MR. CHARLIE SAILLARD (Orcid ID : 0000-0003-3061-839X)
DR. SÉBASTIEN MULÉ (Orcid ID : 0000-0002-6896-6149)

Article type : Original

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[1], Alexis Laurent[2,3], Giuliana Amaddeo[3,4,5], Hélène Regnault[5], Daniele Sommacale[2,3,4], 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], Julien Calderaro[3,4,10].

1. Owkin Lab, Owkin.
2. Assistance Publique-Hôpitaux de Paris, Department of Hepatobiliary and Digestive Surgery, Hôpital Henri Mondor, Créteil, France.
3. Paris Est Créteil University, UPEC, Créteil, France.
4. INSERM U955, Team "Pathophysiology and Therapy of Chronic Viral Hepatitis and Related Cancers", Créteil, France.
5. Assistance Publique-Hôpitaux de Paris, Department of Hepatology, Hôpital Henri Mondor, Créteil, France.
6. Assistance Publique-Hôpitaux de Paris, Department of Pathology, Hôpital Jean Verdier, Bondy, France.
7. Paris 13 University, INSERM-1162, Functional Genomics of Solid Tumors, 75010, Paris, France.
8. National Reference Center for Viral Hepatitis B, C and Delta, Department of Virology, Hôpital Henri Mondor, Créteil, France.
9. Assistance Publique-Hôpitaux de Paris, Department of Medical Imaging, Hôpital Henri Mondor, Créteil, France.

10. Assistance Publique-Hôpitaux de Paris, Department of Pathology, Hôpital Henri Mondor, Créteil, France.

Running title: Deep learning and hepatocellular carcinoma

Corresponding authors:

Dr Julien Calderaro
Département de Pathologie
Hôpital Henri Mondor
51 avenue du Maréchal de Lattre de Tassigny
94010 Créteil, France
tel +33149812732
fax +33149812733
julien.calderaro@aphp.fr

Pierre Courtiol
Owkin Lab
16 rue Nansouty
75014 Paris, France
tel +33687529918
pierre.courtiol@owkin.com

Keywords : hepatocellular carcinoma, deep-learning, survival, artificial intelligence

Conflicts of interest:
CS, BS, OL, MM, ST, MZ, EP, GW, TC, PC are employed by Owkin, Inc.
JC receives consulting fees from Owkin, Inc.

Abstract

Standardized and robust risk stratification systems for patients with hepatocellular carcinoma (HCC) are required to improve therapeutic strategies and investigate the benefits of adjuvant systemic therapies after curative resection/ablation. In this study, we used two deep-learning algorithms based on whole-slide digitized histological slides (WSI) to build models for predicting the survival of patients with HCC treated by surgical resection. Two independent series were investigated: a discovery set (Henri Mondor Hospital, n=194) used to develop our algorithms and an independent validation set (TCGA, n=328). WSIs were first divided into small squares ("tiles") and features were extracted with a pretrained convolutional neural network (preprocessing step). The first deep-learning based algorithm ("SCHMOWDER") uses an attention mechanism on tumoral areas annotated by a pathologist while the second ("CHOWDER") does not require human expertise. In the discovery set, c-indexes for survival prediction of SCHMOWDER and CHOWDER reached 0.78 and 0.75, respectively. Both models outperformed a composite score incorporating all baseline variables associated with survival. The prognostic value of the models was further validated in the TCGA dataset, and, as observed in the discovery series, both models had a higher discriminatory power than a score combining all baseline variables associated with survival. Pathological review showed that the tumoral areas most predictive of poor survival were characterized by vascular spaces, the macrotrabecular architectural pattern and a lack of immune infiltration.

Conclusion

This study shows that artificial intelligence can help refine the prediction of HCC prognosis. It highlights the importance of pathologist/machine interactions for the construction of deep-learning algorithms that benefit from expert knowledge and allow a biological understanding of their output.

Introduction

Hepatocellular carcinoma (HCC) is the fourth cause of cancer-related death worldwide and its incidence is increasing in most Western countries.(1) It is characterized by an aggressive clinical course, with approximately two thirds of patients diagnosed at advanced stages.(1) Curative treatments, such as surgical resection and percutaneous ablation, are hampered by high recurrence rates (up to 70-80%).(1) Improvements in the identification of patients with poor clinical outcomes are required for the development of better therapeutic strategies and the investigation of the potential benefits of adjuvant systemic therapies.(1)

In most human cancers, prognosis is strongly related to pathological criteria.(2) Indeed, the histological analysis of tumor tissues provides crucial information for the stratification of patients and treatment allocation.(2) The recent development of slide digitization and computational/mathematical image processing promises to improve and standardize various pathological and morphological analyses, and may also facilitate the extraction of "hidden" imaging features potentially providing useful clinical and/or biological information.(3) In a recent pioneering study simulating routine pathology workflows, a subset of deep learning-based algorithms outperformed a panel of 11 pathologists at detecting lymph node metastases of breast cancers.(4) Coudray and coworkers have also shown that neural networks applied on whole-slide digitized histological images (WSIs) can diagnose the main histological subtypes of non-small cell lung cancer and predict the mutational status of various genes (e.g. *STK11*, *EGFR*).(5) There is growing evidence to suggest that the computational processing of WSIs will refine the prediction of patient prognosis, thereby improving treatment allocation.(6, 7) Yu and coworkers have demonstrated that digital image analysis software can extract significant prognostic features from hematein and eosin-stained slides in lung cancer.(6) Convolutional neural networks have also been reported to predict the aggressiveness of various human malignancies.(7, 8)

In this study, we investigated two independent cohorts of patients with HCC treated by surgical resection, to determine whether computational approaches of this type could help to refine the prediction of prognosis. We show that models based on convolutional neural networks predict survival more accurately than baseline clinical, biological and pathological features. By analyzing areas classified as "low-risk" and "high-risk" by our model, we were able to obtain insight into the most relevant features used by the network for patient stratification.

Material and Methods

Patients and samples

The discovery set consisted of patients treated by surgical resection at Henri Mondor University Hospital between 2004 and 2016. The independent validation set consisted of patients from The Cancer Genome Atlas (TCGA) public dataset. The inclusion criteria for the two cohorts were as follows: patients treated by surgical resection without prior anti-tumor treatment, available follow-up data, an unequivocal diagnosis of HCC (exclusion of cases with features suggestive of combined hepatocellular-cholangiocarcinoma), available histological slides of formalin-fixed paraffin-embedded material, and an absence of extrahepatic metastatic disease at the time of surgery **(Figure 1)**. The clinical endpoint was overall survival (time from surgery to death).

1. *Discovery set*

One or two representative digital slides of hematein-eosin-saffron (HES)-stained sections were available for each HCC case (NDPI format, 40 x magnification, total of 390 slides from 206 tumors). These slides included tissues from both the HCC and the adjacent surrounding liver tissues. The tumor and non-tumor areas were annotated by an expert liver pathologist (JC). Clinical (age, sex, risk factors for liver disease, disease stage), biological (pre-operative serum alpha-fetoprotein (AFP) serum level) and pathological (tumor size, number and differentiation, Edmondson-Steiner grade, vascular invasion, satellite nodules, fibrosis stage according to METAVIR) features were retrieved from medical records. Informed consent was obtained for each patient and the study was approved by a review board.

2. *Validation set*

We used The Cancer Genome Atlas database as an independent validation set for testing the robustness of our models. We applied the same inclusion criteria. Clinical, biological, pathological data and WSIs were downloaded from the website (https://www.cbioportal.org/study/summary?id=lihc_tcga).(9-11) In total, 342 WSIs from 328 patients were available for analysis (several slides were available for some patients). Slides from this series had been subjected to different staining and scanning protocols (SVS format) from those in our discovery set (the TCGA slides were collected from different centers).

