Improved assessment of donor liver steatosis using Banff consensus recommendations and deep learning algorithms

Research Article
Innovative Diagnostics, Modelling and Digital Hepatology

JOURNAL
OF HEPATOLOGY

Improved assessment of donor liver steatosis using Banff consensus recommendations and deep learning algorithms

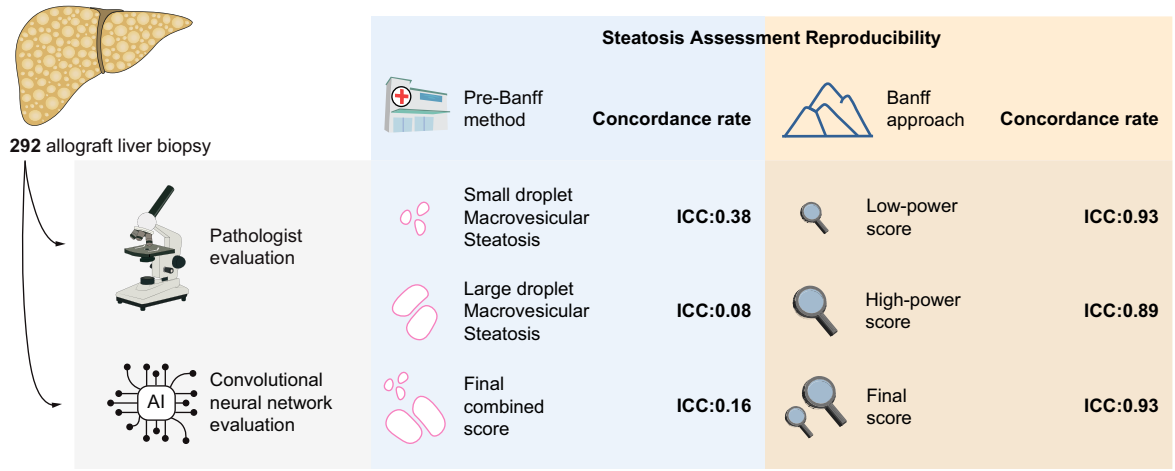
Authors

Alessandro Gambella, Massimo Salvi, Luca Molinaro, …, Mauro Papotti, Renato Romagnoli, Filippo Molinari

Correspondence

alessandro.gambella@unito.it (A. Gambella).

Graphical abstract



Highlights

- We developed and validated automated deep-learning algorithms for steatosis assessment based on Banff consensus recommendations.
- Our algorithm allows for an unbiased automated evaluation of steatosis.
- This will enable analysis of steatosis' effect on organ viability and the identification of clinically relevant cut-offs.
- Implementing our algorithm in daily clinical practice will enable more efficient and safe allocation of donor organs.

Impact and implications

We developed and validated the first automated deep-learning algorithms for standardized steatosis assessment based on the Banff Liver Working Group consensus recommendations. Our algorithm provides an unbiased automated evaluation of steatosis, which will lay the groundwork for granular analysis of steatosis's short- and long-term effects on organ viability, enabling the identification of clinically relevant steatosis cut-offs for donor organ acceptance. Implementing our algorithm in daily clinical practice will allow for a more efficient and safe allocation of donor organs, improving the post-transplant outcomes of patients.

https://doi.org/10.1016/j.jhep.2023.11.013

Research Article
Innovative Diagnostics, Modelling and Digital Hepatology

JOURNAL OF HEPATOLOGY

Improved assessment of donor liver steatosis using Banff consensus recommendations and deep learning algorithms

Alessandro Gambella[1,2,*,†], **Massimo Salvi**[3,†], Luca Molinaro[4], Damiano Patrono[5], Paola Cassoni[1], Mauro Papotti[6], Renato Romagnoli[5], Filippo Molinari[3]

Journal of Hepatology **2024**. vol. 80 | 495–504

Background & Aims: The Banff Liver Working Group recently published consensus recommendations for steatosis assessment in donor liver biopsy, but few studies reported their use and no automated deep-learning algorithms based on the proposed criteria have been developed so far. We evaluated Banff recommendations on a large monocentric series of donor liver needle biopsies by comparing pathologists' scores with those generated by convolutional neural networks (CNNs) we specifically developed for automated steatosis assessment.
Methods: We retrospectively retrieved 292 allograft liver needle biopsies collected between January 2016 and January 2020 and performed steatosis assessment using a former intra-institution method (pre-Banff method) and the newly introduced Banff recommendations. Scores provided by pathologists and CNN models were then compared, and the degree of agreement was measured with the intraclass correlation coefficient (ICC).
Results: Regarding the pre-Banff method, poor agreement was observed between the pathologist and CNN models for small droplet macrovesicular steatosis (ICC: 0.38), large droplet macrovesicular steatosis (ICC: 0.08), and the final combined score (ICC: 0.16) evaluation, but none of these reached statistically significance. Interestingly, significantly improved agreement was observed using the Banff approach: ICC was 0.93 for the low-power score (p <0.001), 0.89 for the high-power score (p <0.001), and 0.93 for the final score (p <0.001). Comparing the pre-Banff method with the Banff approach on the same biopsy, pathologist and CNN model assessment showed a mean (±SD) percentage of discrepancy of 26.89 (±22.16) and 1.20 (±5.58), respectively.
Conclusions: Our findings support the use of Banff recommendations in daily practice and highlight the need for a granular analysis of their effect on liver transplantation outcomes.

Introduction

The introduction of extended donor selection criteria led to the cautious but increasing use of donor livers with moderate steatosis (*i.e.*, steatosis affecting 30%-60% of the hepatic parenchyma).[1] However, livers with steatosis are more vulnerable to ischemia-reperfusion injury and are at a higher risk of early graft dysfunction and primary non-function,[2–5] thus making the pathologist assessment of donor tissue biopsy critical to evaluate organ eligibility and ensure optimal transplant success rates.[5–13] Remarkably, the lack of international guidelines for steatosis evaluation and reporting has resulted in the use of several different institution-tailored methods, thus leading to a heterogenous panorama and a non-negligible poor inter-observer reproducibility among pathologists.[14,15] To address this issue, the Banff Liver Working Group recently published consensus recommendations, now introducing standardized terminology and a specific diagnostic algorithm for steatosis assessment.[16] Their use in daily clinical practice is expected to improve pathologist reproducibility and provide standardized data for organ management and multicentric studies analyzing graft functionality, ultimately aiming to improve the allocation of steatotic livers.[15–17] Since their publication in December 2021, few studies have reported using Banff recommendations,[18,19] and none have tested their implication in the daily diagnostic practice or provided an automated digital algorithm for steatosis assessment based on their standardized criteria.

