Research paper

Deep learning quantification of percent steatosis in donor liver biopsy frozen sections

Lulu Sun[a,#], Jon N. Marsh[a,b,#], Matthew K. Matlock[a], Ling Chen[c], Joseph P. Gaut[a], Elizabeth M. Brunt[a], S. Joshua Swamidass[a,b,*], Ta-Chiang Liu[a,d,*]

[a] Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, United States
[b] Institue for Informatics (I²), Washington University School of Medicine, St. Louis, MO, United States
[c] Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, United States
[d] Lead contact

A R T I C L E   I N F O

A B S T R A C T

Background: Pathologist evaluation of donor liver biopsies provides information for accepting or discarding potential donor livers. Due to the urgent nature of the decision process, this is regularly performed using frozen sectioning at the time of biopsy. The percent steatosis in a donor liver biopsy correlates with transplant outcome, however there is significant inter- and intra-observer variability in quantifying steatosis, compounded by frozen section artifact. We hypothesized that a deep learning model could identify and quantify steatosis in donor liver biopsies.
Methods: We developed a deep learning convolutional neural network that generates a steatosis probability map from an input whole slide image (WSI) of a hematoxylin and eosin-stained frozen section, and subsequently calculates the percent steatosis. Ninety-six WSI of frozen donor liver sections from our transplant pathology service were annotated for steatosis and used to train (*n* = 30 WSI) and test (*n* = 66 WSI) the deep learning model.
Findings: The model had good correlation and agreement with the annotation in both the training set (r of 0.88, intraclass correlation coefficient [ICC] of 0.88) and novel input test sets (*r* = 0.85 and ICC=0.85). These measurements were superior to the estimates of the on-service pathologist at the time of initial evaluation (*r* = 0.52 and ICC=0.52 for the training set, and *r* = 0.74 and ICC=0.72 for the test set).
Interpretation: Use of this deep learning algorithm could be incorporated into routine pathology workflows for fast, accurate, and reproducible donor liver evaluation.
Funding: Mid-America Transplant Society

Introduction

There is a global shortage of donor livers suitable for transplantation. In the United States alone, approximately 11,000 people are added to the wait list for liver transplant every year, yet nearly 10% of livers recovered for transplant are discarded [1]. The decision to discard may be made after histologic examination of biopsies from potential donor livers, which is essential to assess organ suitability prior to transplantation, and to provide predictive information for graft outcome [2–4]. This evaluation must be performed quickly,

since increased warm and cold ischemic time are associated with poor graft outcome [5–8]. Warm ischemic time occurs in donation after circulatory death donors, as the organ remains in the body after the blood supply has been reduced, while cold ischemic time is the time from cross-clamping of the donor liver to removal of the liver from cold storage solution [7,9].

One of the most important features to evaluate is the percentage of macrovesicular steatosis (herein referred to as steatosis). While some non-invasive methods have been proposed [10–12], the gold standard for steatosis evaluation remains liver biopsy and examination by a pathologist [13]. Using histologic assessment of steatosis, some studies have suggested that transplantation of donor livers with ≥30% macrovesicular steatosis is associated with early allograft dysfunction, with conflicting evidence regarding the definitions and role of small versus large droplet macrovesicular steatosis [3,14–21]. However, there is no agreed-upon cutoff value for discarding donor

Research in Context

Evidence before this study

We searched PubMed and Google Scholar databases for peer-reviewed, original research articles and reviews from 1990 to 2020, using combinations of search terms in three areas of interest: 1) value of hepatic steatosis estimation in transplant ("hepatic or liver transplant steatosis outcomes", "steatosis estimation frozen," "liver transplant evaluation"); 2) pathologist estimation of hepatic steatosis ("pathologist estimation hepatic or liver steatosis", "pathologist agreement steatosis," "variability pathologist steatosis"); and 3) computational assessment of steatosis ("computational assessment steatosis histology," "artificial intelligence steatosis," "computer quantification steatosis").

We found that the majority of articles agreed that high amounts of macrovesicular steatosis correlated with poor graft outcome, although the threshold for recommended organ discard varied by study and center. One review highlighted the increasing number of steatotic liver grafts due to the expanding prevalence of fatty liver disease, and how the lack of a reliable method for steatosis quantification contributes to an inconsistent ability to predict graft outcome. Multiple studies measuring intra- and inter-observer variability illustrated that there is poor agreement among pathologists in the quantification of hepatic macrovesicular steatosis. In addition, studies have demonstrated the difficulty and error rate of pathologist evaluation of steatosis in frozen sections specifically, which are used in donor liver assessment due to their rapid preparation, but are subject to substantial artifact. Previous attempts to computationally quantify steatosis in histologic images involved measurement of the fraction of the image area occupied by fat droplets, many using segmentation methods focussing on white space quantification. We found only one method that attempted to quantify steatosis in frozen section slides, and it used Oil Red O staining to highlight fatty areas.

Added value of this study

Our project directly addresses the problem of the lack of a reliable method for steatosis quantification by developing a deep learning model that can accurately and reproducibly quantify liver steatosis. In addition, it performs on digitized frozen section slides, accounting for confounding frozen section artifact. Our model performs the analysis and quantification within minutes, much faster than many whole-slide imaging deep learning models.

Implications of all the available evidence

Donor liver macrovesicular steatosis correlates with graft outcome. Accurate quantification of this steatosis by automated deep learning models could aid in predicting transplant outcome and help prevent unnecessary donor organ discard.

prepared by rapid intra-operative frozen sectioning, a process that is known to contain histologic artifacts that can mimic steatosis (water droplets, holes, etc.) [31]. Comparison to the corresponding formalin-fixed, paraffin-embedded (FFPE) sections can reveal these artifacts. Studies show that agreement between estimates of steatosis in frozen sections versus properly fixed and processed permanent sections is less than 70% [31,32]. In addition, the concordance rate of estimation of percent steatosis between nonspecialist pathologists and expert liver pathologists is only 70% [32]. Inaccurate assessment of steatosis could lead to unnecessary discard and longer transplant wait list time (if overestimated), or poor transplant outcome (if underestimated). Therefore, approaches to standardize the quantification process with rapid, accurate, and reproducible interpretation of steatosis on donor liver biopsies will enhance decision-making for transplant clinicians.

Recently, deep learning, a subset of machine learning, has revolutionized several fields, including medical image analysis, and provides a possible solution for accurate frozen section steatosis quantification [33]. In particular, deep learning convolutional neural networks (CNNs) have been shown to perform as well as human experts in solving many image recognition tasks, including those specific to anatomic pathology. For example, deep learning algorithms can be employed to automate grading of prostate biopsies, classify melanoma, recognize patterns of lung adenocarcinoma, and detect breast cancer lymph node metastases [34−39]. These deep learning CNNs are modeled after neural networks, with multiple connected layers that process the data and transmit the output to the next layer. CNNs for image analysis are trained on input images, and the models automatically learn salient features from the data alone. We therefore hypothesized that a deep learning algorithm would be able to determine steatosis on frozen sections as well as expert liver pathologists.