Convolutional neural networks for predicting patient survival

We used two deep-learning algorithms, "CHOWDER" and "SCHMOWDER", which were specifically designed for the processing of WSIs. CHOWDER is a neural network developed

during a previous study that predicts overall survival from WSIs without the need for local annotation.(12) It can automatically identify very localized survival-related patterns on slides, and calculates a risk score for each WSI analyzed in three successive steps: a preprocessing step, a tile-scoring step, and a prediction step.

The WSI is first divided into small squares, 112 x 112 micrometers in size (224 pixels x 224 pixels), called "tiles", and features are extracted from these tiles with a pretrained convolutional neural network (preprocessing step) **(Figure 2A)**.

During model development, the tiles are then fed into the network architecture along with survival data, and a risk score is assigned to each tile through an iterative learning process. Finally, the network selects a small number of tiles with the highest and lowest survival scores for the prediction of patient survival (**Figure 2B**). The architecture of CHOWDER was designed to retrieve the most predictive tiles from the thousands processed for further analysis by pathologists.

SCHMOWDER is a two-branch neural network combining an unsupervised component and a supervised attention mechanism (**Figure 2C**). The preprocessing step is identical to that of CHOWDER. Annotations provided by the pathologist are then used to train the upper branch to identify tiles as tumoral or non-tumoral. By assigning a tumor score to each tile and applying an attention mechanism to these scores, the upper branch generates a representation of tiles with a high probability of being tumoral. The lower branch is weakly supervised and generates a representation of only a small number of tiles, the most predictive of survival. Representations from the two branches are merged to generate a survival risk as output.

This “expert-driven” approach, using annotations during training, thus combines two views of the WSI to improve the prediction of survival: a supervised approach focusing on tumoral areas, and a weakly supervised approach not dependent on prior knowledge. The lower branch focuses on very localized areas of the WSI, whereas the upper branch gathers information from broader areas (all tiles located in tumoral regions). We believe that this approach enables the model to capture survival-related information from both tumoral and non-tumoral tissues. Local annotations are needed during training, but not for inference; so, once trained, SCHMOWDER can be used on slides for which no annotations are available (such as those of the TCGA dataset).

Histological analysis of tiles of high predictive value

CHOWDER is an interpretable-by-design model, and this particular feature makes pathological assessments of the image tiles most significantly associated with patient outcome possible. This approach can thus provide insights into the features associated with tumor aggressiveness. Tiles with high and low risk scores were thus extracted and further analyzed by a pathologist expert in liver disease (J.C). He was blinded to the risk scores associated with each tile. In total, 27 histological features of tumoral and non-tumoral liver tissues were systematically recorded (**Supplemental Figure 1**).

To obtain a deeper understanding of morphological similarities of tiles reviewed, we then applied UMAP algorithm on these tiles. For both discovery and validations series, the set of low-risk and high-risk tiles reviewed was augmented with a random selection of 500 tiles that cover the whole distribution of predicted risks. UMAP was fitted on the tile features of the discovery set and then used to embed tiles of the discovery and validation sets.

K-Means clustering with 8 clusters was further fitted on the reduced data of the discovery set and applied on the validation set.

Statistical analysis

Survival analyses were performed with univariate and multivariate Cox proportional hazards models implemented in the lifelines package of Python.(13) Log-rank tests were used to compare survival distributions between stratification subgroups. We used Harrel's concordance index (c-index) as a metric for assessing the predictive performance of our model, and to compare the predictive performances of this model and baseline clinical, biological and pathological features.

The results were first validated on the discovery dataset with the following cross-validation strategy: 5 stratified folds, with 10 repeats. Folds were stratified based on censoring. The dataset from the TCGA database was kept entirely separate from the discovery series, and was used only for external validation by the same cross-validation strategy.

Model performances were compared in Student's *t*-tests. The qualitative variables included in the histological analysis of tiles of high predictive value were compared in *Z*-tests of proportions. The Holm-Sidak procedure was used to correct for multiple testing. All tests were two-tailed, and *p*-values < 0.05 were considered statistically significant.

Results

Model development from the discovery set

We included 194 patients in our discovery series **(Figure 1)**. The most frequent risk factors for liver disease were alcohol intake (30%) and HCV (hepatitis C virus) infection (27%) **(Supplemental Table 1)**. Disease stage, according to the Barcelona Clinic of Liver Cancer (BCLC) system, was 0/A in 79% of the patients and B/C in 21% of the patients. As expected liver function was preserved in almost all the patients (191/194 were classified as Child-Pugh A). We did not, therefore, include this variable in the analysis. The frequency of cirrhosis was also low **(Supplemental Table 1)**. Fifty-six deaths were recorded during follow-up (after a median follow up of 26.5 months). The clinical and pathological features associated with shorter overall survival were microvascular invasion (hazard ratio (HR)=4.2, *p*=5E-7), BCLC stage B-C (HR=4.1, *p*=6E-6) and macrovascular invasion (HR=3.7, *p*=4E-5).

One or two representative digital slides of hematein-eosin-saffron (HES)-stained sections were available for each HCC case (NDPI format, 40x magnification, total of 390 slides from 206 tumors; some patients had several tumors). These slides included tissues from both the HCC and the adjacent surrounding liver tissues, and the tumor and non-tumoral areas were identified and annotated by an expert liver pathologist (JC).

We extracted 20 000 tiles (small image patches of 224 x 224 pixels) from each available WSI (total of 7 800 000 tiles), and used a pretrained convolutional neural network to extract relevant features from each tile before training our models **(Figure 2A, Supplemental Methods)**.(14)

We assessed the discriminatory power of the CHOWDER and SCHMOWDER models for predicting overall survival by cross-validation. We found that both models performed well, with mean c-indexes of 0.75 and 0.78, respectively **(Figure 3A)**. Interestingly, both models outperformed baseline variables **(Figure 3A).** We then constructed a composite score (CS) integrating the relevant clinical, biological and pathological features for survival prediction, and showed that both deep learning-based models had a greater discriminatory power than the CS **(Figure 3A and Supplementary Methods)**.

Indeed, CHOWDER and SCHMOWDER outperformed the CS by 4 points (*p*=3.6E-3), and 7 points (*p*=5.9E-7), respectively. SCHMOWDER also significantly outperformed CHOWDER (*p*-value=2.1E-2). A mean aggregation of the predictions from the CS and our models did not improve performance significantly (c-indices stagnated at 0.75, p=0.99 and 0.78, p=0.83 with

CHOWDER and SCHMOWDER respectively), suggesting that the available clinical, biological and pathological variables provided no additional prognostic information.

The output of our neural networks is a continuous risk score. We used the risk score assigned to each patient by CHOWDER and SCHMOWDER to stratify the population into two subgroups. The median risk score of each model was used as a threshold for stratifying patients into low and high-risk subgroups. Our models stratified the population more accurately than the CS or any other clinical or pathological variable (HR=4.0, *p*=1.1e-6 and HR=5.38, *p*=5.2e-9 and HR=5.72, *p*=4.4e-9 for the CS, CHOWDER and SCHMOWDER, respectively) **(Figure 3B and 4)**. The risk score computed by SCHMOWDER was of independent prognostic value (*p*=3.10-5), and predicted survival even after stratification for other features (such as BCLC disease stage, satellite nodules, AFP serum level, vascular invasion) **(Figure 4)**. These observations demonstrate that the model captures complex patterns non-redundant with baseline variables known to affect survival in patients with HCC.

