A Large-Scale Internal Validation Study of Unsupervised Virtual Trichrome Staining Technologies on Non-alcoholic Steatohepatitis Liver Biopsies

Joshua J. Levy[1,2,3,*], Nasim Azizgolshani[2], Michael J. Andersen Jr.[1], Arief Suriawinata[1], Xiaoying Liu[1], Mikhail Lisovsky[1], Bing Ren[1], Brock C. Christensen[2,4,5], Louis J. Vaickus[1]

1. Emerging Diagnostic and Investigative Technologies, Clinical Genomics and Advanced Technologies, Department of Pathology and Laboratory Medicine, Dartmouth Hitchcock Medical Center, Lebanon, NH 03756
2. Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756
3. Program in Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756
4. Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756
5. Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756

* To whom correspondence should be addressed: joshua.j.levy.gr@dartmouth.edu

Abstract
Non-alcoholic steatohepatitis (NASH) is a type of fatty liver disease characterized by accumulation of fat in hepatocytes and concurrent inflammation. It is associated with significant morbidity as it can progress to cirrhosis and liver failure and increases the risk of developing hepatocellular carcinoma. NASH has become increasingly prevalent in the US and is a leading cause of liver transplantation for end stage liver disease and a significant public health burden. Grading and staging NASH involves liver core biopsy, followed by microscopic examination of histologic preparations of the tissue. Tissue sections are stained with hematoxylin and eosin (H&E) to grade NASH activity, and stained with trichrome to help stage fibrosis. Methods to computationally transform one stain into another on digital whole slide images (WSI) can reduce the need for additional physical staining besides H&E, alleviating personnel, equipment, and time costs. Recent computational techniques have shown promise and success for virtual staining of tissue. We conducted a large-scale validation study of the viability of virtual staining technologies for H&E to trichrome conversion on WSI (n=574). Our results demonstrate that virtual trichrome technologies present an immediate software solution that can be readily employed in the clinical setting as a diagnostic decision aid.

Introduction

Non-alcoholic steatohepatitis (NASH), the most serious condition in the Non-Alcoholic Fatty Liver Disease (NAFLD) spectrum, is a liver disease characterized by serologic markers of hepatocyte injury as well as distinct histological features[1]. NASH is associated with significant morbidity and mortality [2–4]. Patients with NASH are at increased risk of progressive fibrosis with end-stage liver disease characterized by reduced hepatic synthetic function, portal hypertension, encephalopathy, and malignancy such as hepatocellular carcinoma. Portal hypertension can lead to variceal bleeding, splenomegaly, ascites, and renal injury [5]. Prevalence of NASH in the United States has been increasing in recent years in parallel with rising levels of obesity and metabolic syndrome[6]. It is predicted to become the leading cause of liver transplantation in the next decade [7].

Definitive diagnosis of NASH requires tissue biopsy. These biopsy findings can motivate patients to make lifestyle changes (such as improved diet and weight loss) that significantly improve their prognosis. Additionally, appropriate staging can determine if a patient is eligible for routine screening for hepatocellular carcinoma or treatment for varices, ascites, amongst other conditions. Therefore, accurate and timely diagnosis is crucial for patient care. For NASH evaluation, liver biopsies are fixed in formalin, embedded in paraffin and sectioned. Tissue sections are subsequently stained with two principle stain combinations: a routine tissue stain called hematoxylin and eosin (H&E)[8] for grading of NASH and a trichrome special stain that highlights fibrous tissue blue for staging of fibrosis.[9,10] Macrovesicular steatosis of hepatocytes, parenchymal inflammatory foci and hepatocyte ballooning are hallmarks of NASH used in grading disease activity while the distribution, extent and interconnectivity of fibrous tissue is used to stage fibrosis.[11] Maintaining the reagents and technical staff to perform dozens of special chemical stains in a CLIA certified lab is costly and resource intensive[12]. It is thus attractive to computationally "infer" or "translate" a digital WSI of a ubiquitous chemical stain (H&E) to a special stain (trichrome)[13].