Here, we describe the development of a deep learning model that can accurately and reproducibly interpret percent macrovesicular steatosis on WSI of donor liver biopsy frozen sections within minutes. Our work advances the field of deep learning-based computer-aided diagnosis by the use of intraoperative frozen sections, for which there are few published studies [40,41], and by applying classification and quantification methods to a critical problem in surgical pathology and transplantation.

Materials and methods

Case selection and cohorts

H&E-stained frozen donor liver biopsies were obtained from cases evaluated between April 2015 and December 2016 by the Washington University School of Medicine Transplant Pathology service, which evaluates donor liver biopsies from multiple organ procurement organizations. During this time, all frozen sections were digitally scanned into WSI. 96 total cases from 91 deceased donors were selected to exhibit a wide range of percent steatosis values (<1% to >95%; Table 1 and Supplemental Table 1). The sample size was determined by the number of available WSI at the time of initial evaluation. Cases with poor image quality were excluded. Cases were selected based on steatosis values alone, determined by the on-service pathologist, without looking at the WSI. 5 donors were the source of 2 frozen sections each − these were treated as independent samples as they were prepared and evaluated separately. 44 of the donors were female, and 47 of the donors were male. For training and testing the deep learning algorithm, the cases were randomly divided into an annotated training set ($n = 30$), and an annotated test set ($n = 66$).

Evaluation of steatosis

Each WSI underwent 5 evaluations: the initial interpretation by the on-service pathology (OS-P) at the time of transplant evaluation, re-review by 3 pathologists, and assessment by the deep learning

livers and in some centers, livers with reportedly much greater than 30% steatosis are successfully transplanted [22−24]. Part of the problem with defining thresholds for steatosis evaluation in donor liver biopsies may be the significant variability that exists between observers, with reported correlation coefficients widely ranging from 0.55−0.98 (continuous percentages), and kappa values ranging from 0.38−0.94 (categorical grades) [25−30]. Furthermore, due to the time-sensitive nature of donor liver evaluation, pathologists currently manually identify and make visual estimates of features of biopsies

Table 1
Percentage steatosis in training and test sets, as determined by the on-service pathologist.

% steatosis	Training set		Test set	
	N	*% of set*	N	*% of set*
<**5%**	7	*23*	32	*48*
5−33%	10	*33*	15	*2*
34−66%	9	*30*	12	*18*
>**66%**	4	*13*	7	*11*
Total	30	*100*	66	*100*

model. Frozen sections of the donor liver biopsies were initially evaluated by the 24hr on-service pathologist, as part of routine clinical practice, and percent steatosis was recorded as part of the initial report. All on-service pathologists were surgical pathologists with or without liver pathology expertise. The frozen WSI were re-reviewed by two expert liver pathologists and a pathology trainee, who also reviewed permanent sections of the frozen liver biopsy remnants, as an "optimized" pathologist assessment. The purpose of reviewing the FFPE permanent sections was to account for frozen section artifact. These assessments were recorded as percentages in 5% increments. If a range of steatosis was reported by any pathologist, the average of the range was used in comparative analyses. Re-reviewers were blinded to the initial on-service pathologist percentage, as well as to other evaluator assessments. The average of the percent steatosis determined by the 3 pathologists after evaluation of both frozen section WSI and permanent sections was recorded as the average re-reviewer assessment.

Whole slide imaging

Frozen sections were scanned at 20X using an Aperio Scanscope CS2 scanner (Leica Biosystems, Wetzlar, Germany) and stored in SVS format, then converted to TIFF format at full resolution (0.495 μm/pixel).

Annotation

Areas of steatosis were manually annotated in series by two pathologists (initial annotation by a pathology trainee, followed by review by an expert liver pathologist). Annotation was performed by outlining and labeling all areas of macrovesicular steatosis (using elliptically-shaped and free-form masks) in each WSI using an in-house plugin written for Fiji [41,42], generating pixel-wise label masks of steatotic regions at the same resolution as the parent WSI. To account for frozen section artifact, artifactual white areas that were present in the frozen section WSI, but not in the FFPE permanent slides, were not labeled. All cells with macrovesicular steatosis were annotated (including large droplets with large vacuoles in the cytoplasm with eccentric nuclear displacement, and small droplets with few/discrete fat vacuoles without nuclear displacement) [21,43]. Microvesicular steatosis was defined as the presence of non-zonal, contiguous patches of foamy hepatocytes and was not included in the annotation, as current evidence does not support its clinical utility in donor liver evaluation [44,45]. Adjacent nuclei and cytoplasm of steatotic hepatocytes were included in the annotation masks to provide additional context for the deep learning model. For some larger images with higher percentages of steatosis, representative bounding box areas were annotated instead of the whole slide. This enabled training on more images to account for differences in slide cutting and staining variables. Bounding box areas were selected to include at least 2 portal tracts and intervening areas, as well as zones with frozen artifact. Areas with tissue folding were not included. For the training set of 30 WSI, 19 had bounding boxes and 11 were completely annotated. For the test set, 33/66 had bounding boxes.

10,653 regions (containing one or more steatotic cells) were labeled and used to generate 75,079,680 target pixels from 30 WSI for training the CNN. For the test set, 12,643 annotation regions comprising 113,626,112 pixels were labeled. Because the CNN output was downsampled 32x relative to the input images, the annotation masks were similarly downsampled to maintain pixelwise registration. Downsampling was accomplished by gridding the full-size annotation mask into 32 × 32 pixel patches and assigning to the corresponding target pixel the fraction of the original patch area covered by the mask.

CNN architecture

Utilizing a pre-trained network as a starting point (that is, transfer learning) has been shown to be a useful technique to reduce the need for large training datasets and decrease training time for deep CNNs [46,47]. In transfer learning, the hundreds of millions of network weights in the deep layers of the model (which encode generic image features) are kept constant, while the weights of the uppermost classification layers are allowed to change in the process of training in the new domain. Thus, a much smaller set of network weights needs to be optimized for model convergence, which is therefore feasible with fewer computational resources and input images. We trained a fully convolutional model based on VGG16, a CNN originally trained (using a vast amount of computational time and resources) on a large database of millions of digital photographs to label image content [48]. No image preprocessing was performed prior to training or testing, other than that prescribed in the VGG16 schema, i.e., subtraction by a precomputed fixed RGB value [48]. Starting with the pre-trained VGG16 CNN with weights frozen below the bottleneck, we replaced the final fully-connected layers with the following sequence of convolutional layers: two 1 × 1 convolution layers (256 and 128 nodes, respectively), followed by a 64-node 3 × 3 convolution layer and another 64-node 5 × 5 convolutional layer (Fig. 3). All convolutional layers used ReLU activation, with no input padding before filtering. Output was fed to a 2-node layer with softmax activation for classification into steatotic and non-steatotic categories.

Storing the activations of a fully convolutional network over an entire WSI is not feasible due to excessive memory requirements, therefore we adopted a sampling approach to training the model. For images used in training and validation, 832 × 832-pixel partially-overlapping image patches (stride=448) having at least 5% non-background area were extracted and presented to the model by weighted sampling. Classes were weighted in a ratio of 4:1 for steatosis:background categories to account for class imbalance.