In the past decade, there has been a growing adoption of digital image analysis methods, particularly those using artificial intelligence algorithms. These methods have enabled pathologists to perform consistent and replicable histopathologic evaluations, reducing the burden of time-consuming and repetitive tasks. Due to their robust learning capabilities and ability to handle complex patterns, deep learning frameworks have rapidly emerged as the leading methodology for medical image

Keywords: digital pathology; steatosis assessment; liver allograft; donor eligibility; Banff consensus recommendations.
Received 28 August 2023; received in revised form 23 October 2023; accepted 3 November 2023; available online 29 November 2023
* Corresponding author. Address: Pathology Unit, Department of Medical Sciences, University of Turin, Turin, Italy; Tel.: +1 412 450 2829, fax: +39 011 633 4633.
E-mail address: alessandro.gambella@unito.it (A. Gambella).
† These authors contributed equally to this work.
https://doi.org/10.1016/j.jhep.2023.11.013

analysis, particularly in digital pathology. Current research is focused on developing fully automated methods based on artificial intelligence for quantitative histological analysis to create tools and instruments with high reproducibility.[20,21] These approaches can reduce inter- and intra-observer variability in the assessment of cellular structures, ultimately increasing the reproducibility of results. Given these advantages, digital image analysis methods are ideal for unbiased and reproducible assessments of steatosis on liver slides. In this regard, multiple deep learning-based approaches have been proposed to automatically detect steatosis.[14,22–24] However, these algorithms often fail to recognize single fat droplet steatosis[24] and perform poorly in the presence of highly overlapping steatosis droplets.[22] In addition, as mentioned above, no automated strategy has been presented so far for the assessment of steatosis according to the new Banff consensus recommendations.

In this study, we propose a novel deep-learning framework for the automated assessment of steatosis in allograft liver biopsy images. The main contributions of this paper can be summarized as follows:

- We propose a segmentation strategy capable of accurately segmenting single liver fat droplets, regardless of their size. To enhance the deep learning framework, we have integrated a stain normalization tool that can standardize the color appearance of the entire biopsy as a pre-processing step.
- We have developed two different strategies for evaluating hepatic steatosis using the same deep neural network. The first follows a traditional segmentation scheme based on an intra-institution method (pre-Banff method), while the second adheres to the recently published Banff consensus recommendations (Banff approach).
- We tested and validated the proposed strategy on a cohort of 292 allograft liver biopsies. We performed an extended validation of the deep learning framework by comparing our approach with the pathologist's score. Our method obtained highly satisfactory results, exhibiting a strong correlation with the pathologist's evaluation when adopting the Banff consensus recommendations.

Materials and methods

Liver biopsy

This monocentric study analyzed consecutive allograft liver needle biopsies performed between January 2016 and January 2020 at the AOU Città Della Salute e Della Scienza di Torino Hospital (Turin, Italy). Liver biopsies were obtained during organ retrieval or at the end of the transplant and then processed according to the routine laboratory procedures of the Pathology Unit as previously described[25,26] and detailed in the Supplementary Methods. Original H&E-stained glass slides were retrospectively retrieved from the Pathology Unit archives, along with the related diagnostic reports. Following tissue adequacy assessment (at least 2 cm-long tissue biopsy), 292 biopsies were included in the study. Before any analysis was performed, patient data were anonymized by a staff member not involved in the study.

Pathologist assessment of donor steatosis

This study used and compared two different procedures to assess steatosis. In particular:

1) *Pre-Banff method:* This is an intra-institutional method that considers the quantity of small droplet macrovesicular steatosis (SDMS), large droplet macrovesicular steatosis (LDMS), and combined SDMS-LDMS (CSL). According to this method, steatosis droplets larger than 200 μm^2 or that displaced the hepatocyte nuclei to the periphery of the cells are defined as LDMS. If these criteria are not met, the droplets are considered SDMS. Notably, the dimension of the droplets and the percentage of SDMS and LDMS are estimated directly by pathologists without using any specific approach or mathematical formula (*i.e.*, "eyeball" assessment). In the final diagnostic report, the pathologist decides whether to report the specific percentages of SDMS, LDMS, or CSL. The original scores provided at the time of diagnosis were collected and reviewed. This approach was used by the Transplant Pathology Unit of the AOU Città Della Salute e Della Scienza di Torino Hospital before the Banff consensus recommendations were published.
2) *Banff approach:* these recommendations provide accurate definitions and a three-step diagnostic approach to report the percentage of steatosis affecting the hepatic parenchyma.[16] Briefly, LDMS is defined as a fat droplet larger than a non-steatotic nearby hepatocyte, and that displaces the nucleus to the periphery of the hepatocyte that contains it. Fat droplets that do not satisfy these criteria are defined as SDMS but are not considered relevant in the overall assessment of organ steatosis due to biological and pathophysiological considerations.[16] Regarding the three-step diagnostic approach, first, the overall percentage of steatosis affecting the whole biopsy is assessed at low power (LP), generating the LP score. Then, a high-power (HP) assessment of the areas of steatosis is performed to determine the exact percentage of LDMS only (HP score). Finally, a final score (LS) is calculated by combining the LP and HP scores (HP of the LP). All the donor needle biopsies were revised and scored according to the Banff approach. The Transplant Pathology Unit of the AOU Città Della Salute e Della Scienza di Torino Hospital now uses this approach in the daily diagnostic routine.

Steatosis assessment was performed by two transplant pathologists (A.G. and L.M.) for both the Pre-Banff Method and the Banff Approach by visual assessment only with no support from any specific software. In case of disagreement, a consensus was reached by joint review and discussion. Definitions used by the two methods are summarized in Table 1.

Image normalization and automatic steatosis segmentation

All the original H&E slides were reviewed to confirm specimen adequacy and then scanned to obtain digital whole slide images (WSIs) as previously described[27,28] and detailed in the Supplementary Methods. Then, WSIs were adjusted for stain normalization. Stain normalization is a common pre-processing step in almost all the deep learning frameworks in digital pathology.[14,29–31] Briefly, a stain normalizing procedure allows for standardization of the color appearance of a source image for the color profile of a template image. This operation reduces the stain variability and improves the robustness of computer-aided diagnostic and image quantification algorithms.[29,30] To obtain precise quantification of steatosis within the WSI, a convolutional neural network (CNN) is employed. In particular, the segmentation is performed with the same CNN architecture we developed in our previous work.[14] To improve the segmentation of fused or touching droplets, we modified our CNN

Table 1. Details of definitions and scores of the two approaches used for steatosis assessment.

	Pre-Banff method	Banff approach
Definitions		
Macrovesicular steatosis		
Large droplet (LDMS)	Fat droplet larger than 200 μm^2 or that displaced the hepatocyte nuclei to the periphery of the cell	Fat droplet larger than a non-steatotic nearby hepatocyte or that displaced the hepatocyte nuclei to the periphery of the cell
Small droplet (SDMS)	Any fat droplets that do not satisfy the criteria for LDMS and are not microvesicular steatosis	Any fat droplets that do not satisfy the criteria for LDMS and are not microvesicular steatosis
Microvesicular steatosis*	Diffuse, faint small lipid droplets (typically <2 μm^2) that do not displace the hepatocyte nuclei and determine an overall "foamy" appearance of the hepatocyte cytoplasm	
Scores		
LDMS score	Average percentage of LDMS affecting the liver parenchyma	n.a.
SDMS score	Average percentage of LDMS affecting the liver parenchyma	n.a.
Combined SDMS-LDMS (CSL)	Combination of LDMS and SDMS scores	n.a.
Low power score	n.a.	Overall percentage of steatosis affecting the hepatic parenchyma at low power
High power score	n.a.	High power assessment of steatosis areas to determine the exact percentage of LDMS
Final score (LS)	n.a	Adjustment of the LP score with the HP score (*i.e.*, the percentage of LDMS determined with the HP score is applied to the LP score)

HP, high power; LDMS, large droplet macrovesicular steatosis; LP, low power; SDMS, small droplet macrovesicular steatosis.
*This type of steatosis is secondary to specific diseases (*e.g.*, acute fatty liver of pregnancy, valproic acid toxicity) that are unlikely to affect a donor liver evaluated for organ transplantation.

task by adopting a three-class segmentation approach: (i) steatosis, (ii) steatosis boundary, and (iii) background. The reason for choosing both object and edge detection is to define the spatial limit of each area of steatosis based on the information on the location and contour of each object (Fig. S2). Further technical details on stain normalization and steatosis segmentation are reported in the Supplementary Methods. Data, intermediate steps, and final results of the algorithm for quantification of hepatic steatosis are available online at the following link: 10.17632/cjgd4wr2tz.1.