Validation of our models with the TCGA dataset

We assessed the robustness of our models by testing them on an independent series from The Cancer Genome Atlas (TCGA). Clinical, biological, and pathological data were downloaded from the database, together with digital slides (9-11). In total, 328 patients met the inclusion criteria used for the discovery set **(Supplemental Table 2 and Methods),** and 115 deaths were recorded among these patients. Overall survival was significantly lower than that for the discovery set (median of 58 months vs. 91 months, *p*=0.02). The slides had been treated by different staining protocols and were encoded in a different format from the discovery set (they were also collected from different clinical centers). The following variables were available and included in the analysis: disease stage according to the American Joint Committee on Cancer (AJCC), age at diagnosis, sex, serum alpha fetoprotein (AFP), alcohol consumption, HBV (hepatitis B virus) or HCV infection, other etiologies, undetermined etiology, tumor differentiation according to the World Health Organization criteria, macrovascular and microvascular invasion, positive surgical margins, and non-tumoral liver fibrosis (cirrhosis). The clinical, biological and pathological features associated with shorter survival were AJCC stage (HR=2.35, *p*=1.6E-5), and undetermined etiology (HR=1.97, *p*=3.8E-3) **(Figure 5)**. As for the discovery set, a CS integrating relevant baseline variables was calculated **(Figure 5)**.

Tiles from 342 WSIs corresponding to the 328 patients were extracted and processed by our two models. Both models were validated, with mean c-indexes for the prediction of survival of 0.68 for

CHOWDER and 0.70 for SCHMOWDER **(Figure 5A)**. As for the discovery set, both models outperformed all the other available features and the CS integrating all relevant baseline characteristics.The CS indeed gave a mean c-index of 0.63, which represents a 26% increase with respect to a random prediction, while CHOWDER and SCHMOWDER featured an increase of 36% (p=3.6e-5) and 40% (p=3.4e-9), respectively. **(Figure 5A)**. SCHMOWDER significantly outperformed CHOWDER (*p*=4.5E-2) and was an independent prognostic factor (*p*=3.2E-7). We used the threshold established with the discovery set to classify patients into high and low-risk subclasses. This classification stratified the population of the validation set more effectively than the CS or any other variable (HR=2.6, *p*=3.2E-7 and HR=4.3, *p*=2.5E-12 and HR=3.4, *p*=5.3E-11 for the CS, CHOWDER and SCHMOWDER, respectively) **(Figure 5B)**.

SCHMOWDER also predicted survival even after stratification for other common prognostic features (microvascular invasion, cirrhosis, serum AFP concentration and AJCC tumor stage). Survival curves for the risk subgroups are provided in **Figure 6.**

Analysis of tiles

The architecture of CHOWDER enables it to retrieve the most predictive tiles from among the thousands processed. We investigated the main histological determinants of survival, by extracting the 400 most predictive tiles (high risk of death: 200, low risk of death: 200) from 245 WSIs with CHOWDER and having them reviewed by a pathologist with expertise in liver disease and tumors (JC). Nineteen histological features were recorded in tumoral areas, and the features most predictive of a high risk of death were the presence of vascular spaces (p=1.3E-10), a macrotrabecular architectural pattern (*p*=8.6E-5), a high degree of cytological atypia (*p*=1.4E-8) and nuclear hyperchomasia (*p*=1.4E-8). In both tumoral and non-tumoral areas, fibrosis and immune cells were associated with a low risk score (*p*=1.1E-3 and 3.4E-8, respectively) **(Figure 7)**. Overall, these results demonstrate that our deep learning model can detect both known and novel histological patterns associated with survival in HCC patients.

We further performed UMAP on the subsets of low and high risk tiles reviewed, augmented with a random selection of tiles that cover the whole distribution of predicted risks. Panel B of **Figure 7** shows the distribution of tiles from both datasets, using a 2D embedding of the 256-dimensional autoencoded features. Interestingly, the generated graphs show that tiles tend to cluster according to their risk class. To further analyze the morphological variations caught by the model, we also performed a clustering of the data into 8 clusters using K-means algorithm. Among those clusters, 3 were clearly associated with an increased or reduced risk (clusters 0, 1 and 3 on the

figure), one contained essentially artefacts such as folds or ink (cluster 2). The 4 remaining clusters, in the bulk of the distribution, were not specifically associated with high or low-risk or a well-defined pattern, so we decided to merge them together (cluster 4). The same clustering was subsequently applied to the TCGA data.

In the discovery set, most high-risk tiles are grouped in cluster 1, that is essentially defined visually by the presence of vascular spaces **(Supplemental table 3)**. Other common features of this cluster are blood, atypia and macrotrabecular pattern. A majority of low-risk tiles are located in cluster 0, that is enriched in fibrosis and immune cells. Cluster 3 is associated with a low risk, and contains essentially tiles with steatosis. Finally, both high and low-risk tiles are found in Cluster 4, including a majority of tiles with fibrosis or immune cells **(Supplemental table 3)**.

The same clustering gives overall consistent results on the validation set, with a few differences **(Supplemental table 3)**. Cluster 3, characterized by steatosis, is almost absent and cluster 4 is associated with an increased risk. The patterns found in cluster 4 are also different, with a predominance of atypia and hyperchromasia.

Discussion

In this study, we developed two deep learning models from histological slides for predicting survival after the surgical resection of HCC. We show that our algorithms predict survival more accurately than classical clinical, biological and pathological features. Both models had a greater discriminatory performance than the combination of all common variables associated with HCC prognosis.

Artificial intelligence-driven approaches for processing medical images provide us with a unique opportunity to standardize the diagnosis of cancer subtypes and improve patient stratification.(3) In a recent pioneering study, Coudray and coworkers used a large collection of lung cancer slides and showed that a deep learning-based model could classify them into non-tumoral lung tissue, adenocarcinoma and squamous carcinoma with an accuracy similar to that of pathologists.(5) Moreover, six of the most frequent genetic alterations occurring in these malignancies could be predicted directly from the slides.(5)

Here, we provide additional evidence suggesting that these approaches will improve precision medicine. Indeed, our models outperformed all other common clinical or pathological features for predicting survival. Slides from the TCGA dataset were indeed processed as they were, without the annotation of tumoral or non-tumoral areas or the removal of artifacts (e.g. blurring, annotation markers or folds). Slide formats and staining protocols also differed between the two series of slides (NDPI format and hematein-eosin-saffron staining for the discovery set and SVS format and hematein-eosin staining for the validation set). These findings demonstrate the robustness of our models and suggest that they may be suitable for use in different technical settings. However, both models performed less well on the validation set, for which the prognostic power of classical clinical and pathological variables was also lower (difference in c-indexes and hazard ratio values between our dataset and the TCGA). This may be explained by the lower overall survival of the patients from the TCGA dataset. We cannot however exclude an overfitting issue during the development of the model in the training set. Indeed, features specific of the discovery set (such as noise related to the staining techniques or slides format) are likely to be also incorporated by the neural networks in the decision process and may thus impact the performance on external and new datasets. This is particularly true for image analysis, and the limitation of this phenomenon is an active area of research. Data augmentation approaches, which artificially inflate the training set size by data transformation and/or oversampling, may be considered, however they significantly increase the computation time.

The major difference between the two neural networks used in this study was the inclusion of an attention mechanism in SCHMOWDER. Whole-slide images contain large amounts of information, not all of which is relevant to the prediction of survival. Our model acts in a similar way to pathologists analyzing images to make a diagnosis. It focuses on specific areas of the slide identified as potentially containing key survival-related information. Indeed we hypothesized that the most important features for survival prediction were located in the tumoral areas, and we showed that the addition of this attention mechanism significantly improved the performance of the model. Once trained, these networks no longer require annotation, as they can automatically identify the regions of interest on slides. These results highlight the importance of human-machine interactions for the development of artificial intelligence-based algorithms.

