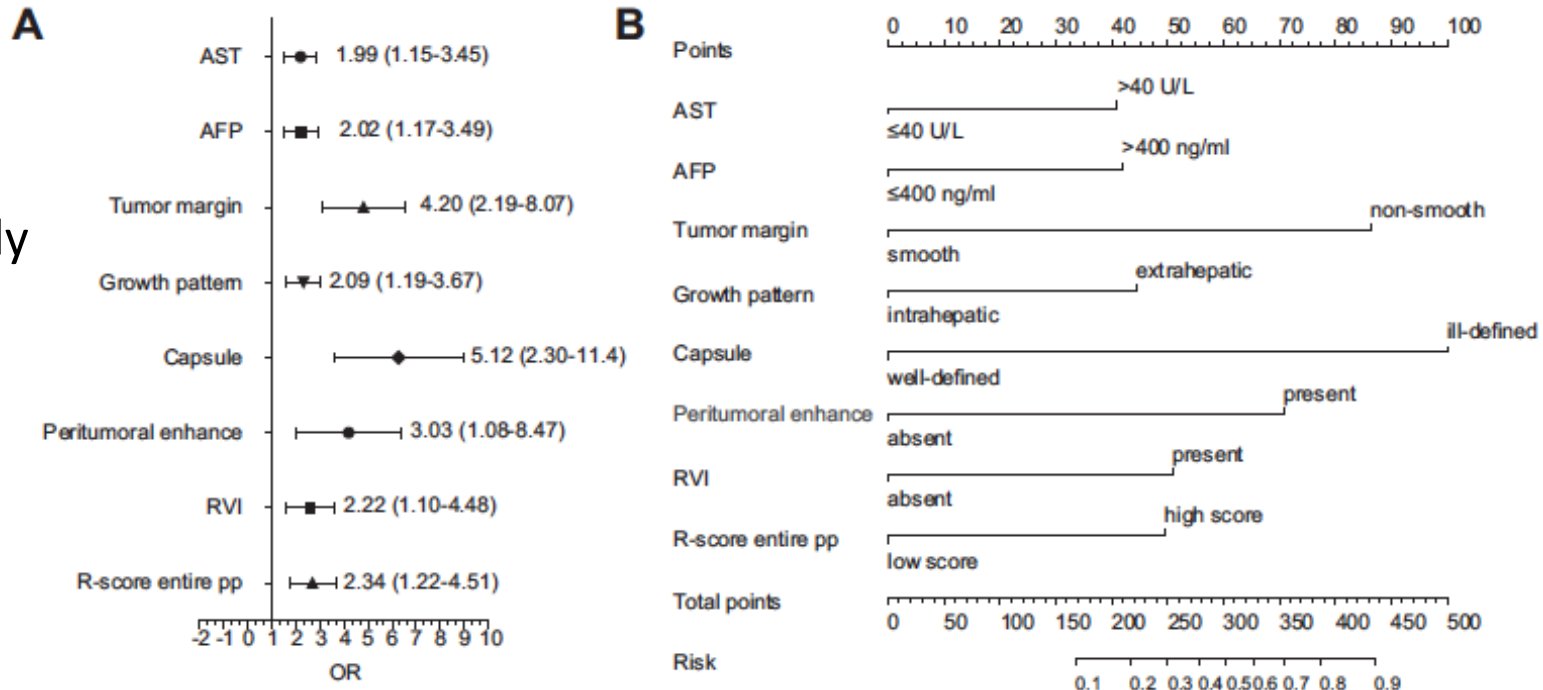
Methods: In total, 495 surgically resected patients were retrospectively included. MVI-related radiomic scores (R-scores) were built from 7,260 radiomic features in 6 target volumes. Six R-scores, 15 clinical factors, and 12 radiographic scores were integrated into a predictive model, the radiographic-radiomic (RR) model, with multivariate logistic regression.