WSI analysis before and after the Banff consensus recommendations

This section explains how the segmentation algorithm was employed to replicate the former steatosis assessment method (Pre-Banff method) and the newly introduced approach according to the Banff consensus recommendations (Banff approach).

WSI analysis before the Banff consensus recommendations (Pre-Banff method)

The first step involved separating the histological tissue from the background using a thresholding operation. Next, the segmentation network described in the previous section was applied to detect steatosis in the WSI. The segmentation map for the entire slide is created patch-wise using a sliding window approach. Using criteria similar to those available in the literature[32] and following the pre-Banff method used by pathologists, SDMS and LDMS were separated based on an area criterion. Structures with an area smaller than 200 μm^2 were labeled as SDMS, while the others were labeled as LDMS as previously reported.[4,32] Finally, the percentages of SDMS and LDMS were calculated as the ratio of the area occupied by fat droplets to histologic tissue. We also calculated the percentage occupied by SDMS and LDMS combined (overall steatosis), regardless of their classification. The pipeline followed for the Pre-Banff method is shown in Fig. 1.

WSI analysis after the Banff consensus recommendations (Banff approach)

The first step consists of histological tissue and steatosis segmentation on WSI. For the quantification of the LP score, the algorithm works at a magnification of 25x, thus simulating the low-magnification evaluation as described by the Banff consensus recommendations. Steatosis regions close to each other were joined using morphological operators, while isolated steatosis was removed, and the remaining regions' perimeter was interpolated to obtain the raw surface area occupied by fat. The LP score was then calculated by dividing the raw surface area by the tissue area.

To identify the area for high-magnification assessment, an iterative approach was used. A sliding window over the entire WSI was used, and the tile of size 190x190 at 25x (equivalent to 1520x1520 at 200x) with the maximum area occupied by steatosis according to the LP score was selected. Cell nuclei were then identified on this tile using a previously published algorithm.[28,33] Identified structures were subjected to morphological cleaning, where all nuclei with an area less than 22 μm^2 and an axis ratio greater than 0.75 were deleted. This was done to keep only the nuclei of hepatocytes and remove other cells not of interest, such as immune cells, Kupffer cells, endothelial cells, cholangiocytes, and hepatic stellate cells. The area of a hepatocyte (nucleus and cytoplasm) is approximately five times larger than the area of its nucleus,[34,35] and this was used as a cut-off to recognize LDMS, following the Banff

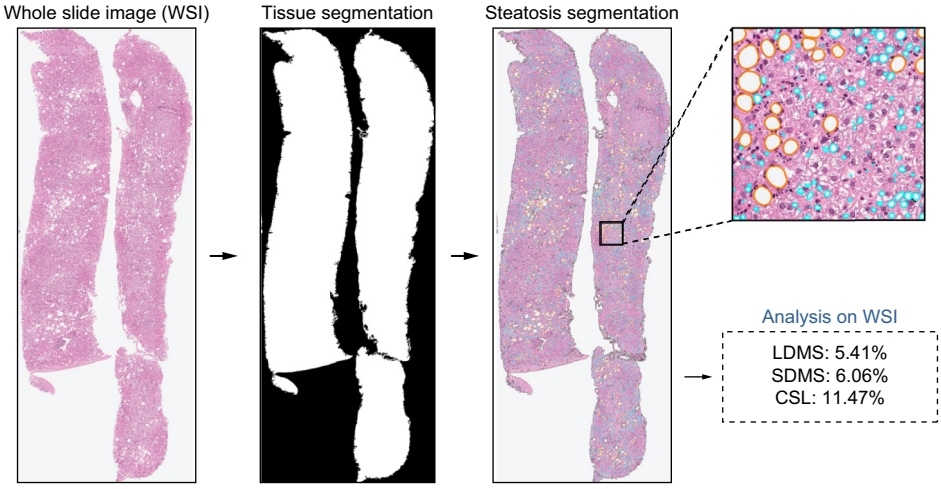


Fig. 1. Steps followed by the algorithm during steatosis assessment (Pre-Banff method). The automatic method provides three outputs: percentage of area occupied by LDMS, percentage of area occupied by SDMS, and percentage area occupied by SDMS and LDMS combined (overall steatosis). LDMS is displayed in orange, while SDMS is shown in cyan. CSL, combined small droplet macrovesicular steatosis-large droplet macrovesicular steatosis; LDMS, large droplet macrovesicular steatosis; SDMS, small droplet macrovesicular steatosis; WSI, whole slide image. (This figure appears in color on the web.)

consensus recommendations. All steatosis with smaller areas was eliminated, and the HP score was calculated as follows:

$$HP_{SCORE} = \frac{N_{macro}}{N_{macro}+N_{nuclei}} \quad (1)$$

where N_{macro} and N_{nuclei} indicate the number of areas of macrosteatosis and nuclei, respectively. The final score (LS) was obtained as the product of the LP and HP scores. A visual representation of the entire method is illustrated in Fig. 2. The codes developed and performed are detailed in the Supplementary Methods.

Performance metrics

Automatic and manual masks are compared to evaluate the performance of our method in the segmentation of liver steatosis. To assess the segmentation performance, various pixel-based metrics were calculated, including accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and Dice score (Table 2).

Ethics approval

All procedures were in accordance with the ethical standards of the responsible committee on human experimentation (national and institutional) and with the World Medical Association Declaration of Helsinki of 1964 and later versions. Written informed consent for participation was waived for this study due to the retrospective nature of the research protocol and considering that it had no impact on patients' care.

Statistical analysis

Statistical analysis was performed using R Software (version 4.2.2; The R Foundation for Statistical Computing, Vienna, Austria) and RStudio (version 2022.12.0+353; RStudio, Boston (MA), USA). Results of the Pre-Banff method (SDMS, LDMS, and CSL) and the Banff approach (LP, HP, and LS) were reported as mean ± SD. The difference in assessment between the pathologist and the CNN model was considered as follows: minimal (<1%), mild (1-10%), moderate (10-30%), and severe (>30%). The degree of agreement between pathologists and the CNN model in assessing donor steatosis was measured with the intraclass correlation coefficient ("irr" package) using the two-way random effects model with single measures [ICC(2,1)]. The ICC(2,1) was calculated using a two-way ANOVA model, providing a numerical value between 0 and 1, where higher values indicate higher agreement between the raters. Specifically, the following thresholds for ICC value were used for agreement interpretation: <0.50: poor; 0.50-0.75: moderate; 0.75-0.90: good; >0.90: excellent.[36] The ICC(2,1) of the agreement was calculated for both the Pre-Banff method and the Banff approach. Based on the literature, we assumed an agreement of 0.5 between evaluators and set the null hypothesis accordingly (H0: r0 = 0.5; H1: r0 >0.5). All tests were two-sided, and a p value <0.05 was considered statistically significant.