The task of the CNN was to assign labels to individual pixels in the WSI to match the provided annotations. Because the model's output was downsampled by a factor of 32 relative to the original image, the CNN was trained against a similarly downsampled annotation map, by computing the ratio of the number of annotation mask pixels to total pixel number within each cell of a 26 × 26 pixel grid in the full-size annotation mask.

The fully convolutional CNN was trained in several phases in which successively deeper convolution block weights were unfrozen and allowed to be modified in training. In the first phase, the pre-trained VGG16 layers up to the bottleneck were frozen and the top layers were trained for 15 epochs using the Adam optimizer with a learning rate of 1e-4 [49]. In the next phase, the weights from the last convolutional block (composed of 3 convolutional layers) before the bottleneck were then unfrozen, and the model trained for 5 epochs using stochastic gradient descent (SGD) with a learning rate of 1e-5. This was repeated twice more, each time unfreezing the next deeper convolutional block for training. The final phase entailed unfreezing all model weights and training for 15 epochs. Training in phases as described, with low learning rates and SGD optimizer, was necessary to prevent overfitting and training loss divergence as the model's

entropic capacity increased at each phase. Prior exploration indicated this set of training parameters prevented overfitting while yielding satisfactory categorical accuracy.

The model was trained in a hold-one-out cross-validation scheme. For each fold, the corresponding trained model was applied to the withheld WSI by sampling image patches in a raster pattern from a sliding window (832 $\times$ 832-pixels) with a stride of 448 pixels, yielding a 26 $\times$ 26-pixel categorical probability map associated with each respective image patch. These patches were stitched together to yield a categorical probability map of the complete WSI, downsampled by a factor of 32.

All WSI image analyses were performed using an Nvidia Tesla K20X graphics processing unit (GPU).

Evaluation of pixel predictions

Evaluation of model performance was computed at several stages of the modeling process. An annotation-based evaluation of steatosis fraction was derived from CNN-generated pixel maps by computing the sum of the steatosis probability from all pixels and dividing by the area of the downsampled image associated with tissue (determined from intensity thresholding). In the test sets, model estimates were computed for both the bounding box area as well as the whole slide. This value was compared to the corresponding value derived from the annotation maps to compute Pearson's correlation coefficient, and intraclass correlation coefficient (ICC).

Statistics

Correlations and agreement measures between pathologist interpretation of frozen section and FFPE histology, pathologist annotation and CNN prediction, as well as CNN prediction and pathologist interpretation of FFPE histology were performed by using the Pearson correlation coefficient and ICC, respectively [50]. The intraclass correlation coefficient is a statistical measure of rater agreement on the same subjects, with values less than 0.5, 0.5 to 0.75, 0.75 to 0.9, and >0.90 considered as poor, moderate, good and excellent agreement, respectively [50,51]. The mean difference between evaluators was compared using a linear mixed effects model. Statistical analyses were performed using R, SPSS Statistics (IBM, Armonk, New York) and GraphPad Prism v.8 (GraphPad Software, San Diego, California). A p value of <0.05 was considered significant.

Materials availability statement

This study did not generate new unique reagents.

Ethics statement

The study was approved by the Institutional Review Board of Washington University School of Medicine with a waiver of consent (IRB number 201703119).

Results

Pathologist quantification of steatosis in donor liver biopsies may be confounded by frozen section artifact and shows high inter-observer variability

We retrieved 96 WSI with frozen and FFPE sections, corresponding to 91 individual donors, evaluated by our institutional transplant pathology service during a 20 month period. These cases exhibited a wide range of macrovesicular steatosis. They were evaluated by 15 different on-service pathologists (OS-P), 6 of whom were gastrointestinal/liver specialty pathologists. Each OS-P evaluated from 1 to 34 frozen sections, with an average of 6.6 cases and a median of 3 cases. To determine whether frozen section artifact may have affected the OS-P estimates of steatosis, the slides were re-evaluated by 3 pathologists who reviewed both the frozen section (available to the OS-P at the time of their assessment), as well as the FFPE slides of the frozen section remnants (not available to the OS-P). In addition, the 3 pathologists had unconstrained time to examine the slides, and performed their evaluation during their normal working hours, whereas the OS-P assessed the frozen sections using current standard of care, evaluating frozen section WSI at any time of the day or night, with time pressure imposed by the urgent nature of these specimens. The purpose of the re-review was to determine the discrepancy between the on-service pathologist, faced with time constraints, possible fatigue, and frozen section artifacts, and an "optimized" pathologist assessment.

Inspection of the data demonstrated a few notable patterns, including at the 30% threshold associated with poor graft survival [17,20]. Of the 39 WSI in which the average re-reviewer steatosis percentage was <5% (almost no steatosis), the OS-P overestimated steatosis ($\geq$ 5%) in 19 cases (49%) (Fig. 1a and Supplemental Table 1). The OS-P reported steatosis $\geq$30% in 2 of these cases (5%) (Fig. 1a and Supplemental Table 1). Manual re-review of these cases suggested that processing artifact seen in the frozen sections, but not the FFPE sections, may have resulted in the higher OS-P estimate (example shown in Fig. 2). In addition, for the 31 cases in which the average re-reviewer steatosis percentage was >30%, the OS-P underestimated steatosis (<30%) in 14 cases (45%) (Fig. 1a and Supplemental Table 1). Manual re-review of these cases did not suggest any evident reasons for the discrepancies.
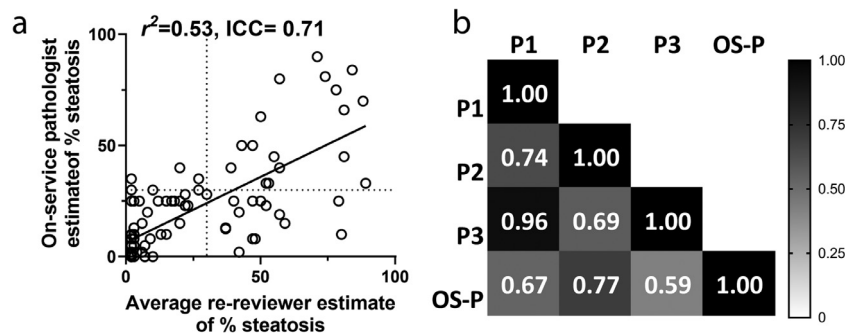