The past 20 years have seen the development of sophisticated computational pipelines that operate on digitized representations of slides, referred to as Whole Slide Images (WSI). Prior to the age of machine learning many of the traditional computational approaches to medical image analysis relied on overly simplistic assumptions as to the shape, color and structure of the tissue features of interest (e.g. nuclei are round, the cytoplasm is pink).[14] These rigid assumptions often lead to subpar performance on morphometric tasks and fail to capture more informative features that are too esoteric, or high dimensional for pathologists and programmers to effectively capture in rules based programs.[15] Deep learning approaches, heuristics that utilize artificial neural networks (ANN) and convolutional neural networks (CNN), circumvent this issue in that they are able to find important shapes and patterns for prediction without human specification by capturing and integrating low level features into higher level abstractions[16,17]. Given the advent of these powerful machine learning tools, the Department of Pathology and Laboratory Medicine at Dartmouth Hitchcock Medical Center (DHMC) has fully embraced digital pathology with the caveat that it must leverage advanced computational technologies to provide significant benefits over

traditional glass slides.[18–20] To this end, we have been developing and validating various deep learning technologies in histopathology and designing them to clinical scale[18,21–25].

Generative image translation techniques transform an image from a source domain A to a target domain B. Recently, these techniques have seen much use in the clinical space, from converting MRI to CT scans to the ability to generate retinal fundus images from blood vessel scans for diabetic retinopathy[26,27]. As such, these technologies present an attractive opportunity in digital histopathology stain translation. The most promising of these methodologies rely on the use of Generative Adversarial Networks (GAN), which generate realistic images from noise using two ANNs; one, the generator, which "generates" the images, and another, the discriminator, which decides whether an incoming image is real or synthetic[28]. The quality of images created by GANs improves as the generator “learns" to fool the discriminator through an iterative process of trial and error. GANs have been applied in histopathology with success, including for stain translation from autofluorescence imaging[29], to removal of technical artifacts, as well as for nucleus detection and augmentation of deep learning datasets with synthetic images to improve prediction accuracy.[13,29–41]

Our initial preliminary evaluation of technologies for inferring trichrome stains from H&E featured a small dataset of images to test the viability of utilizing such techniques at our institution.[23] Here, we perform a full-scale internal validation study of inferring trichrome stains from H&E at our institution for near-term incorporation into our clinical workflow. We describe an image translation model for the purpose of converting H&E stains into trichrome stains, rigorously validate the technology, and highlight technical improvements that demonstrate its readiness for clinical deployment.

Methods

Study Population

Our study cohort consisted of 273 individuals with biopsy proven NASH diagnosis from Dartmouth Hitchcock Medical Center in Lebanon, New Hampshire. All patients had paired H&E and trichrome sections (574 total WSI; 14 patients had repeat biopsies; 21 of the 287 liver biopsies were wedge biopsies), taken from July 2007 to March 2019. Retrospective demographic data, serologic markers, and symptoms of end stage liver disease were collected from patient charts. We chose to collect data on patient characteristics that have been shown to correlate with fibrosis score and prognosis such as BMI [42], status of having a diagnosis of impaired fasting glucose, and diabetes mellitus[43]. Additionally, in order to address potential discrepancies between scores of traditionally stained tissues and their virtual counterparts, we collected data on non-invasive markers of hepatic fibrosis that are widely used. AST and ALT are well known independent predictors of NASH induced liver fibrosis[44]. Markers of reduced hepatic function such as INR, albumin and downstream effects of portal hypertension such as hyponatremia, thrombocytopenia, creatinine levels were also recorded. Sodium levels were capped at 137 milliequivalents per liter as values greater than that are not found to be predictive of disease severity and this is the value used to calculate the MELD score [45]. We collected the serologic data from the time of diagnosis and last clinical encounter.

Lastly, we collected serologic markers in order to calculate MELD and Fib4 scores which have been shown to correlate with outcomes of steatohepatitis. Fib-4 combines platelet count, ALT, AST and age to calculate a score which has been shown to have good predictive accuracy for advanced fibrosis in NASH [46–48]. The MELD score is a prospectively developed and validated measure of liver disease severity commonly used by United Network for Organ Sharing (UNOS) to assess the need for transplantation [49]. One-Way Analysis of Variance (ANOVA) and Chi-Squared statistical tests were performed to compare these demographic statistics between the mean fibrosis stages assigned to the real slides.