Results

Steatosis droplet segmentation using deep learning

Overall, 292 consecutive allograft liver biopsies were retrospectively collected and included in this study. The fully automated results provided by the deep learning method are compared with manual masks drawn by an expert operator. A quantitative comparison is carried out by evaluating the accuracy, sensitivity, and specificity of the segmentation of liver steatosis. Table 2 shows the segmentation algorithm's performance on the train set (504 image tiles) and test set (56 image tiles).

Correlation with pathologist assessment using the Pre-Banff method

This first analysis aimed to evaluate the agreement between the pathologist and the CNN model using the pre-Banff method for steatosis evaluation (Table 3). We first compared the data regarding the SDMS assessment, which was available for 221 biopsies. The mean (±SD) difference in percentage values obtained between the pathologist and the CNN model was 6.20

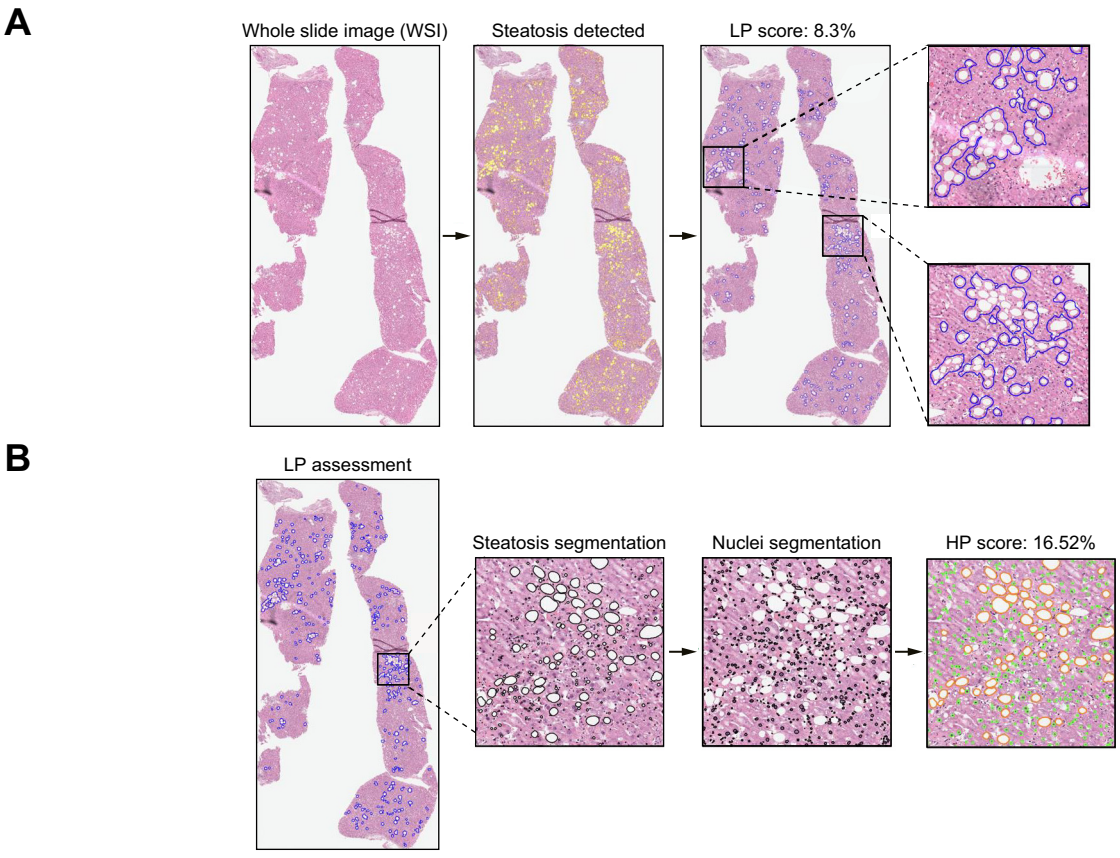


Fig. 2. Steps followed by the algorithm during steatosis assessment (Banff approach). (A) LP score assessment: the algorithm groups areas of steatosis close to each other and computes the LP score. (B) HP score assessment: steatosis and nuclei are automatically segmented through deep learning. After morphological cleaning, the HP score is computed. (This figure appears in color on the web.)

Table 2. Segmentation performance of the deep learning method for steatosis segmentation on both training and testing sets.

Subset	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Dice score (%)
Training set	99.04 ± 1.42	73.06 ± 22.76	99.28 ± 1.20	70.61 ± 14.73	98.03 ± 1.12	84.38 ± 11.08
Testing set	99.18 ± 1.39	73.36 ± 28.17	99.37 ± 1.18	72.95 ± 13.84	98.05 ± 1.08	85.28 ± 10.62

Values are reported as mean ± SD.
PPV, positive predictive value; NPV, negative predictive value.

(±8.17). Minimal, mild, moderate, and severe differences in SDMS assessment were observed in 88 (39.82%), 83 (37.56%), 46 (20.81%), and 4 (1.81%) cases, respectively (Fig. 3A). Notably, the pathologist recorded a higher percentage of SDMS than the algorithm in all the mild, moderate, and severe cases. The highest difference was observed in one case where the pathologist identified 60% of SDMS while the CNN model scored 11.70% (Fig. 3B). The pathologist and the CNN model reported the same SDMS value in two biopsies.

A more relevant difference was observed in the evaluation of LDMS, which was available for 225 biopsies. The mean discordant percentage was 19.41 (±20.07). Minimal, mild, moderate, and severe differences were observed in 33 cases (14.67%), 74 cases (32.89%), 75 cases (33.33%), and 43 cases (19.11%), respectively (Fig. 3C). In two cases with mild differences, the CNN model reported a higher LDMS than the pathologist (difference of 1.00% and 1.26%, respectively), whereas, in all moderate and severe cases, the pathologist

Table 3. Comparison between pathologist and CNN model assessment using the pre-Banff method.

	Pathologist assessment (mean ± SD)	CNN model (mean ± SD)	Mean difference (±SD)	Minimal (%)	Mild (%)	Moderate (%)	Severe (%)	ICC score	*p* value
SDMS (n = 221)	7.61 (±10.38)	1.13 (±2.49)	6.20 (±8.17)	88 (39.82%)	83 (37.56%)	46 (20.81%)	4 (1.81%)	0.38	0.938
LDMS (n = 225)	20.90 (±21.07)	1.31 (±1.30)	19.41 (±20.07)	33 (14.67%)	74 (32.89%)	75 (33.33%)	43 (19.11%)	0.08	1
CSL (n = 211)	27.77 (±22.55)	2.44 (±3.34)	25.00 (±21.47)	13 (6.16%)	58 (27.49%)	68 (32.23%)	72 (34.12%)	0.16	1

In addition to the mean percentage (±SD) of steatosis assessment reported by the pathologist and the CNN model, the table shows the mean difference, the number of cases presenting a minimal (<1%), mild (1-10%), moderate (10-30%), and severe (>30%) difference of steatosis, and the ICC score and related *p* value. *P* values <0.05 were considered statistically significant (ICC score).
CNN, convolutional neural network; CSL, combined small droplet macrovesicular steatosis-large droplet macrovesicular steatosis; ICC, intraclass correlation coefficient; LDMS, large droplet macrovesicular steatosis; SDMS, small droplet macrovesicular steatosis.

recorded a higher percentage of LDMS than the CNN model. The most notable variation was observed in one case where the pathologist assigned a value of 90%, while the algorithm reported 3.86%. The pathologist and the CNN model scored no biopsies with the same LDMS value.