Fig. 1. Pathologist quantification of steatosis in donor liver biopsies shows high inter-observer variability. a) Plot with linear regression of estimates of % steatosis by on-service pathologists on WSI of frozen donor liver biopsies versus re-reviewer estimates (the average of estimates from 3 pathologists who had reviewed the corresponding FFPE slides). Each dot represents the percent steatosis estimated from one WSI of a donor liver biopsy. *N* = 96 WSI. The dotted lines indicate the 30% threshold reported in the literature for organ discard. r^2 value = squared Pearson correlation coefficient. ICC = intraclass correlation coefficient. b) Heat map of ICC between each pair of pathologists for steatosis quantification for 96 WSI. P1, P2, and P3 were re-reviewers, and OS-P was the on-service pathologist. ICC between all pathologists was 0.74 (95% CI 0.67–0.80).
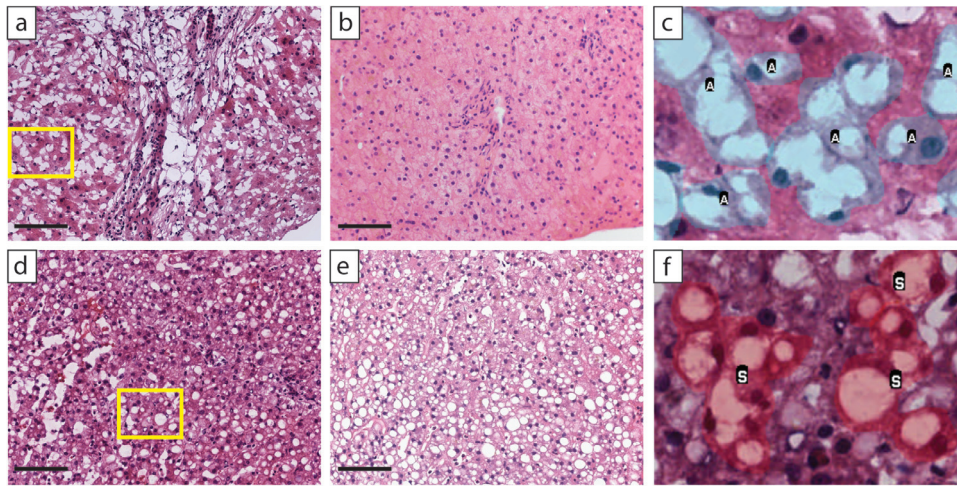
Fig. 2. Frozen section preparation of liver biopsies result in histologic artifacts that may be confused with steatosis. Scanned whole slide images of a) an H&E-stained frozen section of a liver biopsy with <5% steatosis demonstrating preparation artifact, and b) the corresponding H&E-stained formalin fixed and paraffin-embedded frozen section remnant. Black scale bar = 100 μm. c) Area in yellow box in A with artifactual white spaces mimicking steatosis highlighted in blue and annotated "A." Scanned whole slide images of d) an H&E-stained frozen section of a liver biopsy with 80% steatosis, and e) the corresponding H&E-stained formalin fixed and paraffin-embedded frozen section remnant. Black scale bar = 100 μm. f) Area in yellow box in A with steatosis highlighted in red and annotated "S.".

To investigate interobserver agreement, the intraclass correlation coefficient (ICC) and Pearson correlation coefficient of percent steatosis values was calculated between each pair of pathologists, including the 3 re-reviewers (P1, P2, and P3), and OS-Ps (Fig. 1b). Pearson correlation coefficients between each pair ranged from 0.63 (P3 vs OS-P) to 0.97 (P1 vs P3). ICC between each pair ranged from 0.59 to 0.96. Although the degree of correlation between the 3 pathologists with optimized assessment were, on average, higher than those for the OS-Ps, variation still existed between individuals. The observed values are comparable to those previously reported in the literature regarding frozen vs. FFPE liver sections, and between pathologists [25,31,32].

Deep learning algorithm predictions of steatosis correlate with annotations

To tackle the task of reproducibly quantifying donor liver biopsy steatosis, taking into account frozen section artifact, a deep learning algorithm/model was designed. The model was constructed using the pre-trained VGG16 architecture, which was trained on millions of high-resolution internet images [48]. The use of a pre-trained network as a starting point, referred to as transfer learning, is commonly used to reduce the need for large training datasets and decrease training time[47]. The VGG16 convolutional layers and pretrained weights were utilized as an initial framework, with the original fully-connected layers replaced by a set of fully-convolutional upper layers (Fig. 3). The model received WSI as input, processed the image via layers of computations, and produced an output consisting of steatosis probabilities mapped to the input image. It then calculated the percent steatosis from these probabilities. The model first had to be trained to "recognize" areas of steatosis on a set of training images. It was then tested for its performance on the training images again, as well as de novo test images.

Our initially retrieved 96 WSI were divided into a training set (n = 30 WSI) and a test set (n = 66 WSI), corresponding to 91 liver donors (Table 1). As "ground truth," for the model to recognize steatosis, instead of visual estimation of percent steatosis, the WSI were manually annotated by 2 pathologists for steatotic cells and steatotic areas. All cells with macrovesicular steatosis were annotated. Inclusion of the surrounding cell in the annotation of fat droplets ensured that this information was presented to the model as meaningful input for positive steatosis labeling, as opposed to voids arising from

random freezing artifact. Annotations were performed on the frozen section WSI serially by two pathologists who had also seen the FFPE permanent slides, to account for frozen artifact. This yielded 10,653 annotation areas comprising 75,079,680 pixel training targets.

The 30 annotated training images were presented to the model in a hold-one-out cross validation scheme, in which the CNN was trained using all images but one, which was then used to validate the resulting model prediction. This was done in turn for all 30 training images. This cross validation decreases overfitting and enables checking of the model performance during the training phase. Probability maps for the holdout slides indicating areas of steatosis were generated to compare model performance with manual annotations and H&E slide images (Fig. 4). The model was able to predict areas of steatosis, while ignoring other white areas on the slide, such as dilated sinusoids and tears in the tissue.

A percent steatosis value was calculated from model-generated pixel maps by summation of the steatosis probability from all pixels and dividing by total area of tissue. At the WSI level, the model prediction of percent steatosis for each WSI had good correlation and agreement with the annotation % steatosis, with an r of 0.88 (95% CI 0.75−0.94) and ICC of 0.88 (95% CI 0.75−0.94) (Fig. 5a). After training, the model was able to rapidly analyze the WSI and calculate percent steatosis within 5−7 min, using a 6 year-old GPU (Nvidia Tesla K20X).

Once the training regimen and model architecture had been evaluated in the cross-validation scheme, the model was trained anew using all 30 slides in a single group. The network weights were then kept constant, and the model was subsequently tested with the 66 annotated test WSI, which it had not previously received as input. The model prediction of percent steatosis for the entire WSI again correlated well with the percent steatosis calculated from the annotation, with an r of 0.85 (95% CI 0.77−0.91) and an ICC of 0.85 (95% CI 0.76−0.90) (Fig. 5b), indicating the architecture's fitness for inference on new data. The model, once trained, was completely deterministic, yielding an identical result every time for the same input WSI, and did not require additional annotation by the pathologist.