Biopsy Collection and Digitization

Liver tissue from our cohort of NASH patients were obtained by core needle biopsy, fine needle biopsy, and wedge resection. Samples were fixed and embedded in paraffin blocks and 5-micron sections were stained with H&E and trichrome. Slides were scanned using the Leica Aperio-AT2 (Leica Microsystems, Buffalo Grove, IL) scanner at 20x magnification, stored in SVS image format, from which they were extracted and converted to NPY format via our in house pipeline, *PathFlowAI.* [18]

We utilized 20 H&E/Trichrome WSI pairs for training the deep learning model. WSI are large images that can span hundreds of thousands of pixels in any given spatial dimension. As such, they are currently too big to fit into memory for a deep learning analysis and must be broken into smaller sub-images. In this case, a whitespace detector was used to extract 33,456 non-overlapping 512-pixel subimages that contained up to 8% of white space, correspondent to the 40 training WSI.

Model Selection

There are two major GAN translation techniques that may be utilized for stain translation depending on the type of data available, Pix2Pix or CycleGAN. Pix2Pix is able to translate one type of stain to another based on perfectly aligned paired images. Because of the one to one correspondence of pixels in the source and target domains, Pix2Pix is better able to retain the original image structure in the translation image.[50] Analogously, for unpaired images, CycleGAN is utilized. Because the source and target images are unpaired and may show highly disparate scenes (e.g. a horse and a zebra in different poses), CycleGAN must compare the generated synthetic images to the real images and back translates to the original image to retain the original structure (see Supplementary Material section “Mathematical Description of CycleGAN”) [51]. In our research, CycleGAN translated images tend to differ structurally from their source images more significantly than Pix2Pix translations, presumably due to the lack of perfectly paired source material.

We had to decide which of these two aforementioned image translation deep learning algorithms to utilize. The use of a Pix2Pix model would require paired images to be supplied and necessitate perfect alignment of the matching H&E and trichrome WSI. It proved too difficult to register WSI with enough fidelity for Pix2Pix, primarily as a result of tissue changes between adjacent sections [22]. Thus, we opted to train a CycleGAN

model due to imperfect registration [22] and different tissue features in adjacent tissue sections.

Experimental Design

We trained a CycleGAN model on the 33,456 subimages from 20 paired WSI, randomly cropping 256-pixel subimages from each of these 512-pixel tiles to augment the amount of usable data. GANs can train unpredictably and thus may not have an objective to converge. Instead, the user must visually inspect images during the training process to verify training iterations where the model is producing highly realistic images. However, judging the quality of the any given subimage was not sufficient to decide whether the translation technique could produce highly realistic trichrome WSI. Therefore, we built into the currently existing CycleGAN framework the ability to test the latest iteration of the model on an entire WSI by dynamically extracting and translating overlapping subimages from the H&E WSI, then blending overlapped areas as they were used to construct the virtual trichrome. This ensured maximum realism and reduced the appearance of obvious "tiling artifact". During training, we selected two representative H&E/Trichrome pairs (demonstrating different distributions and intensities of trichrome staining) to dynamically translate at various points during model training. A pathologist then visually compared the predicted trichrome WSIs to the actual trichrome WSIs at a variety of points during model training (Figure 1a). After training the model across 200,000 batch iterations (Figure 1b), a pathologist selected the top model through visual comparison of the real and virtual trichrome WSI generated for these two images (Figure 1c). We applied the chosen CycleGAN model to virtually trichrome stain 287 H&E WSI using an inhouse pipeline that encases the selected model, performs the translation and exports WSI into formats that pathologists can read through open-source applications like QuPath or ASAP [52]. Our in-house pipeline also deploys the slides onto a password-protected image server that allows for tissue-staging from any private web browser through the use of openseadragon[53] (Figure 1d).

Four board-certified pathologists independently staged the virtual and real trichrome stains via the NASH CRN system [54]. The inspection of synthetic and real images was separated by a period of two weeks to reduce the risk of the pathologist recognizing previously staged counterparts in the synthetic image set. The real image set was also numbered and ordered differently from the synthetic set. All pathologists staged separately the virtual and real stains for each of the slides, (8 total assessments per slide (4 for real, 4 for virtual); one pathologist was not able to determine stage for two slides; 287 slides; 4*287-2=1146 observations). Images were staged using the NASH CRN system, a five-point scale with F0 indicating no fibrosis, F1 as portal fibrosis without septa, F2 as few septa, F3 as numerous septa, bridging fibrosis without cirrhosis, and F4 as cirrhosis) [54,55].