Finally, we assessed the CSL, which considered both SDMS and LDMS and was available for 211 biopsies. The mean percentage of discrepancy was 25.00 (±21.47), with a minimal difference observed in 13 cases (6.16%), a mild difference in 58 cases (27.49%), a moderate difference in 68 cases (32.23%), and a severe difference in 72 cases (34.12%) (Fig. 3D). In 9 of the 13 mild cases and all moderate and severe cases, the pathologist reported a higher percentage than the CNN model. The highest difference was observed in a biopsy where the pathologist reported a CSL of 90% and the CNN model of 3.95%. The pathologist and the CNN model reported the same CSL value in six biopsies.

As expected, the coefficient of agreement between the pathologist assessment and the CNN model using the Pre-Banff method was poor when considering either the SDMS (0.38; p >0.05), the LDMS (0.08; p >0.05), or the CSL score (0.16; p >0.05). These findings highlighted the poor concordance between the pathologist and the deep learning algorithm when adopting the Pre-Banff method.

Correlation with pathologist assessment using the Banff approach

We assessed whether the newly introduced consensus recommendations (Banff approach) presented similar or different concordance rates compared to the Pre-Banff method. All 292 biopsies were evaluated with the Banff approach, specifically reporting the LP, HP, and LS scores (Table 4).

First, we addressed the difference between the pathologist and CNN model in assessing the LP score, registering a mean (±SD) difference of 1.10 (±0.22). Minimal, mild, and moderate differences in LP scores were observed in 172 (58.90%), 111 (38.01%), and 9 (3.08%) cases, respectively (Fig. 4A). No severe differences were registered. In 38 of the 172 minimal cases (22.09%), 70 of the 111 mild cases (63.06%), and all nine moderate cases (100%), the pathologist recorded a higher

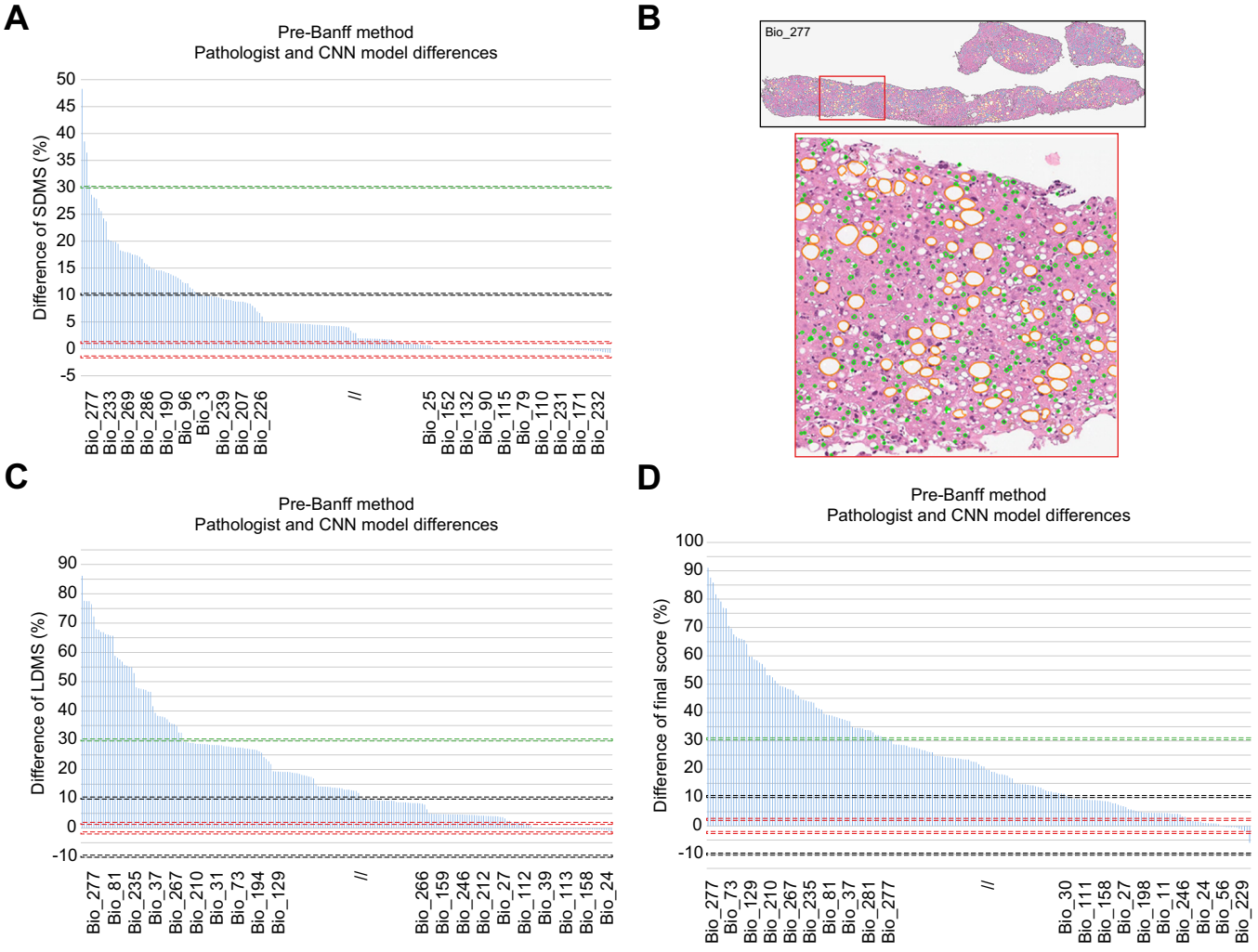


Fig. 3. Comparison between pathologists and the CNN model using the Pre-Banff method. (A, C, D) Waterfall plots showing the difference in SDMS (A), LDMS (C), and CSL (D) between the pathologist and CNN models for each biopsy. Y axis represents the difference of SDMS percentage where positive values indicate that the percentage of SDMS reported by the pathologist was greater than the CNN model and negative values indicate that the percentage of SDMS reported by the pathologist was lower than the CNN model. (B) Representative image at low power (upper part) and high power (lower part) of the biopsy with the highest value of discordant SDMS score. CNN, convolutional neural network; LDMS, large droplet macrovesicular steatosis; SDMS, small droplet macrovesicular steatosis. (This figure appears in color on the web.)

Table 4. Comparison between pathologist and CNN model assessment using the Banff approach.