Deep learning model has superior performance in steatosis estimation compared to pathologists

The average re-reviewer and OS-P steatosis estimates were compared to the percent steatosis calculated from the annotations. The
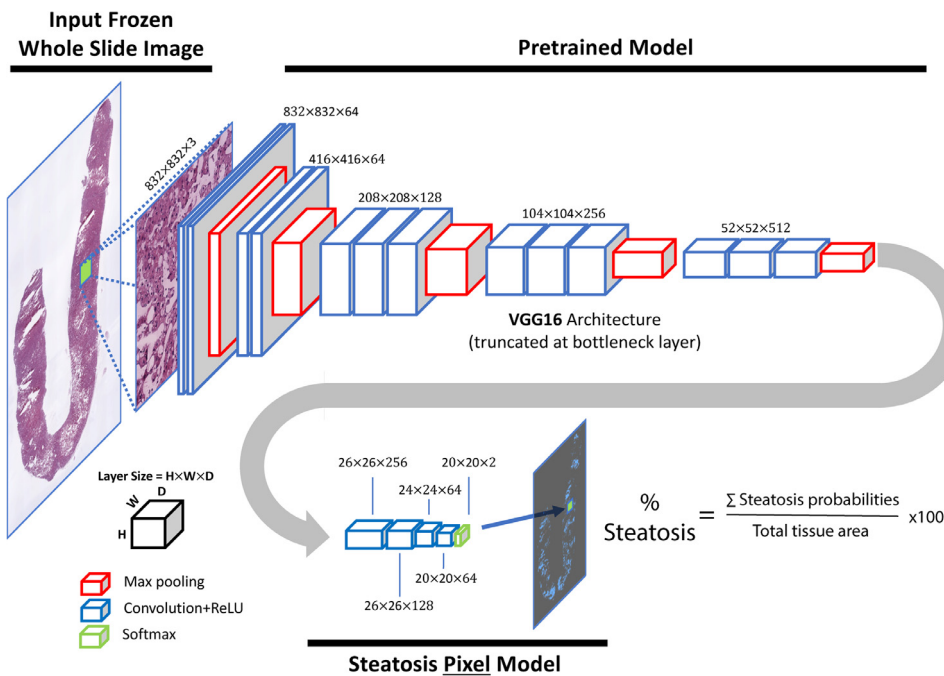
Fig. 3. Schematic of fully convolutional neural network and steatosis fraction model. Each layer of the convolutional neural network is displayed as a 3-D box (red = max pooling layer, blue = convolution+ReLU layer, green = softmax activation layer). A WSI was used as input and divided into partially-overlapping square image patches 832 pixels wide, which were processed individually by the deep learning model. Output patches were stitched together to form steatosis probability maps. The VGG16 CNN base architecture was employed, with weights frozen below the bottleneck (upper half of architecture). The fully-connected classification layers were replaced with 4 convolution+ReLu layers, with output fed to a 2-node layer with softmax activation to classify each pixel into steatotic and non-steatotic categories. The percent steatosis for each WSI was calculated from the model-generated pixel map by summation of the steatosis probabilities from all pixels divided by the total tissue area.

OS-P, representing the current standard of care, had the lowest correlation and agreement to the annotations, with an r of 0.52 (95% CI 0.20−0.74) and ICC of 0.52 (95% CI 0.20−0.74) for the training set, and an r of 0.74 (95% CI 0.61−0.83) and ICC of 0.72 (95% CI 0.58−0.82) for the test set (Fig. 5). The average of the re-reviewer

Fig. 4. Deep learning algorithm predictions of steatosis correlate with annotations. Comparison of H&E-stained donor liver wedge biopsy (left column), mask of annotated steatotic areas (black) within bounding boxes (green) (middle column), and pixel-based model predictions of areas of steatosis across the WSI (grayscale probability map, with 100% probability of steatosis in black, and 0% probability in white) (right column). Black bar = 1 mm. Higher magnification insets shown in second row; black bar = 0.5 mm.

estimates had an r of 0.78 (95% CI 0.58−0.89) and ICC of 0.75 (95% CI 0.55−0.87) for the training set, and an r of 0.87 (95% CI 0.79−0.92) and ICC of 0.79 (95% CI 0.68−0.87) for the test set, compared to the annotations. Overall, the deep learning model had better agreement with the percent steatosis calculated from annotations than the pathologists.

The difference between the percent steatosis by annotation and by the pathologist or model was calculated for each WSI. The means of these differences were compared using a linear mixed effects model. There was no significant difference across the means for the training set ($p = 0.19$). For the test set, there was a significant pairwise difference between the mean difference for the model compared to the OS-P ($p = 0.001$) and for the model compared to the average re-reviewer ($p=<0.0001$). There was no significant pairwise difference between the OS-P and the average re-reviewer. These analyses show that the model had significantly superior agreement with the annotations compared to the pathologists.

Deep learning model performance at the 30% threshold for steatosis

Some institutions report using a 30% steatosis cutoff for rejection of donor livers, although this threshold varies by center. If the manual annotation is used as the gold standard, the sensitivity of the deep learning algorithm to detect steatosis >30% in the 96 annotated WSI was 15/21=71.4%, with a specificity of 73/75=97.3%. The positive predictive value (PPV) was 15/17=88.2%, and the negative predictive value (NPV) was 73/79=92.4%. In contrast, the sensitivity and specificity of the OS-P to detect steatosis >30% was 17/21=80.9% and 64/75=85.3%, respectively. The PPV and NPV of the OS-P were 17/28=60.7% and 64/68=94.1%, respectively. While the deep learning model had lower sensitivity than the OS-P for identifying a case with >30% steatosis, it had much higher specificity. Overall, the deep learning model incorrectly classified 2 WSI as >30% compared to 11 WSI by the OS-P. Functionally, if a 30% steatosis cutoff were used as a threshold for organ discard, the deep learning model would result in
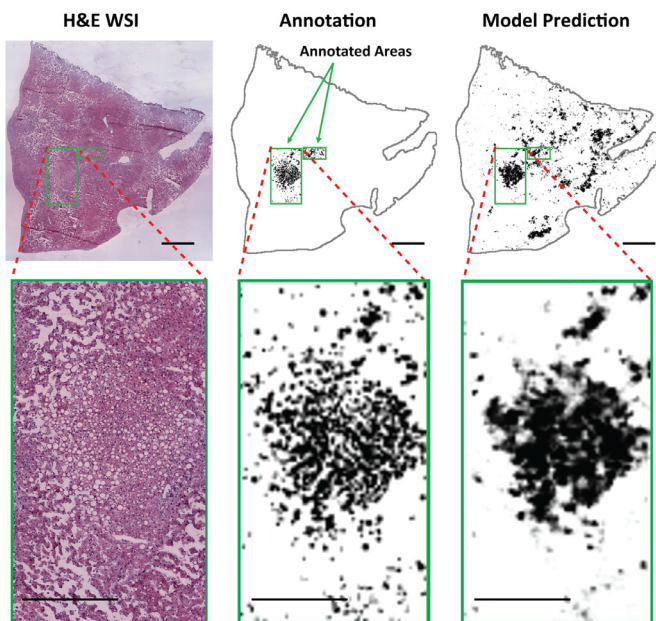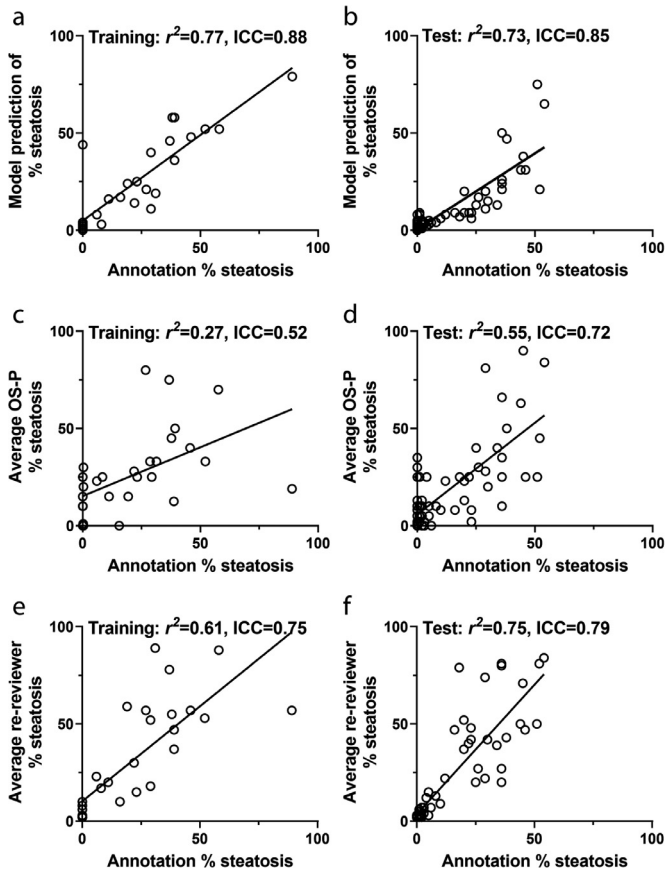
Fig. 5. Deep learning model performs superior to current standard of care in assessing percent steatosis. a) Plot with linear regression of cross-validated model prediction of percent steatosis compared to percent steatosis calculated from annotated area in the training set. Each dot represents the percent steatosis for one WSI (n = 30 WSI). b) Plot with linear regression of model prediction of percent steatosis compared to percent steatosis calculated from annotated area in the test set (n = 66 WSI). c) Plot with linear regression of OS-P prediction of percent steatosis compared to percent steatosis calculated from annotated area in the training set (n = 30 WSI). d) Plot with linear regression of OS-P prediction of percent steatosis compared to percent steatosis calculated from annotated area in the test set (n = 66 WSI). e) Plot with linear regression of average re-reviewer prediction of percent steatosis compared to percent steatosis calculated from annotated area in the training set (n = 30 WSI). f) Plot with linear regression of average re-reviewer prediction of percent steatosis compared to percent steatosis calculated from annotated area in the test set (n = 66 WSI). r^2 value = squared Pearson correlation coefficient. ICC = intraclass correlation coefficient.