Several studies based on gene expression profiling and/or genetic testing have reported scores predicting the survival of patients with HCC.(15) In a recent and elegant study, Chaudhary and collaborators used deep-learning approaches on RNA/miRNA sequencing and methylation data from the TCGA database to predict prognosis after HCC resection. Their model was able to identify two distinct subgroups of patients with significant survival differences, and was further validated in five external cohorts. The implementation of these high-throughput gene expression profiling/sequencing technologies in clinical care is however currently hampered by their cost, the need for nucleic acid extraction, and standardization and reproducibility issues. Our model requires only stained slides, which are easily available in a surgical treatment setting. We believe that this approach would thus facilitate the implementation of risk stratification systems in clinical practice. The use of such models will however require particular workflows in pathology departments that allow efficient, rapid and reproducible glass slides scanning. The WSI processing and computing time should also be short enough not to delay therapeutic decisions. Both our discovery and validation series consisted of patients treated by surgical resection for whom abundant histological material is available, and the validation of our models in other clinical settings remains a challenge for future studies.

Convolutional neural networks are generally seen as “black boxes”.(16) Indeed, data are processed through complex layers of computation, and it is difficult to identify the most relevant features used by trained models for sample classification.(16) This lack of interpretability is considered a major limitation of deep learning-based networks. We used an innovative approach allowing extraction of the most pertinent image tiles and their subsequent analysis by trained pathologists. We show that the classification obtained is at least partly based on pathological features known to be associated with the clinical outcome of HCC, such as the macrotrabecular architectural pattern, the presence of intratumoral immune infiltration, or high levels of cellular

atypia.(17, 18) We also identified a new feature, the presence of vascular spaces, as strongly linked to a high risk of poor overall survival. This observation suggests that the vascular architecture of HCC is a key determinant of the aggressiveness of this cancer, warranting further biological research. It is also consistent with previous histological and biological studies showing an association between pro-angiogenic phenotype and poor clinical outcome.(19-22) An analysis of tiles extracted from the non-tumoral liver yielded unexpected results, with an enrichment in fibrosis and inflammatory infiltrates, features commonly related to severe liver disease, in tiles classified as low risk.(23) However, this model may have identified a particular type of fibrosis that can regress and immune cell subsets preventing liver damage. We thus believe that disrupting new scientific hypotheses can be generated with models of this type, facilitating analyses of the most important features affecting tile classification.

In conclusion, using convolutional neural networks on HCC digital slides, we have developed a model predicting survival with a greater discriminatory power than an optimal combination of all classical and relevant clinical, biological and pathological features. Approaches of this type may lead to improvements in the assessment of patient prognosis, and will also add to our scientific knowledge, through analyses of the most relevant tumor areas.

Acknowledgments

We warmly thank all the physicians (surgeons, hepatologists, oncologists, radiologists and pathologists) involved in the clinical care of the patients.

References

1. Llovet JM, Zucman-Rossi J, Pikarsky E, Sangro B, Schwartz M, Sherman M, Gores G. Hepatocellular carcinoma. Nat Rev Dis Primers 2016;2:16018.
2. Goldblum JR, Lamps LW, McKenney JK, Myers JL, Ackerman LV, Rosai J. Rosai and Ackerman's surgical pathology. Eleventh edition. ed. Philadelphia, PA: Elsevier, 2018: 2 volumes (xiv, 2142 pages).
3. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol 2019;20:e253-e261.
4. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak J, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 2017;318:2199-2210.
5. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyo D, Moreira AL, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559-1567.
6. Yu KH, Zhang C, Berry GJ, Altman RB, Re C, Rubin DL, Snyder M. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 2016;7:12474.
7. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velazquez Vega JE, Brat DJ, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A 2018;115:E2970-E2979.
8. Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis CA, Gaiser T, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Med 2019;16:e1002730.
9. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401-404.
10. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6:pl1.
11. Cancer Genome Atlas Research Network. Electronic address wbe, Cancer Genome Atlas Research N. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell 2017;169:1327-1341 e1323.

12. Pierre Courtiol EWT, Marc Sanselme, Gilles Wainrib. Classification and disease localization in histopathology using only global labels : a weakly supervised approach. In: arXiv; 2018.
13. Davidson-Pilon C K, J, Zivich P, Kuhn B, Fiore-Gartland A, Moneda L, Gabriel, WIlson D, Parij A, Stark K, Anton S, Besson L, Jona, Gadgil H, Golland D, Hussey S, Kumar R, Noorbakhsh J, Klintberg A, Martin E, Ochoa E, Albrecht D, dhuynh, Medvinsky D, Zgonjanin D, Chen D, Ahern C, Fournier C, Arturo, Rendeiro AF. CamDavidsonPilon/lifelines: v0.22.0.
14. He K ZX, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv 2015.
15. Zucman-Rossi J, Villanueva A, Nault JC, Llovet JM. Genetic Landscape and Biomarkers of Hepatocellular Carcinoma. Gastroenterology 2015;149:1226-1239 e1224.
16. Price WN. Big data and black-box medical algorithms. Sci Transl Med 2018;10.
17. Calderaro J, Couchy G, Imbeaud S, Amaddeo G, Letouze E, Blanc JF, Laurent C, et al. Histological Subtypes of Hepatocellular Carcinoma Are Related To Gene Mutations and Molecular Tumour Classification. J Hepatol 2017.
18. Ziol M, Pote N, Amaddeo G, Laurent A, Nault JC, Oberti F, Costentin C, et al. Macrotrabecular-massive hepatocellular carcinoma: A distinctive histological subtype with clinical relevance. Hepatology 2018;68:103-112.
19. Fang JH, Zhou HC, Zhang C, Shang LR, Zhang L, Xu J, Zheng L, et al. A novel vascular pattern promotes metastasis of hepatocellular carcinoma in an epithelial-mesenchymal transition-independent manner. Hepatology 2015;62:452-465.
20. Renne SL, Woo HY, Allegra S, Rudini N, Yano H, Donadon M, Vigano L, et al. Vessels Encapsulating Tumor Clusters (VETC) Is a Powerful Predictor of Aggressive Hepatocellular Carcinoma. Hepatology 2019.
21. Calderaro J, Ziol M, Paradis V, Zucman-Rossi J. Molecular and histological correlations in liver cancer. J Hepatol 2019.
22. Villa E, Critelli R, Lei B, Marzocchi G, Camma C, Giannelli G, Pontisso P, et al. Neoangiogenesis-related genes are hallmarks of fast-growing hepatocellular carcinomas and worst survival. Results from a prospective study. Gut 2015.
23. MacSween RNM, Burt AD, Portmann B, Ferrell LD. MacSween's pathology of the liver. In. 6th ed. Edinburgh: Churchill Livingstone,; 2011. p. 1 online resource (1 v.).

Figure legends

Figure 1. Flow-chart and global methodology of the study. The models were first developed and cross-validated in a series of patients with HCC treated by surgical resection at Henri Mondor University Hospital. They were then validated with patients from The Cancer Genome Atlas.