	Pathologist assessment (mean ± SD)	CNN model (mean ± SD)	Mean (±SD)	Minimal (%)	Mild (%)	Moderate (%)	Severe (%)	ICC score	*p* value
LP score (n = 292)	5.74 (±10.98)	4.64 (±8.77)	1.10 (±0.22)	172 (58.90%)	111 (38.01%)	9 (3.08%)	0 (0%)	0.93	<0.001
HP score (n = 292)	6.4 (±13.06)	6.10 (±11.74)	0.30 (±0.34)	115 (39.38%)	158 (54.11%)	16 (5.48%)	3 (1.03%)	0.89	<0.001
LS (n = 292)	1.62 (±5.39)	1.23 (±4.24)	0.39 (±0.11)	257 (88.01%)	32 (10.96%)	3 (1.03%)	0 (0%)	0.93	<0.001

In addition to the mean percentage (±SD) of steatosis assessment reported by the pathologist and the CNN model, the table shows the mean difference, the number of cases presenting a minimal (<1%), mild (1-10%), moderate (10-30%), and severe (>30%) difference of steatosis, and the ICC score and related *p* value. *P* values <0.05 were considered statistically significant (ICC score).

CNN, convolutional neural network; HP, high power; ICC, intraclass correlation coefficient; LP, low power; LS, final score.

percentage of LP than the CNN model. The highest difference was observed in a biopsy where the pathologist identified 60% steatosis on LP while the CNN model scored 38% (Fig. 4B). The pathologist and the CNN model reported the same LP value in three biopsies.

Regarding the HP score, the mean (±SD) difference was 0.30 (±0.34). In 115 cases (39.38%), the difference between the HP score provided by the pathologist and the CNN model was minimal. A mild difference was observed in 158 cases (54.11%), a moderate difference in 16 cases (5.48%), and a severe difference in 3 cases (1.03%; Fig. 4C). The pathologist provided a higher HP score than the CNN model in 47 of the 158 mild cases (29.75%), 11 of the 16 moderate cases (68.75%), and all the severe cases (100%). The highest difference was observed in a biopsy where the pathologist identified 60% steatosis on HP while the CNN model scored 17% (Fig. 4D). In 115 biopsies, the pathologist and the CNN model reported the same HP score.

Finally, we assessed the LS, which considered both LP and HP scores (HP of LP). The mean (±SD) difference of percentage reported by the pathologist and the CNN model was 0.39 (±0.11), with a minimal difference observed in 257 cases (88.01%), a mild difference in 32 cases (10.96%), and a moderate difference in 3 cases (1.03%). No biopsies with a severe difference were observed (Fig. 4E). The pathologist reported a higher percentage than the CNN model in 106 of the 257 minimal cases (41.24%), 25 of the 32 mild cases (78.12%), and all moderate cases (100%). The highest difference was observed in a biopsy where the pathologist reported an HP score of 32%, while the CNN model scored 16%. The pathologist and the CNN model reported the same LS score in 121 biopsies (Fig. 4F).

The coefficient of agreement between the pathologist assessment and the CNN model using the Banff approach was good for the HP score (0.89; p <0.001) and excellent when considering both the LP score (0.93; p <0.001) and the LS (0.93; p <0.001).

Same biopsy, different fates: Comparison of the Pre-Banff method and the Banff approach

We aimed to evaluate the difference in steatosis percentage on the same biopsy when using the Pre-Banff method and the Banff approach. Due to the structural differences between these two scores, we decided to consider and compare only the final combined scores, namely the CSL (Pre-Banff method) and the LS (Banff approach).

Starting with the pathologist assessment, the mean percentage of discrepancy was 26.89 (±22.16), with a minimal difference observed in 11 cases (5.21%), a mild difference in 57 cases (27.01%), a moderate difference in 68 cases (32.23%), and a severe difference in 75 cases (35.54%; supplementary information). In 8 of the 11 minimal cases (72.73%), 54 of the 57 mild cases (94.74%), 67 of the 68 moderate cases (98.53%), and all severe cases (100%), the CSL (Pre-Banff method) was higher than the LS (Banff approach). The highest difference was observed in a biopsy where the CSL was 95% and the LS was 3%. Six cases presented the same CSL and LS values.

We performed a similar analysis considering the data provided by the CNN models. We considered only the CSL and the LS scores as we did with the pathologist's evaluations. The mean percentage of discrepancy was 1.20 (±5.58), with a minimal difference observed in 124 cases (42.47%), a mild difference in 142 cases (48.63%), and a moderate difference in 26 cases (8.90%; supplementary information). No cases presented a severe difference between the CSL and LS values provided by the CNN models. In 12 of the 124 minimal cases (9.68%), 20 of the 142 mild cases (14.10%), and 11 of the 26 moderate cases (42.31%), the CSL (Pre-Banff method) was higher than the LS (Banff approach). The highest difference was observed in a biopsy where the CSL was 0.72% and the LS was 27.01%. No cases presented the same CSL and LS values.

Discussion

This study produced innovative evidence and considerations, including (1) the first application of the Banff consensus recommendations in a large series of allograft liver biopsies and (2) the first published deep-learning algorithms for automated steatosis assessment based on Banff recommendations. Furthermore, by comparing this innovative algorithm with a model based on a non-standardized method lacking an analytical approach, we demonstrated the importance of using the Banff recommendations in daily clinical practice, particularly if supported by an automated deep-learning algorithm. Using a standardized approach for steatosis definition and quantification following the Banff recommendations led to excellent concordance rates between the pathologist assessment and the CNN model. This finding is particularly relevant for the LS score (ICC: 0.93; p <0.001), which represents the final assessment of overall donor organ steatosis and ultimately provides the most relevant data for subsequent organ allocation.

When biopsies were re-evaluated by the pathologist for steatosis assessment with the Banff recommendations, the percentage of steatosis generally decreased compared to the pre-Banff method, with some cases presenting dramatic changes (95% with the pre-Banff method, 3% with the Banff approach). This difference may be due to the different definitions used by the two systems and, in particular, the absence of the SDMS from the Banff approach, which was still considered in the pre-Banff method. Compared to the CNN model, pathologist assessment seems to overestimate steatosis regardless of the approach used, as reported in the literature.[32] The lower sensitivity of our CNN model compared to its high specificity, as highlighted in Table 2, may have contributed to an underestimation of steatosis percentages. Conversely,