9/96=9% fewer organs unnecessarily discarded in this cohort, compared to the current standard of care.

Discussion

We created a deep learning algorithm that is capable of identifying and quantifying macrovesicular steatosis in WSI of donor liver biopsy frozen sections. We validated our model on *de novo* whole slide images, with performance superior to those of the current standard of care on-service pathologist evaluations and expert pathologists in optimized evaluation conditions.

Our results were consistent with previously published reports demonstrating inter-observer variability in quantifying percent macrovesicular steatosis, and in differences between frozen section and permanent section steatosis percentages [25,31,32]. We noted specific instances in which pathologist overestimation on frozen sectioning were most likely attributed to frozen preparation artifact. The reasons for pathologist underestimation on frozen sectioning were less evident. Some possibilities include an attempt to account for frozen section artifact, misclassification of small droplet

macrosteatosis as microvesicular steatosis, or simple inaccuracy in visual estimation.

To attempt to overcome inter-observer variability, there have been multiple prior efforts to automate quantification of liver steatosis [52−57]. These methods usually measure steatosis by the fraction/percent of the image area that is occupied by fat droplets. Many use segmentation methods that focus on white space quantification. These previous models cannot account for frozen section artifact, while our model was designed to perform on frozen sections. We trained the algorithm on annotations that included only areas of steatosis, not frozen artifactual white areas, which were verified during annotation using the corresponding FFPE permanent sections. We used annotation masks that included information surrounding the white space, such as nuclei and cytoplasm, reasoning that this contextual information could provide the deep learning model with additional cues for classification. While this study did not identify the pertinent features that the algorithm used to classify an area as steatotic, future studies could attempt to discover these features through identification of activated regions of the image after each convolution [58]. Notably, our model removes intra-observer variability, as it is fully reproducible, producing the same percent steatosis for a given WSI. We speculate that being able to accurately and precisely assess percent steatosis with our deep learning model could allow for the determination of clinically useful steatotic thresholds for predicting graft outcome, instead of the contested 30% cutoff. To accomplish this, our model could be tested on larger datasets with WSI and clinical outcomes. Further models combining this unbiased measurement of steatosis with additional clinical factors and risk scores could facilitate optimal allocation of the limited donor livers [59,60].

Our method directly modeled a pixel-based prediction of steatosis to manual annotations, which are arguably a more objective measure of steatosis than pathologist visual estimates. However, manual annotation of images for training and testing is time and labor-intensive. The use of representative bounding boxes for some WSI with high amounts of steatosis, due to the labor intensiveness of annotating these slides, could also have affected the results. For future training and testing, computer-assisted semi-automated annotation of the entire slide could be useful.

We designed our deep learning model as a fully convolutional network that could rapidly process WSIs within 5−7 min, producing a steatosis probability map registered pixel-wise to the input image, and a resultant percent steatosis quantification. Because of memory constraints, current GPU hardware cannot process an entire WSI (typically 500 MB in size), requiring that an input image be subdivided into "patches," which are processed individually and stitched together upon completion. In the naive CNN architecture, each input patch is then assigned a label. We adapted this architecture by associating each input patch with an output patch, such that every output pixel was associated with a label, rather than an entire patch. Thus, although our number of WSI was relatively small, the amount of input data was high, with greater than 10,000 annotation areas yielding nearly 4 million pixel training targets, facilitating accurate training and prediction. Naïve CNN implementations that associate an entire patch with a single label require far finer input image sampling for equivalent output resolution, and can take far longer to produce a result [41]. These technical innovations could be employed in other histologic image analysis applications.

Our study was limited by the use of slides and images collected at a single medical center, so the model performance may be weaker on external datasets. The expansion of our dataset including the addition of WSI from other institutions and pathology collections in our training and testing sets will be pursued as future research. It would also be interesting to investigate whether annotation of non-steatotic regions could be utilized as additional input to further refine our deep learning model.

The implementation of our model in a clinical setting would be in an automated platform for quantification of steatosis in scanned donor liver biopsy frozen sections, followed by pathologist verification and quality assurance. Many pathology centers already routinely use digital slide scanning for frozen intraoperative consultations and donor biopsy evaluation [61–63]. The deep learning model could be hosted on an online application as a cloud computing service, such that the user could upload the scanned WSI, run the assessment, and receive the result within minutes. Pathologist evaluation of tissue would still be necessary, both for quality control as well as for interpretation of other histologic features, including fibrosis, necrosis, and inflammation.