Figure 2. Schematic representation of the CHOWDER and SCHMOWDER models. A) Whole-slide images were first cut into small image patches of tissue ("tiles"), and features were extracted from these tiles by a pretrained convolutional neural network. This preprocessing step was identical for CHOWDER and SCHMOWDER. B) CHOWDER is a neural network that predicts a risk score from WSIs, without the need for local annotations. Tile features are fed into the network along with survival data, and a risk score is assigned to each tile through an iterative learning process. The network then selects the 25 tiles with the highest and lowest scores for the prediction of patient survival. One of the key features of CHOWDER is its interpretability, as the most predictive tiles can be extracted and reviewed. C) SCHMOWDER is a two-branch neural network combining an unsupervised branch (lower part) and a supervised attention mechanism (upper part). Using annotations provided by the pathologist (JC), the upper branch is trained to determine whether each tile is tumoral or non-tumoral. By assigning a tumoral score to each tile and applying an attention mechanism to these scores, the upper branch generates a representation of tiles with a high probability of being tumoral. The lower branch, like CHOWDER, is weakly supervised and generates a representation of a small number of tiles that are the most predictive of survival. Representations from the two branches are merged and a survival risk is calculated.

Figure 3. The SCHOWDER and CHOWDER models can predict survival after hepatocellular carcinoma resection more effectively than all other clinical, biological or pathological variables. A) C-indices for the prediction of survival by the deep learning-based SCHOWDER and CHOWDER models, the composite score and the most relevant clinical, biological and pathological features. The two models significantly outperformed the composite score and all baseline variables. B) Hazard ratios and 95% confidence intervals for the prediction of survival by CHOWDER and SCHMOWDER, the composite score and baseline features. The SCHMOWDER model score was converted into a binary score (high or low risk), using the median as a threshold. HR: hazard ratio; CI: confidence interval. HBV: hepatitis B virus, HCV: hepatitis C virus, NASH:

nonalcoholic steatohepatitis, BCLB: Barcelona Clinic liver cancer, AFP: alpha fetoprotein. ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, +: $p < 0.1$, -: $p > 0.1$.

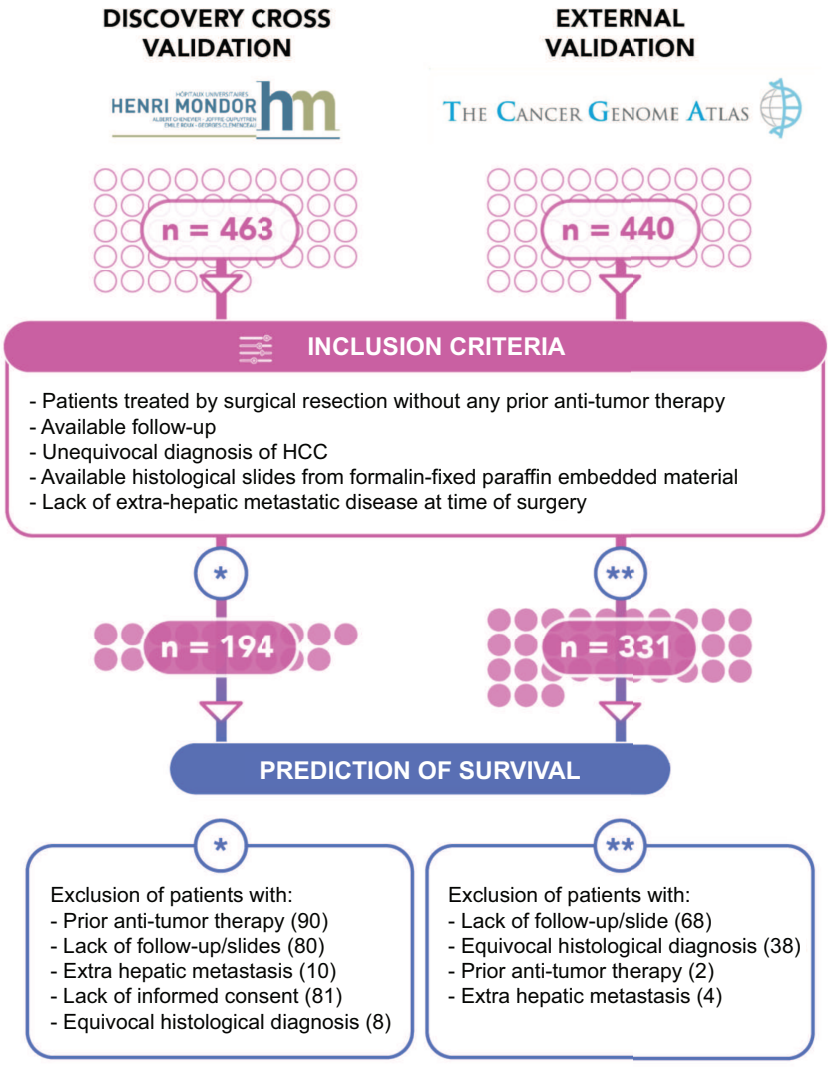
Figure 4. Prognostic value of SCHMOWDER risk subclasses in the whole discovery series and after stratification for common baseline variables. The median SCHMOWDER risk score was used as a threshold to categorize patients into low-risk and high-risk subgroups. The prognostic value of SCHMOWDER was conserved,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

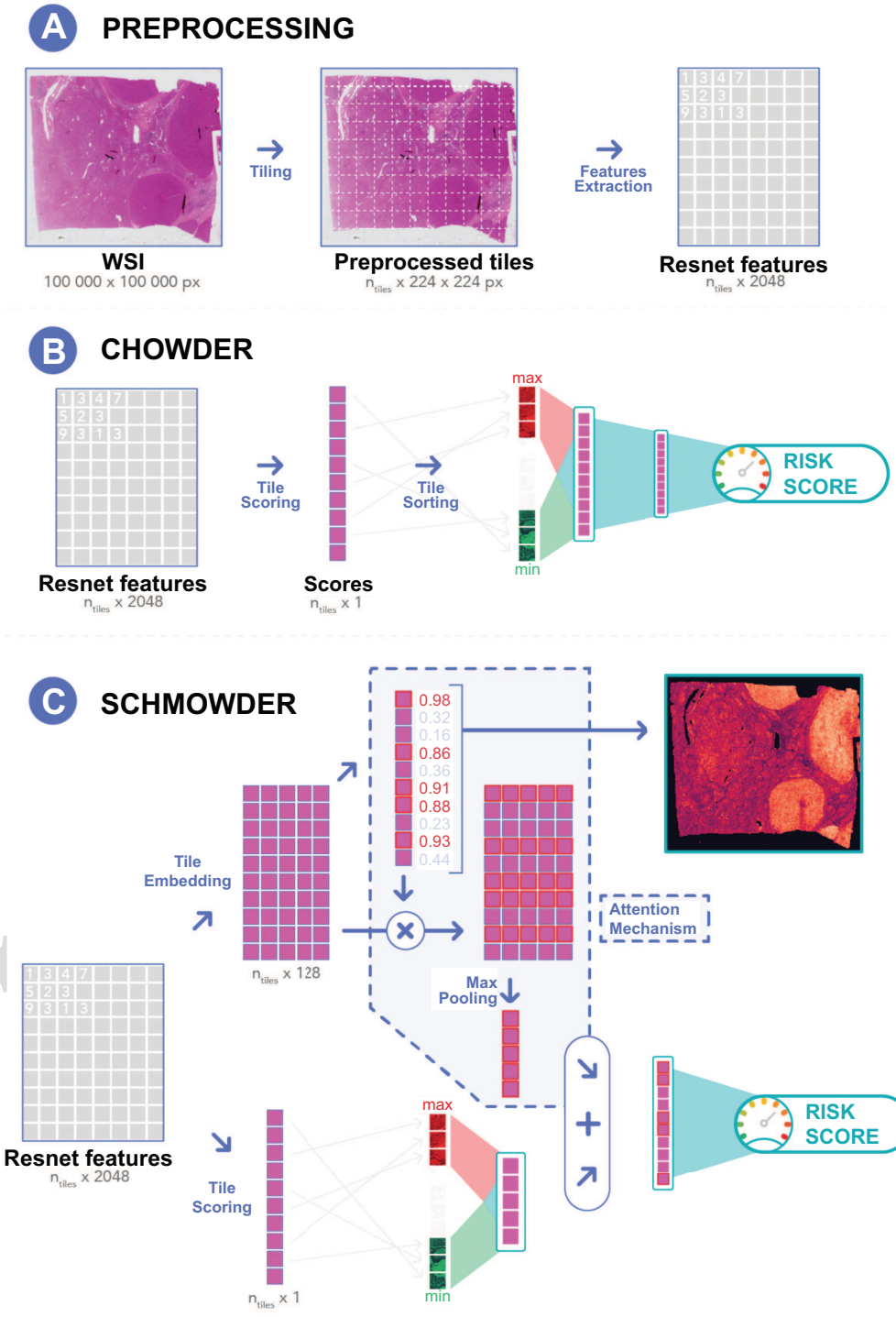
Figure 5. The SCHMOWDER and CHOWDER models can predict survival after hepatocellular carcinoma resection, outperforming all other clinical, biological or pathological variables in the validation set (The Cancer Genome Atlas: TCGA) A) C-indices for the prediction of survival of the SCHMOWDER and CHOWDER models trained on the discovery set and tested on the validation set, the composite score and the most relevant clinical, biological and pathological features of the TCGA cohort. Both models significantly outperformed the composite score and all baseline variables. B) Hazard ratios and 95% confidence intervals for the prediction of survival by the CHOWDER and SCHMOWDER models, the composite score and baseline features. The SCHMOWDER model score was converted into a binary score (high or low risk), using the median value for the discovery set as a threshold. HR: hazard ratio; CI: confidence interval. HBV: hepatitis B virus, HCV: hepatitis C virus, NASH: nonalcoholic steatohepatitis, BCLB: Barcelona Clinic liver cancer, AFP: alpha fetoprotein. ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, +: $p < 0.1$, -: $p > 0.1$.