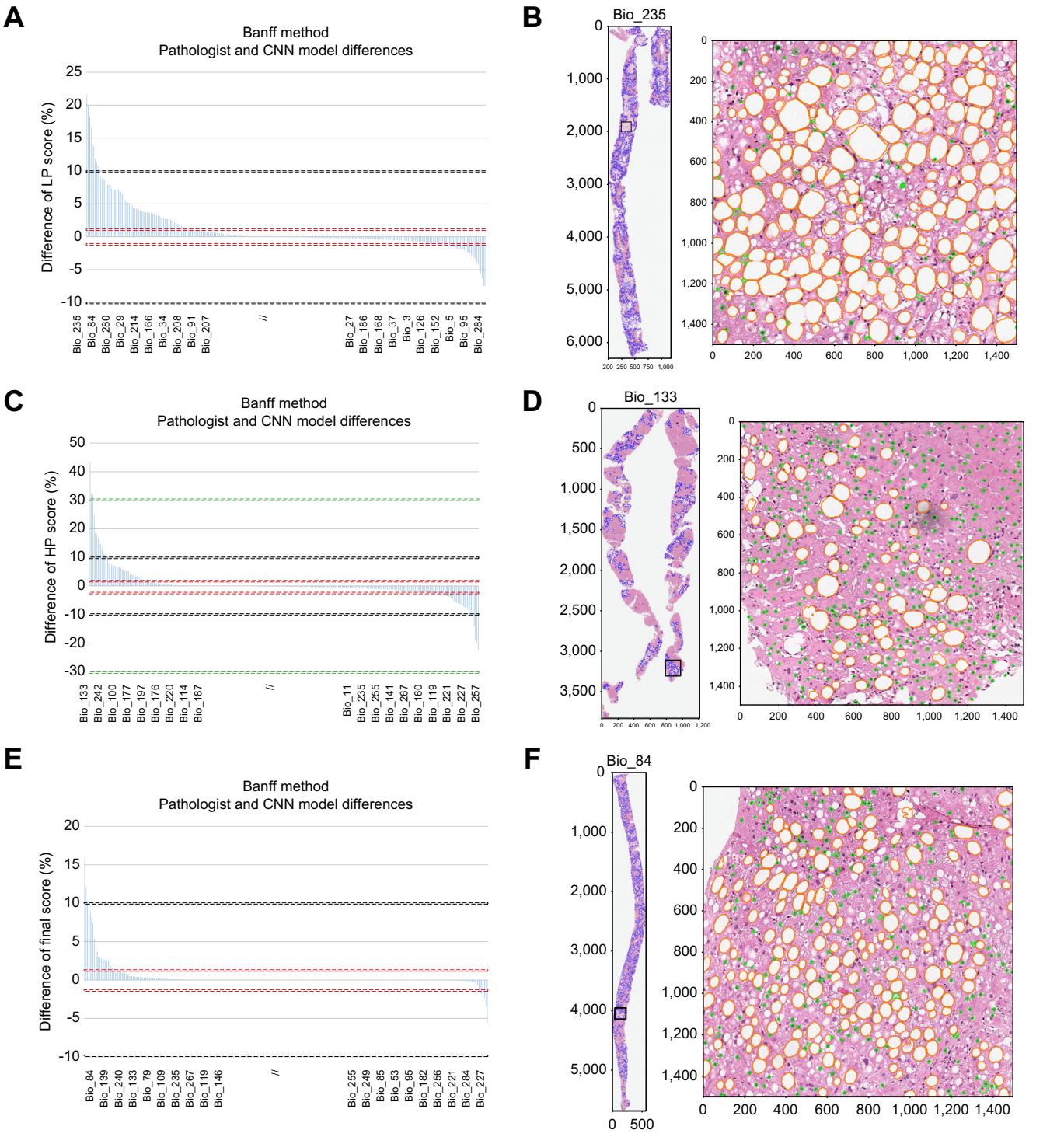


Fig. 4. Comparison between pathologists and CNN model using the Banff approach. (A, C, E) Waterfall plot showing the difference of LP score (A), HP score (C), and LS (E) between the pathologist and CNN models for each biopsy. Y axis represents the difference of LP score, where positive values indicate that the percentage of LP score reported by the pathologist was greater than the CNN model and negative values indicate that the percentage of LP score reported by the pathologist was lower than the CNN model. (B, D, F) Representative image at low power (left part) and high power (right part) of the biopsy with (B) the highest value of discordant LP score, (D) the highest value of discordant HP score, and (F) the highest value of discordant LS. CNN, convolutional neural network; HP, high power; LP, low power; LS, final score. (This figure appears in color on the web.)

pathologist's visual assessment potentially provides overestimated evaluations when performing repetitive tasks, especially those requiring specific quantification of numerous objects. This limitation can be mitigated using a standardized algorithmic approach as introduced by the Banff recommendations, partially explaining why, in our study, the pathologist's

overestimation was notably more pronounced with the pre-Banff method. These findings highlight the possibility that we may be discarding more steatotic livers than necessary. To confirm this hypothesis, we are currently analyzing the clinical outcomes of our liver transplant series using the Banff approach, aiming to identify more reliable and clinically significant steatosis cut-offs.

Compared to previously published algorithms for steatosis assessment,[14,22–24] our CNN model introduced significant improvement. The segmentation framework was designed to identify individual liver droplets regardless of their size, which is a crucial feature for accurately differentiating SDMS and LDMS following the new Banff recommendations. This approach also enables the segmentation of nearby droplets, preventing them from merging into a single, larger droplet. Additionally, implementing the stain normalization add-on in the CNN model further enhances the standardization of analysis, in line with the recommendation of the Banff consensus meeting.[16]

Although our study biopsies were obtained from one of the leading liver transplantation programs in Italy, with the highest number of transplants performed,[37] our study is limited by its monocentric and retrospective nature. To further strengthen the evidence and validate our findings, it is essential that future studies are conducted using external datasets. Furthermore, our pathology laboratory routinely performs rapid microwave-assisted tissue processing on transplant biopsies, which generate formalin-fixed paraffin-embedded tissue even in urgent settings. Although this procedure enables us to evaluate tissue samples without frozen artifacts and maintain a rapid turnaround time, it is not widely available in all centers. Therefore, further confirmation of our findings on a series of frozen sections is required.

In conclusion, our study demonstrates that introducing the standardized definition and analytical approach provided by the Banff consensus recommendations improves steatosis assessment in donor liver biopsies, especially when utilizing an automated digital pathology algorithm. The use and implementation of these guidelines in the daily diagnostic routine will lead to a new interpretation and clinical management of hepatic steatosis in liver transplantation.

Affiliations

[1]Pathology Unit, Department of Medical Sciences, University of Turin, Turin, Italy; [2]Division of Liver and Transplant Pathology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; [3]Department of Electronics and Telecommunications, PolitoBIOMed Lab, Politecnico di Torino, Biolab, Corso Duca degli Abruzzi 24, 10129 Turin, Italy; [4]Division of Pathology, AOU Città Della Salute e Della Scienza di Torino, Turin, Italy; [5]General Surgery 2U, Liver Transplant Center, AOU Città Della Salute e Della Scienza di Torino, University of Turin, Turin, Italy; [6]Division of Pathology, Department of Oncology, University of Turin, Turin, Italy

Abbreviations

CNNs, convolutional neural networks; CSL, combined small droplet macrovesicular steatosis-large droplet macrovesicular steatosis; HP, high power; ICC, intraclass correlation coefficient; LDMS, large droplet macrovesicular steatosis; LP, low power; LS, final score; SDMS, small droplet macrovesicular steatosis; WSIs, whole slide images.

Financial support

The author(s) received no specific funding for this work.

Conflict of interest

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

Please refer to the accompanying ICMJE disclosure forms for further details.

Authors' contributions

LM and FM performed the study concept and design; AG, MS, and LM provided acquisition, analysis, and interpretation of data and statistical analysis. AG and MS developed the methodology and wrote the original draft; DP, PC, MP, RR, and FM supervised study development. All authors read and approved the final paper.

Data availability statement

Part of the dataset and the results generated during the current study are available in the Mendeley Dataset repository (https://doi.org/10.17632/cjgd4wr2tz.1). The source codes and the complete dataset are available from the authors upon reasonable request.