As the use of digital pathology and WSI expands, algorithms to improve and facilitate pathologists' diagnoses can be developed and implemented. Computer-aided diagnosis for intraoperative consultations using frozen sectioning would be a highly useful application, due to challenging frozen artifacts and the need for rapid evaluation. Our deep learning model performed more accurately compared to current standard of care on-call pathologists, and has the potential to change the paradigm of donor liver examination. This model could also be used in non-frozen section settings, as quantification of steatosis is important in multiple other liver diseases. Our work further advances this burgeoning field of histopathologic image analysis using deep learning algorithms.

Funding sources

Declaration of interests

Besides the research grant described in the funding sources, the authors declare no financial conflicts of interest relevant to this work. Outside of this work, EMB reports financial relationships with Pfizer Ad Board, Histoindex, Cymabay, Intercept, NGM, and Alynam-Regeneron.

Author contributions

Conceptualization, LS, JNM, JPG, SJS, and TCL; Methodology and Software, JNM, MKM, and SJS; Investigation, LS, TCL, EMB, and JNM; Formal Analysis, LS, JNM, and LC; Writing – Original Draft, LS and JNM; Writing – Review and Editing, JPG, SJS, EMB, and TCL; Visualization, LS and JNM; Supervision, SJS and TCL; Funding Acquisition, JPG, SJS and TCL.

Data sharing

Data collected for the study will not be shared.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ebiom.2020.103029.

References

[1] Kim WR, Lake JR, Smith JM, Schladt DP, Skeans MA, Harper AM, et al. OPTN/SRTR 2016 Annual Data Report: liver. Am J Transplant 2018;18:172–253.

[2] Flechtenmacher C, Schirmacher P, Schemmer P. Donor liver histology—A valuable tool in graft selection. Langenbeck's Arch Surg 2015;400:551–7.

[3] Melin C, Miick R, Young NA, Ortiz J, Balasubramanian M. Approach to Intraoperative Consultation for Donor Liver Biopsies. Arch Pathol Lab Med 2013:1375858.

[4] Markin RS, Wisecarver JL, Radio SJ, Stratta RJ, Langnas AN, Hirst K, et al. Frozen section evaluation of donor livers before transplantation. Transplantation 1993;56:1403–9.

[5] Coffey JC, Wanis KN, Monbaliu D, Gilbo N, Selzner M, Vachharajani N, et al. The influence of functional warm ischemia time on DCD liver transplant recipients' outcomes. Clin Transplant 2017;31:e13068.

[6] Blok JJ, Detry O, Putter H, Rogiers X, Porte RJ, van Hoek B, et al. Longterm results of liver transplantation from donation after circulatory death. Liver Transplant 2016;22:1107–14.

[7] Stahl JE, Kreke JE, Malek FAA, Schaefer AJ, Vacanti J. Consequences of Cold-Ischemia Time on Primary Nonfunction and Patient and Graft Survival in Liver Transplantation: a Meta-Analysis. Timmer A, editor. PLoS ONE 2008;3:e2468.

[8] Mathur AK, Heimbach J, Steffick DE, Sonnenday CJ, Goodrich NP, Merion RM. Donation after Cardiac Death Liver Transplantation: predictors of Outcome. Am J Transplant 2010;10:2512–9.

[9] Halazun KJ, Al-Mukhtar A, Aldouri A, Willis S, Ahmad N. Warm Ischemia in Transplantation: search for a Consensus Definition. Transplant Proc 2007;39:1329–31.

[10] Golse N, Cosse C, Allard MA, Laurenzi A, Tedeschi M, Guglielmo N, et al. Evaluation of a micro-spectrometer for the real-time assessment of liver graft with mild-to-moderate macrosteatosis: a proof of concept study. J Hepatol 2019;70:423–30.

[11] Cesaretti M, Poté N, Cauchy F, Dondero F, Dokmak S, Sepulveda A, et al. Noninvasive assessment of liver steatosis in deceased donors: a pilot study. Liver Transplant 2018;24:551–6.

[12] Swelam A, Adam R, Lauka L, Basilio Rodrigues L, Elgarf S, Sebagh M, et al. A Model to Predict Significant Macrosteatosis in Hepatic Grafts. World J Surg 2020;44: 1270–6.

[13] Cesaretti M, Addeo P, Schiavo L, Anty R, Iannelli A. Assessment of Liver Graft Steatosis: where Do We Stand? Liver Transplant 2019;25:500–9.

[14] Fernández-Merino F, Nuño-Garza J, López-Hervás P, López-Buenadicha A, Moreno-Caparrós A, Quijano-Collazo Y, et al. Impact of donor, recipient, and graft features on the development of primary dysfunction in liver transplants. Transplant Proc 2003;35:1793–4.

[15] Ploeg RJ, D'Alessandro AM, Knechtle SJ, Stegall MD, Pirsch JD, Hoffmann RM, et al. Risk factors for primary dysfunction after liver transplantation—a multivariate analysis. Transplantation 1993;55:807–13.

[16] Verran D, Kusyk T, Painter D, Fisher J, Koorey D, Strasser S, et al. Clinical experience gained from the use of 120 steatotic donor livers for orthotopic liver transplantation. Liver Transplant 2003;9:500–5.

[17] Spitzer AL, Lao OB, Dick AAS, Bakthavatsalam R, Halldorson JB, Yeh MM, et al. The biopsied donor liver: incorporating macrosteatosis into high-risk donor assessment. Liver Transplant 2010;16:874–84.

[18] Choi W-T, Jen K-Y, Wang D, Tavakol M, Roberts JP, Gill RM. Donor Liver Small Droplet Macrovesicular Steatosis Is Associated With Increased Risk for Recipient Allograft Rejection. Am J Surg Pathol 2017;41:365–73.

[19] Briceño J, Ciria R, Pleguezuelo M, de la Mata M, Muntané J, Naranjo Á, et al. Impact of donor graft steatosis on overall outcome and viral recurrence after liver transplantation for hepatitis C virus cirrhosis. Liver Transplant 2009;15:37–48.

[20] de Graaf EL, Kench J, Dilworth P, Shackel NA, Strasser SI, Joseph D, et al. Grade of deceased donor liver macrovesicular steatosis impacts graft and recipient outcomes more than the Donor Risk Index. J Gastroenterol Hepatol 2012;27:540–6.

[21] Ferri F, Lai Q, Molinaro A, Poli E, Parlati L, Lattanzi B, et al. Donor Small-Droplet Macrovesicular Steatosis Affects Liver Transplant Outcome in HCV-Negative Recipients. Can J Gastroenterol Hepatol 2019:2019.

[22] McCormack L, Petrowsky H, Jochum W, Mullhaupt B, Weber M, Clavien P-A. Use of severely steatotic grafts in liver transplantation: a matched case-control study. Ann Surg 2007;246:940–6 discussion 946-8.

[23] Afonso R, Saad W, Parra O, Leitão R, Ferraz-Neto B. Impact of steatotic grafts on initial function and prognosis after liver transplantation. Transplant Proc 2004;36:909–11.