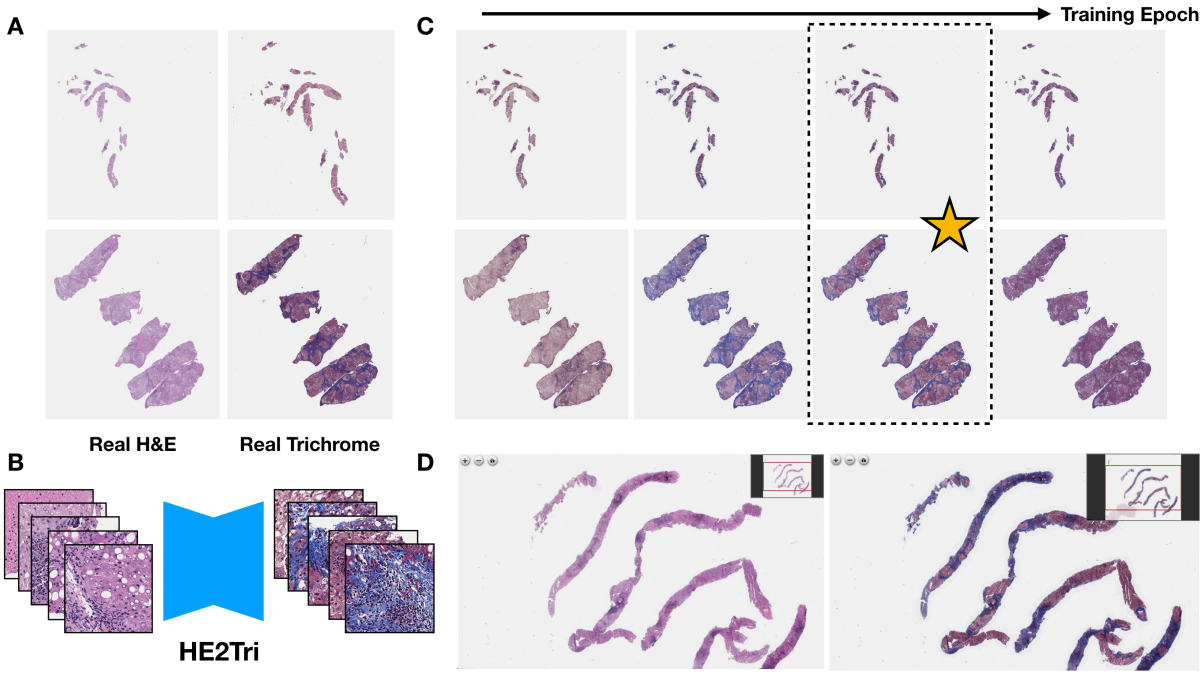

Figure 1: Overview of Workflow: a) Original H&E and trichrome stained slides; b) H&E patches are converted into trichrome patches using CycleGAN; c) Entire trichrome WSI are generated at various training iterations; ideal model is selected; d) Final model is deployed and real time generated/virtual wsi (right) are presented next to original WSI (left) in OpenSeaDragon-based image server for pathologist review [53]

Quantitative Assessment of Concordance

To validate our technology, we utilized various statistical modeling approaches to demonstrate agreement between pathologists staging of the real trichromes and the corresponding virtual trichromes. As a coarse measure of agreement, we averaged the assigned stages across the pathologists for both the real trichromes and virtual trichromes, under the premise that the average of the real stages presents the “ground truth” or “consensus opinion” stage to be compared against. Under the assumption that the cases were drawn from a random population of cases, and there were multiple raters and slides per case, we assumed clustering on the case level (random effect) and measured the relationship between real and virtual staining using a linear mixed effect model (LMM). The conditional R^2 of the LMM model served as the estimate for the strength of the relationship between the real and virtual stain stage. Since averaging the scores may inflate the apparent performance of our technology, we also correlated the real to virtual assigned stages, with clustering on the rater and case level. Finally, we considered a modeling approach that would account for the bias of a given pathologist. This was actualized by whether they were likely to assign a higher or lower stage based on the average stage assigned for the reals. We now assumed the rater to be a fixed effect, the interaction between the rater and mean real stage would indicate the bias or divergence of the rater from the group of pathologists, and related these terms with the virtual stages using a robust linear mixed effects model to control for additional bias/outliers. The model was used to make patch level predictions and was thus not tied to staging being done on the WSI-level. In this spirit (level of analysis), and with regards

to preliminary results that demonstrated that including the training slides in the evaluation set had negligible impact on concordance, we utilized the entire set of slides for the correlation analyses. We also correlated and clustered the stages assigned by each pathologist for the real and virtual slides to further assess bias.