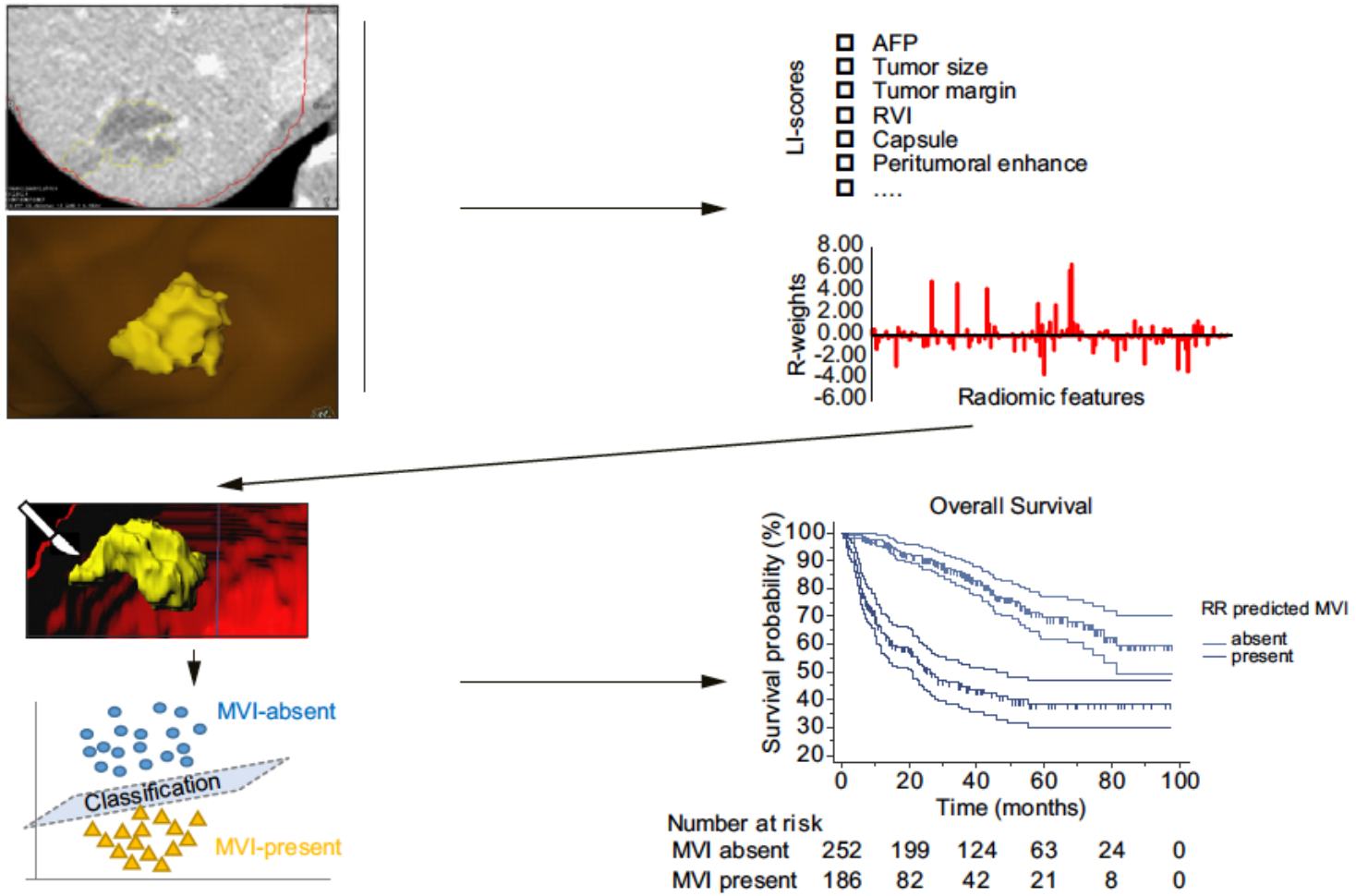
Lesion segmentation for radiomics analysis



Forest plot and nomogram of independent predictors of MVI.

In the multivariate regression model, 8 predictors were independent prognostic factors of histologic MVI. These independently associated risk factors were used to form the RR model.





Radiomic features were selected and quantitatively integrated into 6 R-scores.
The related R-scores showed significant differences according to MVI status ($p < 0.001$).

The RR model using the predictors achieved an area under the curve (AUC) of 0.909 in training/validation and 0.889 in the test set.

Progression-free survival (PFS) and overall survival (OS) were significantly different between the RR-predicted MVI-absent and MVI-present groups (median PFS: 49.5 vs. 12.9 months; median OS: 76.3 vs. 47.3 months).

POLICLINICO DI SANT'ORSOLA

SERVIZIO SANITARIO REGIONALE
EMILIA-ROMAGNA
Azienda Ospedaliero - Universitaria di Bologna
IRCCS Istituto di Ricovero e Cura a Carattere Scientifico

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Thank you for your kind attention

fabio.piscaglia@unibo.it
Internal Medicine, University of Bologna