[24] McCormack L, Dutkowski P, El-Badry AM, Clavien PA. Liver transplantation using fatty livers: always feasible? Vol. 54. J Hepatology Elsevier B.V. 2011:1055–62.

[25] El-Badry AM, Breitenstein S, Jochum W, Washington K, Paradis V, Rubbia-Brandt L, et al. Assessment of Hepatic Steatosis by Expert Pathologists. Ann Surg 2009;250:691–7.

[26] D'Alessandro E, Calabrese F, Gringeri E, Valente M. Frozen-Section Diagnosis in Donor Livers: error Rate Estimation of Steatosis Degree. Transplant Proc 2010;42:2226–8.

[27] Pournik O, Alavian SM, Ghalichi L, Seifizarei B, Mehrnoush L, Aslani A, et al. Interobserver and Intra-observer Agreement in Pathological Evaluation of Non-alcoholic Fatty Liver Disease Suspected Liver Biopsies. Hepat Mon 2014;14:e15167.

[28] Jung ES, Lee K, Yu E, Kang YK, Cho M-Y, Kim JM, et al. Interobserver Agreement on Pathologic Features of Liver Biopsy Tissue in Patients with Nonalcoholic Fatty Liver Disease. J Pathol Transl Med 2016;50:190–6.

[29] Lee MJ, Bagci P, Kong J, Vos MB, Sharma P, Kalb B, et al. Liver steatosis assessment: correlations among pathology, radiology, clinical data and automated image analysis software. Pathol - Res Pract 2013;209:371–9.

[30] Biesterfeld S, Knapp J, Bittinger F, Götte H, Schramm M, Otto G. Frozen section diagnosis in donor liver biopsies: observer variation of semiquantitative and quantitative steatosis assessment. Virchows Arch 2012;461:177–83.

[31] Heller B, Peters S. Assessment of liver transplant donor biopsies for steatosis using frozen section: accuracy and possible impact on transplantation. J Clin Med Res 2011;3:191–4.

[32] Lo IJ, Lefkowitch JH, Feirt N, Alkofer B, Kin C, Samstein B, et al. Utility of liver allograft biopsy obtained at procurement. Liver Transplant 2008;14:639–46.

[33] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A Survey on Deep Learning in Medical Image Analysis. Med Image Anal 2017;42:60–88.

[34] Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol 2020;21:233–41.

[35] Hekler A, Utikal JS, Enk AH, Berking C, Klode J, Schadendorf D, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. Eur J Cancer 2019;115:79–83.

[36] Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Sci Rep 2019;9:3358.

[37] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non−small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559–67.

[38] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;1.

[39] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 2017;318:2199.

[40] Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. Comput Struct Biotechnol J 2018;16:34–42.

[41] Marsh JN, Matlock MK, Kudose S, Liu T-C, Stappenbeck TS, Gaut JP, et al. Deep Learning Global Glomerulosclerosis in Transplant Kidney Frozen Sections. IEEE Trans Med Imaging 2018;37:2718–28.

[42] Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: An open-source platform for biological-image analysis, 9. Nature Methods Nature Publishing Group; 2012. p. 676–82.

[43] Brunt EM. Pathology of fatty liver disease. Mod Pathol 2007;20:40–8.

[44] Tandra S, Yeh MM, Brunt EM, Vuppalanchi R, Cummings OW, Ünalp-Arida A, et al. Presence and significance of microvesicular steatosis in nonalcoholic fatty liver disease. J Hepatol 2011;55:654–9.

[45] Fishbein TM, Fiel MI, Emre S, Cubukcu O, Guy SR, Schwartz ME, et al. Use of livers with microvesicular fat safely expands the donor pool. Transplantation 1997;64:248–51.

[46] Lu J, Behbood V, Hao P, Zuo H, Xue S, Zhang G. Transfer learning using computational intelligence: a survey. Knowledge-Based Syst 2015;80:14–23.

[47] Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging 2016;35:1285–98.

[48] Simonyan K., Zisserman A.Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:14091556[csCV].

[49] Kingma D.P., Ba J. Adam: a Method for Stochastic Optimization. arXiv:14126980 [csLG].

[50] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.

[51] Liu J, Tang W, Chen G, Lu Y, Feng C, Tu XM. Correlation and agreement: overview and clarification of competing concepts and measures. Shanghai Arch Psychiatry 2016;28:115–20.

[52] Nativ NI, Chen AI, Yarmush G, Henry SD, Lefkowitch JH, Klein KM, et al. Automated image analysis method for detecting and quantifying macrovesicular steatosis in hematoxylin and eosin-stained histology images of human livers. Liver Transpl 2014;20:228–36.

[53] Batool N. Detection and spatial analysis of hepatic steatosis in histopathology images using sparse linear models. 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA) IEEE; 2016. p. 1–6.

[54] Munsterman ID, van Erp M, Weijers G, Bronkhorst C, de Korte CL, Drenth JPH, et al. A Novel Automatic Digital Algorithm that Accurately Quantifies Steatosis in NAFLD on Histopathological Whole-Slide Images. Cytom Part B Clin Cytom 2019;9999B:1–8.

[55] Vanderbeck S, Bockhorst J, Komorowski R, Kleiner DE, Gawrieh S. Automatic classification of white regions in liver biopsies by supervised machine learning. Hum Pathol 2014;45:785–92.

[56] Fiorini RN, Kirtz J, Periyasamy B, Evans Z, Haines JK, Cheng G, et al. Development of an unbiased method for the estimation of liver steatosis. Clin Transplant 2004;18:700–6.

[57] Guo X, Wang F, Teodoro G, Farris AB, Kong J. Liver Steatosis Segmentation With Deep Learning Methods. In Institute of Electrical and Electronics Engineers (IEEE); 2019. p. 24–7.

[58] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag; 2014. p. 818–33.

[59] Ershoff BD, Lee CK, Wray CL, Agopian VG, Urban G, Baldi P, et al. Training and Validation of Deep Neural Networks for the Prediction of 90-Day Post-Liver Transplant Mortality Using UNOS Registry Data. Transplant Proc 2020.

[60] Lau L, Kankanige Y, Rubinstein B, Jones R, Christophi C, Muralidharan V, et al. Machine-Learning Algorithms Predict Graft Failure after Liver Transplantation. Transplantation 2017;101:e125–32.

[61] Zarella MD, Bowman D, Aeffner F, Farahani N, Xthona A, Syeda, et al., et al. A Practical Guide to Whole Slide Imaging A White Paper From the Digital Pathology Association A Practical Guide to Whole Slide Imaging-Zarella et al. Arch Pathol Lab Med 2019;143:222–34.

[62] Evans AJ, Salama ME, Henricks WH, Pantanowitz L. Implementation of whole slide imaging for clinical purposes issues to consider from the perspective of early adopters, 141. Archives of Pathology and Laboratory Medicine College of American Pathologists; 2017. p. 944–59.

[63] Bauer TW, Slaw RJ. Validating whole-slide imaging for consultation diagnoses in surgical pathology. Arch Pathol Lab Med 2014;138:1459–65.