After performing a correlation analysis, which assessed overall agreement across a wide range of stages, we wanted to understand whether pathologists could accurately assess virtual slides for the patients with the highest potential for progression to cirrhosis. We defined advanced fibrosis slides as slides where the real trichromes were staged in high agreement between physicians, with mean fibrosis stage greater than or equal to F3.5 and a median absolute deviation (inter-pathologist variation) from the median stage of 0.35. We retained these slides (where there is strong pathologist agreement for advanced fibrosis) for analysis. We assessed agreement in staging advanced fibrosis for these slides by recording proportion of times that each pathologist staged a given virtual trichrome as advanced fibrosis. We averaged these accuracy scores to yield a final agreement measure for advanced fibrosis. We then performed a sensitivity / specificity assessment for the presence or absence of advanced fibrosis for the entire set of pathologist determined real and virtual fibrosis stages. A receiver operating curve (ROC) and correspondent area under the curve (AUROC) was generated for this assessment.

Assessing Relationships with Clinical Endpoints

In order to further test the accuracy of the virtual trichrome stain, we aimed to assess if the staging obtained by virtual and real stains were not only correlated with each other but with serologic markers of advanced fibrosis. These markers provide us with a means of assessing discordance between the virtual and real staging means against quantitative serological data which is known to be correlated with fibrosis / NASH severity. We utilized Spearman's Correlation Coefficient to correlate known serological markers with virtual staging, real staging, the residual between the real and virtual assigned stage, and the variation in pathologist staging of the tissue (as defined by the median absolute deviation). For the sodium biomarker, we capped the maximum value at 137 and any value greater was marked as 137 as is the protocol in calculating MELD scores[45]. We also correlated fibrosis staging with the AST/ALT ratio, of which an AST/ALT ratio greater than one is associated with bridging fibrosis [44].

We additionally used Kaplan-Meier survival analysis to determine the relationship between virtual and real fibrosis stage with development of end stage liver disease. Our model adjusted for age and BMI. An event was defined as the occurrence of ascites, hospitalization due to liver disease, encephalopathy, variceal bleed, evidence of varices with imaging, a diagnosis of hepatocellular carcinoma, dialysis, or the prescription of non-specific beta blocker propranolol as a prophylaxis for esophageal variceal bleed during follow up. Multivariate cox proportional hazard models adjusting for age and sex were also used. As these are markers of late stage liver disease, we also dichotomized our cohort into those with a virtual fibrosis stage greater than or equal to F3.5 and less than F3.5 and re-ran our analyses.

Results

Results Overview

We conducted a large-scale internal validation study of H&E to trichrome conversion by training a CycleGAN model on subimages from 20 paired WSI. Then, we converted our entire set of 287 H&E stained WSI into trichrome stained WSI; when processed in series, it took an average of 5 minutes to convert an entire H&E WSI into a virtual trichrome stained WSI. While other studies have chosen to assess the quality of the stain as the means of assessing the synthetic images, which is only tangentially related to the true utility of the virtual stain, we believe that the best way to assess the viability of the approach is whether pathologists can still assign the same fibrosis stage as if they were examining the true trichrome stain on an adjacent tissue section (after accounting for bias/interrater variability). Assessments of virtual trichrome stains should also correlate with actual clinical endpoints (eg. progression to cirrhosis, serological markers of liver injury and failure) in the same manner as the real stain. Our results indicate that the CycleGAN model we trained is capable of generating virtual trichrome stains from H&E which are sufficient for clinical use.

Patient Demographics

We tabulated patient demographic information (age, sex, BMI, various serological markers such as Fib4, AST, ALT) from our study cohort, which is summarized in Supplementary Table 1. Calculation of chi-squared statistics for categorical demographics and ANOVA tests for continuous outcomes demonstrated statistically significant differences in some but not all of the serological markers across the mean Fibrosis stages for the real trichrome slides.