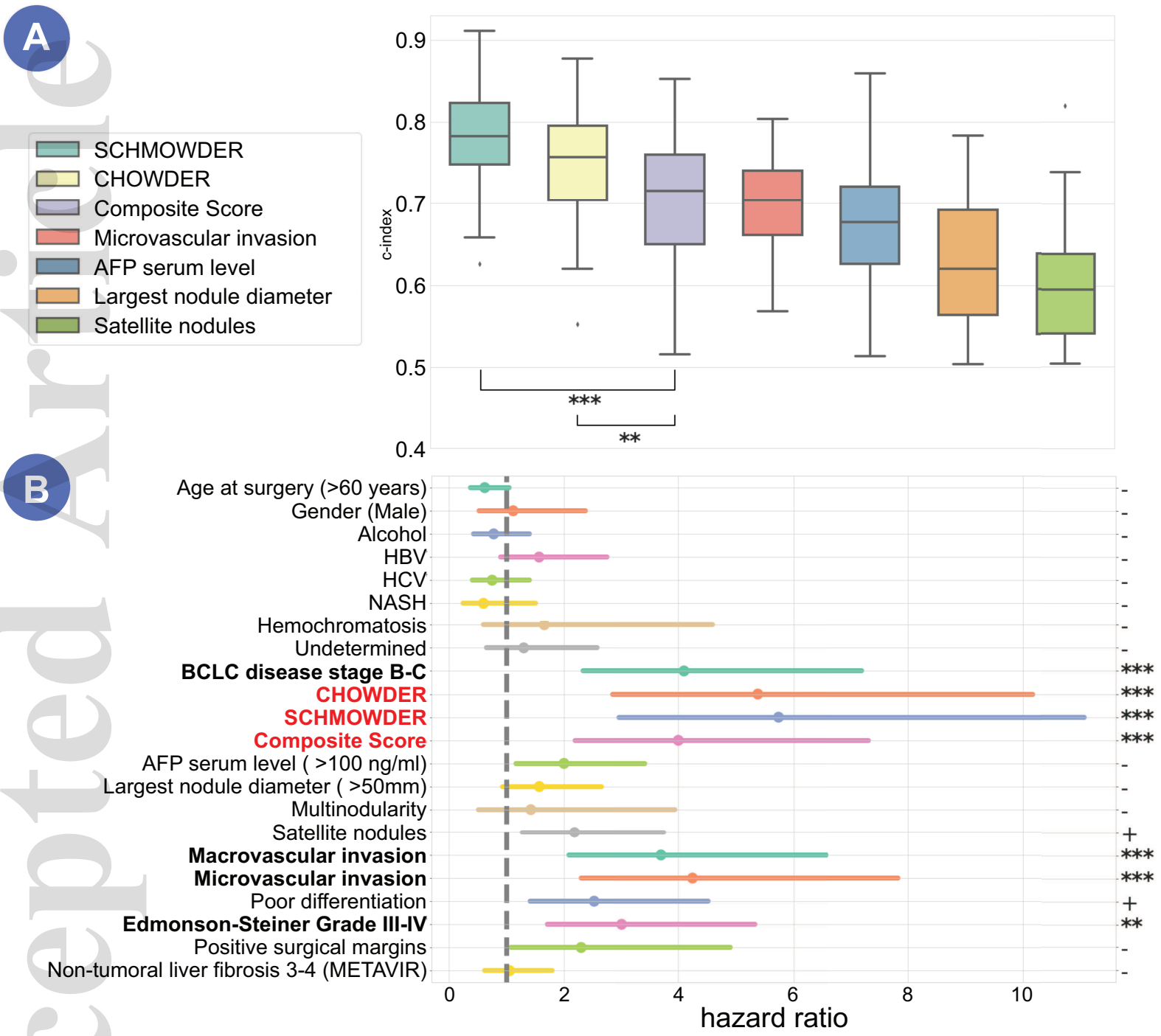
Figure 6. Prognostic of SCHMOWDER risk subclasses for the whole validation set and after stratification for common baseline variables. The median SCHMOWDER risk score for the discovery set was used as a threshold to categorize patients into low-risk and high-risk subgroups. The risk scores obtained also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

Figure 7. Examples of tiles classified as low or high risk by CHOWDER. An expert pathologist (JC) blinded to risk scores analyzed the 400 most predictive tiles. A) The features