Acknowledgments

We would like to express our gratitude to the laboratory staff of the Pathology Unit at the AOU Città della Salute e della Scienza Hospital for their outstanding work in handling and processing tissue specimens in the urgent setting of transplantation.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhep.2023.11.013.

References

Author names in bold designate shared co-first authorship

[1] Hashimoto K, Miller C. The use of marginal grafts in liver transplantation. J Hepatobiliary Pancreat Surg 2008;15:92–101.
[2] Todo S, Demetris AJ, Makowka L, et al. Primary nonfunction of hepatic allografts with preexisting fatty infiltration. Transplantation 1989;47:903–905.
[3] Angelico M. Donor liver steatosis and graft selection for liver transplantation: a short review. Eur Rev Med Pharmacol Sci 2005;9:295–297.
[4] Choi WT, Jen KY, Wang D, et al. Donor liver small droplet macrovesicular steatosis is associated with increased risk for recipient allograft rejection. Am J Surg Pathol 2017;41:365–373.
[5] Spitzer AL, Lao OB, Dick AA, et al. The biopsied donor liver: incorporating macrosteatosis into high-risk donor assessment. Liver Transpl 2010;16:874–884.
[6] Saidi RF. Utilization of expanded criteria donors in liver transplantation. Int J Organ Transpl Med 2013;4:46–59.
[7] Hamar M, Selzner M. Steatotic donor livers: where is the risk-benefit maximized? Liver Transpl 2017;23:S34–S39.
[8] Croome KP, Lee DD, Taner CB. The "skinny" on assessment and utilization of steatotic liver grafts: a systematic review. Liver Transpl 2019;25:488–499.
[9] de Graaf EL, Kench J, Dilworth P, et al. Grade of deceased donor liver macrovesicular steatosis impacts graft and recipient outcomes more than the Donor Risk Index. J Gastroenterol Hepatol 2012;27:540–546.
[10] Lo IJ, Lefkowitch JH, Feirt N, et al. Utility of liver allograft biopsy obtained at procurement. Liver Transpl 2008;14:639–646.
[11] **Fiorentino M**, **Vasuri F**, Ravaioli M, et al. Predictive value of frozen-section analysis in the histological assessment of steatosis before liver transplantation. Liver Transpl 2009;15:1821–1825.
[12] Holowko W, Mazurkiewicz M, Grat M, et al. Reliability of frozen section in the assessment of allograft steatosis in liver transplantation. Transpl Proc 2014;46:2755–2757.
[13] Flechtenmacher C, Schirmacher P, Schemmer P. Donor liver histology–a valuable tool in graft selection. Langenbecks Arch Surg 2015;400:551–557.

[14] Salvi M, Molinaro L, Metovic J, et al. Fully automated quantitative assessment of hepatic steatosis in liver transplants. Comput Biol Med 2020;123:103836.
[15] Ho S, Kuo E, Allende D, et al. Heterogeneity of hepatic steatosis definitions and reporting of donor liver frozen sections among pathologists: a multicenter survey. Liver Transpl 2022;28:1540–1542.
[16] Neil DAH, Minervini M, Smith ML, et al. Banff consensus recommendations for steatosis assessment in donor livers. Hepatology 2022;75:1014–1025.
[17] Gambella A, Mastracci L, Caporalini C, et al. Not only a small liver - the pathologist's perspective in the pediatric liver transplant setting. Pathologica 2022;114:89–103.
[18] Patrono D, Cussa D, Sciannameo V, et al. Outcome of liver transplantation with grafts from brain-dead donors treated with dual hypothermic oxygenated machine perfusion, with particular reference to elderly donors. Am J Transpl 2022;22:1382–1395.
[19] Patrono D, De Carlis R, Gambella A, et al. Viability assessment and transplantation of fatty liver grafts using end-ischemic normothermic machine perfusion. Liver Transpl 2023;29:508–520.
[20] Kleiner DE, Brunt EM, Van Natta M, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology 2005;41:1313–1321.
[21] Xiayu X, Kyungmoo L, Li Z, et al. Stratified sampling voxel classification for segmentation of intraretinal and subretinal fluid in longitudinal clinical OCT data. IEEE Trans Med Imaging 2015;34:1616–1623.
[22] Roy M, Wang F, Vo H, et al. Deep-learning-based accurate hepatic steatosis quantification for histological assessment of liver biopsies. Lab Invest 2020;100:1367–1383.
[23] Guo X, Wang F, Teodoro G, et al. Liver steatosis segmentation with deep learning methods. Proc IEEE Int Symp Biomed Imaging 2019;2019:24–27.
[24] **Sun L**, **Marsh JN**, Matlock MK, et al. Deep learning quantification of percent steatosis in donor liver biopsy frozen sections. EBioMedicine 2020;60:103029.
[25] De Stefano N, Navarro-Tableros V, Roggio D, et al. Human liver stem cell-derived extracellular vesicles reduce injury in a model of normothermic machine perfusion of rat livers previously exposed to a prolonged warm ischemia. Transpl Int 2021;34:1607–1617.
[26] Patrono D, Roggio D, Mazzeo AT, et al. Clinical assessment of liver metabolism during hypothermic oxygenated machine perfusion using microdialysis. Artif Organs 2022;46:281–295.
[27] Salvi M, Bosco M, Molinaro L, et al. A hybrid deep learning approach for gland segmentation in prostate histopathological images. Artif Intell Med 2021;115:102076.
[28] Salvi M, Mogetta A, Gambella A, et al. Automated assessment of glomerulosclerosis and tubular atrophy using deep learning. Comput Med Imaging Graph 2021;90:101930.
[29] Salvi M, Michielli N, Molinari F. Stain Color Adaptive Normalization (SCAN) algorithm: separation and standardization of histological stains in digital pathology. Comput Methods Programs Biomed 2020;193:105506.
[30] Salvi M, Acharya UR, Molinari F, et al. The impact of pre- and post-image processing techniques on deep learning frameworks: a comprehensive review for digital pathology image analysis. Comput Biol Med 2021;128:104129.
[31] Michielli N, Caputo A, Scotto M, et al. Stain normalization in digital pathology: clinical multi-center evaluation of image quality. J Pathol Inform 2022;13:100145.
[32] Long JJ, Nijhar K, Jenkins RT, et al. Digital imaging software versus the "eyeball" method in quantifying steatosis in a liver biopsy. Liver Transpl 2023;29:268–278.
[33] Salvi M, Molinari F, Iussich S, et al. Histopathological classification of canine cutaneous round cell tumors using deep learning: a multi-center study. Front Vet Sci 2021;8:640944.
[34] Watanabe T, Shimada H, Tanaka Y. Human hepatocytes and aging: a cytophotometrical analysis in 35 sudden-death cases. Virchows Arch B Cel Pathol 1978;27:307–316.
[35] Watanabe T, Tanaka Y. Age-related alterations in the size of human hepatocytes. A study of mononuclear and binucleate cells. Virchows Arch B Cel Pathol Incl Mol Pathol 1982;39:9–20.
[36] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15:155–163.
[37] Cardillo M, Ricci A, Filippetti M, et al. Annual activity of the Italian national transplant netwok. 2022.