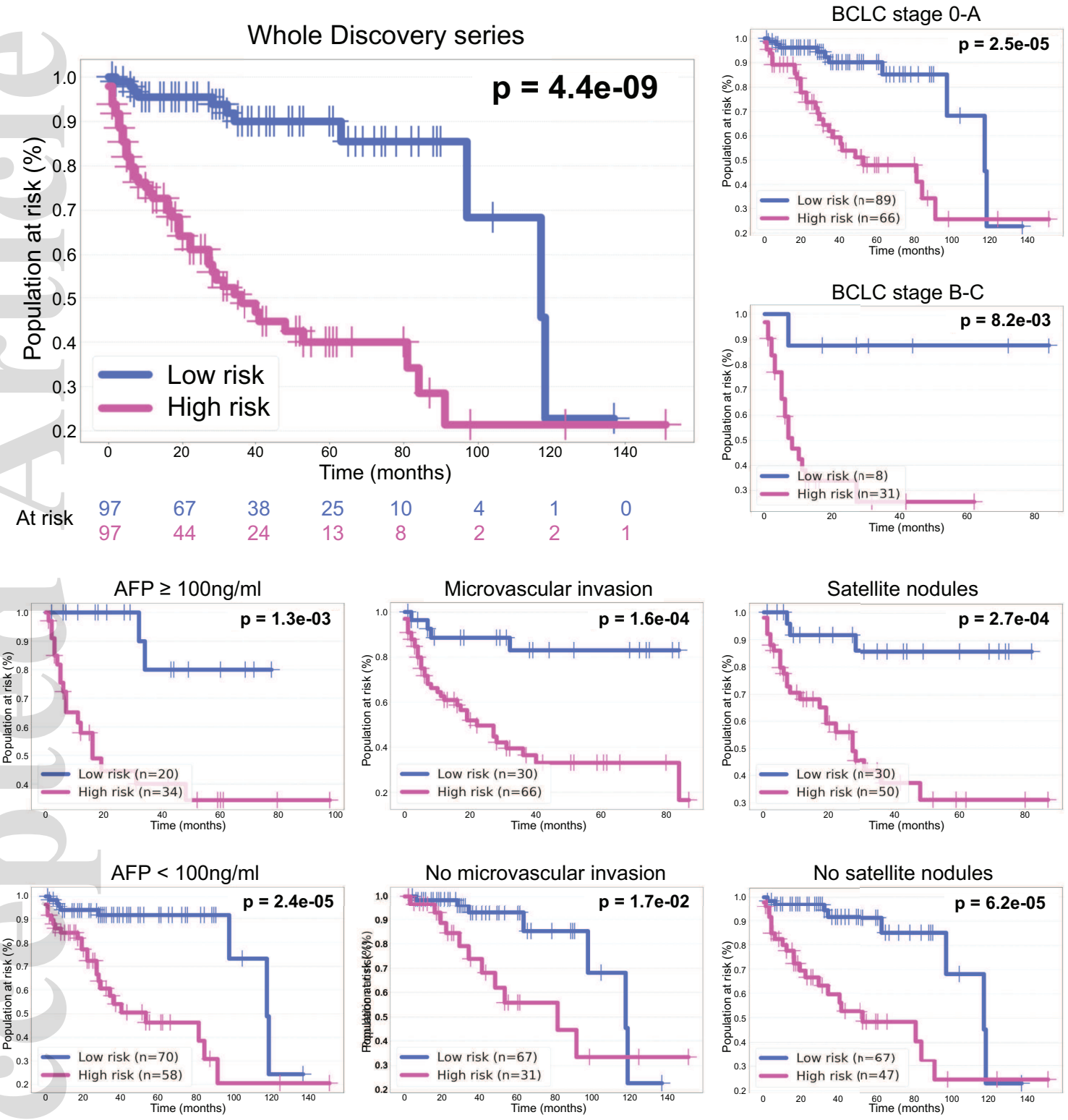
predictive of a high risk of death included cellular atypia, vascular spaces, and macrotrabecular architectural pattern (all identified on tiles extracted from tumoral areas). The features predictive of a low risk of death included tumoral fibrotic stroma, the presence of immune cells and fibrosis (in both tumor and non-tumor areas). In each case, the tissue area containing the tile is also shown. B) Features of low and high risk tiles reviewed by an expert pathologist, as well as an additional set of tiles selected randomly, were analyzed using UMAP. K-means clustering was further applied on the reduced data. Tiles tend to cluster according to their risk class, and clustering of Mondor data show that most high-risk tiles are grouped in cluster 1, while the majority of low-risk tiles are located in cluster 0 (enriched in fibrosis and immune cells). Cluster 3 is also associated with a lower risk, and essentially includes tiles with steatosis. Cluster 4 contains both high and low-risk tiles, including a majority of tiles with fibrosis or immune cells. When applied to TCGA data, the same clustering gives overall consistent results, with a few differences. Cluster 3, characterized by steatosis, is almost absent and cluster 4 is associated with an increased risk.



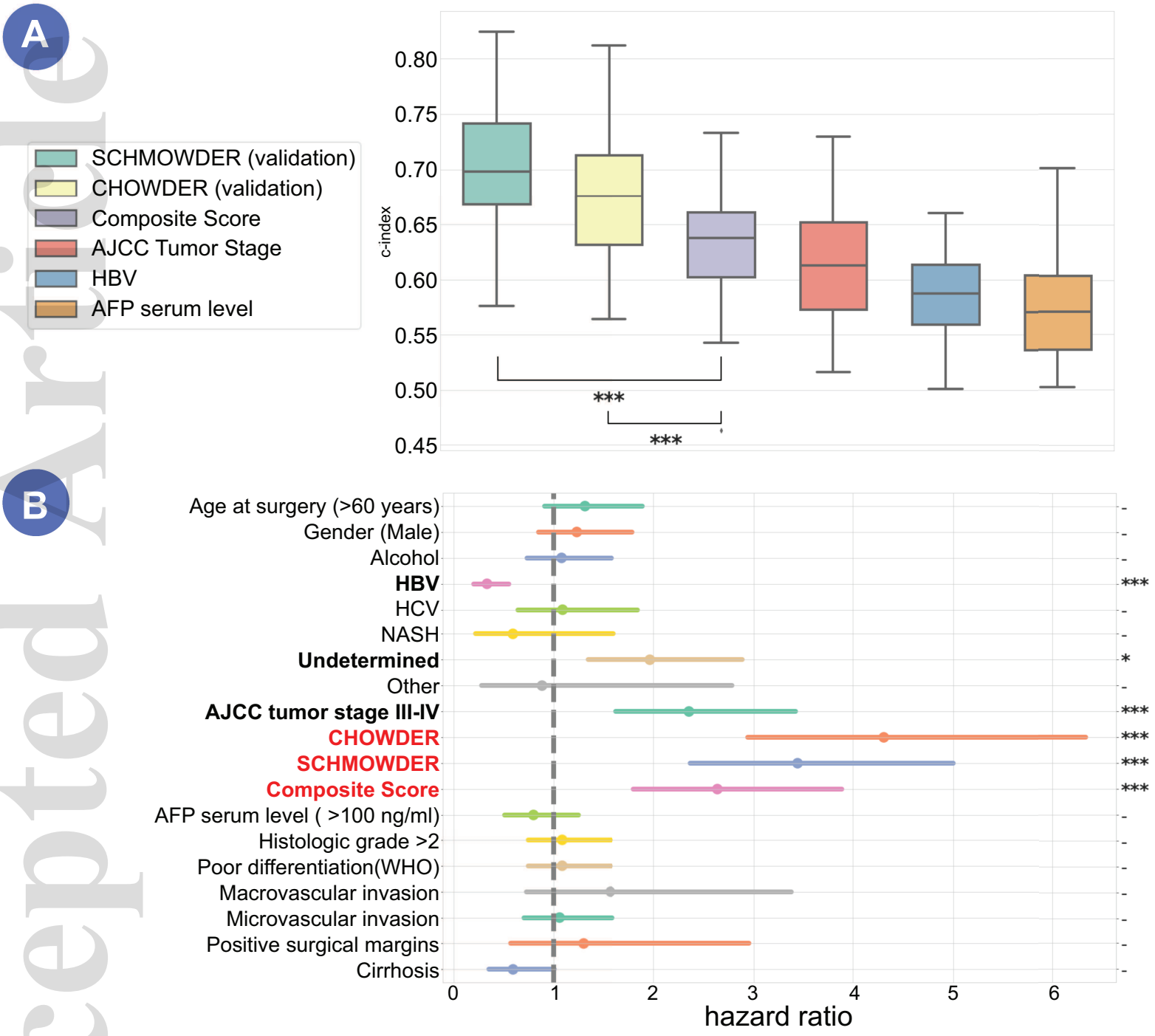




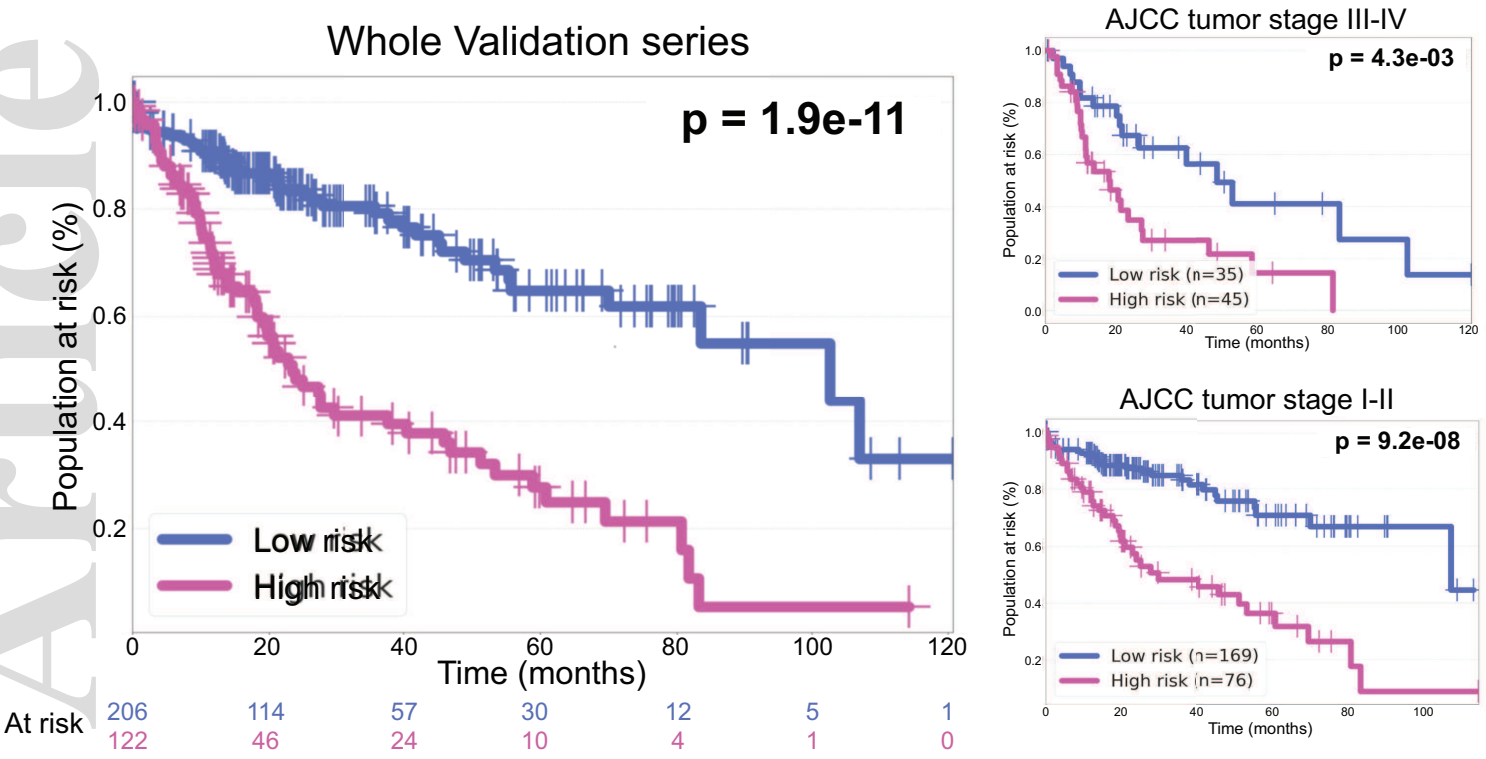
hep_31207_f3.eps

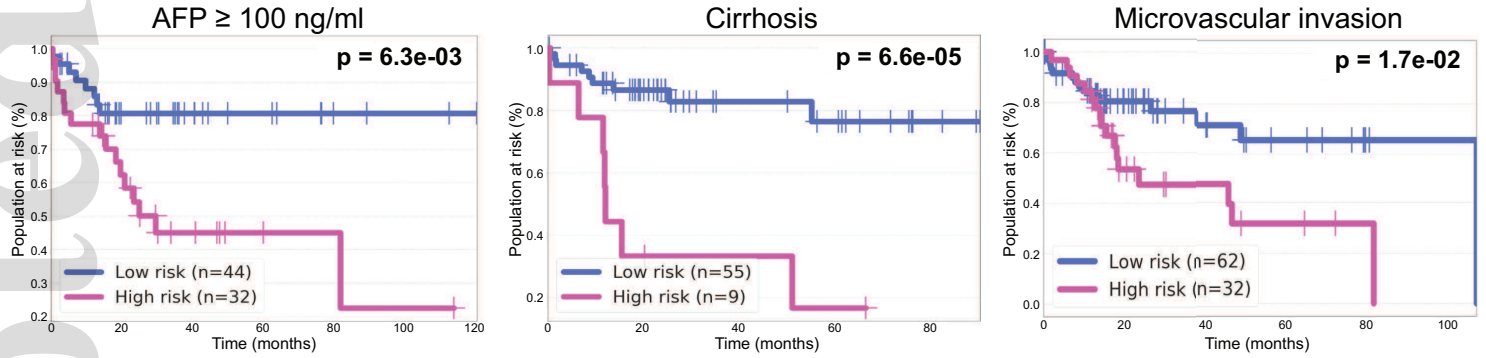


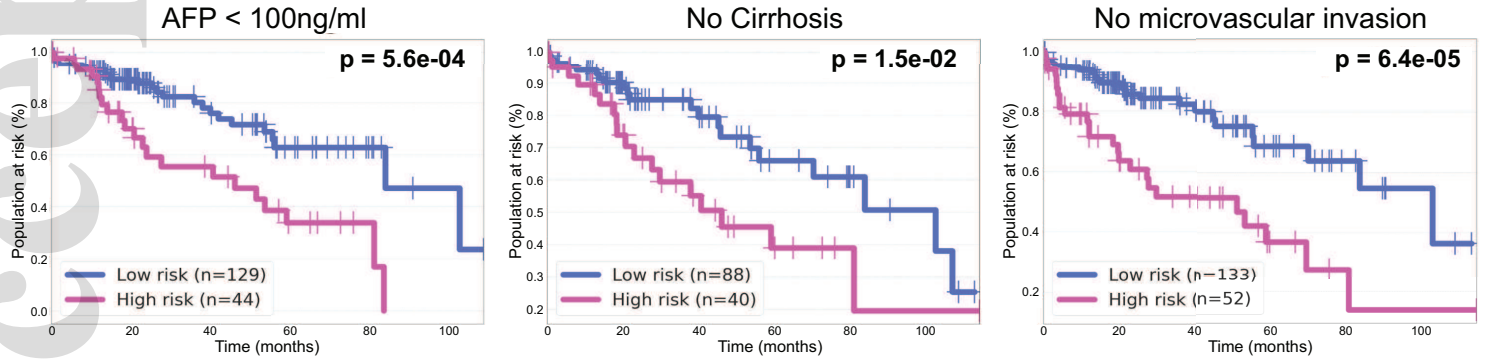
hep_31207_f4.eps



hep_31207_f5.eps







hep_31207_f6.eps