The mean age of our cohort was 53, ranging from 8-81. Consistent with previous studies, our cohort had an elevated BMI with a mean of 34 [42]. Table 1 displays characteristics of our study population stratified by fibrosis score obtained by inspection of the real trichrome. The majority (60%) of our patient cohort had significant fibrosis with a NASH CRN score greater than or equal to F2.

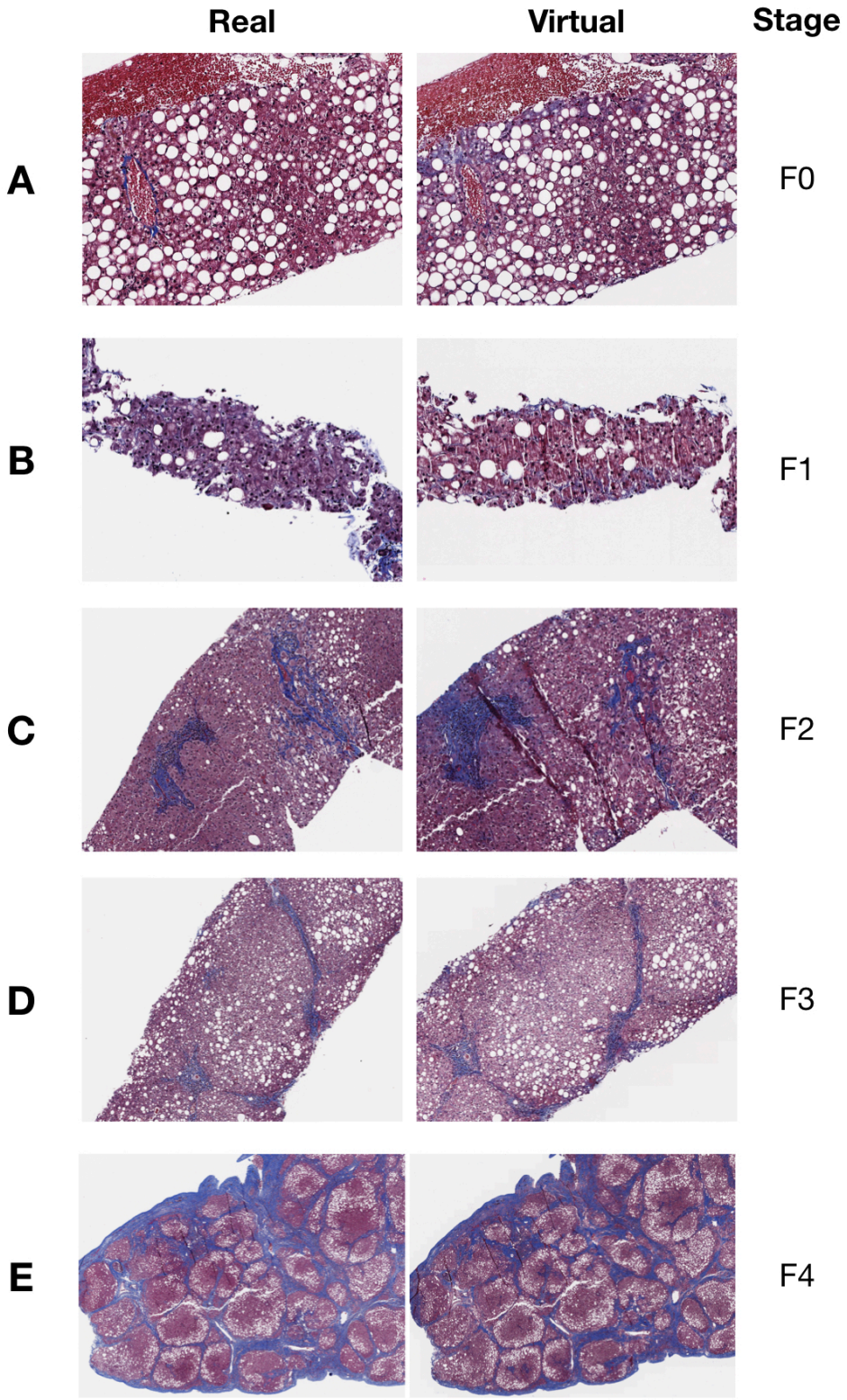

Figure 2: Select Regions of WSI across different accepted Fibrosis stages by: a) Real trichrome, b) Virtual trichrome stains; Entire WSI (including H&E stain) are found in Supplementary Figures 1-5

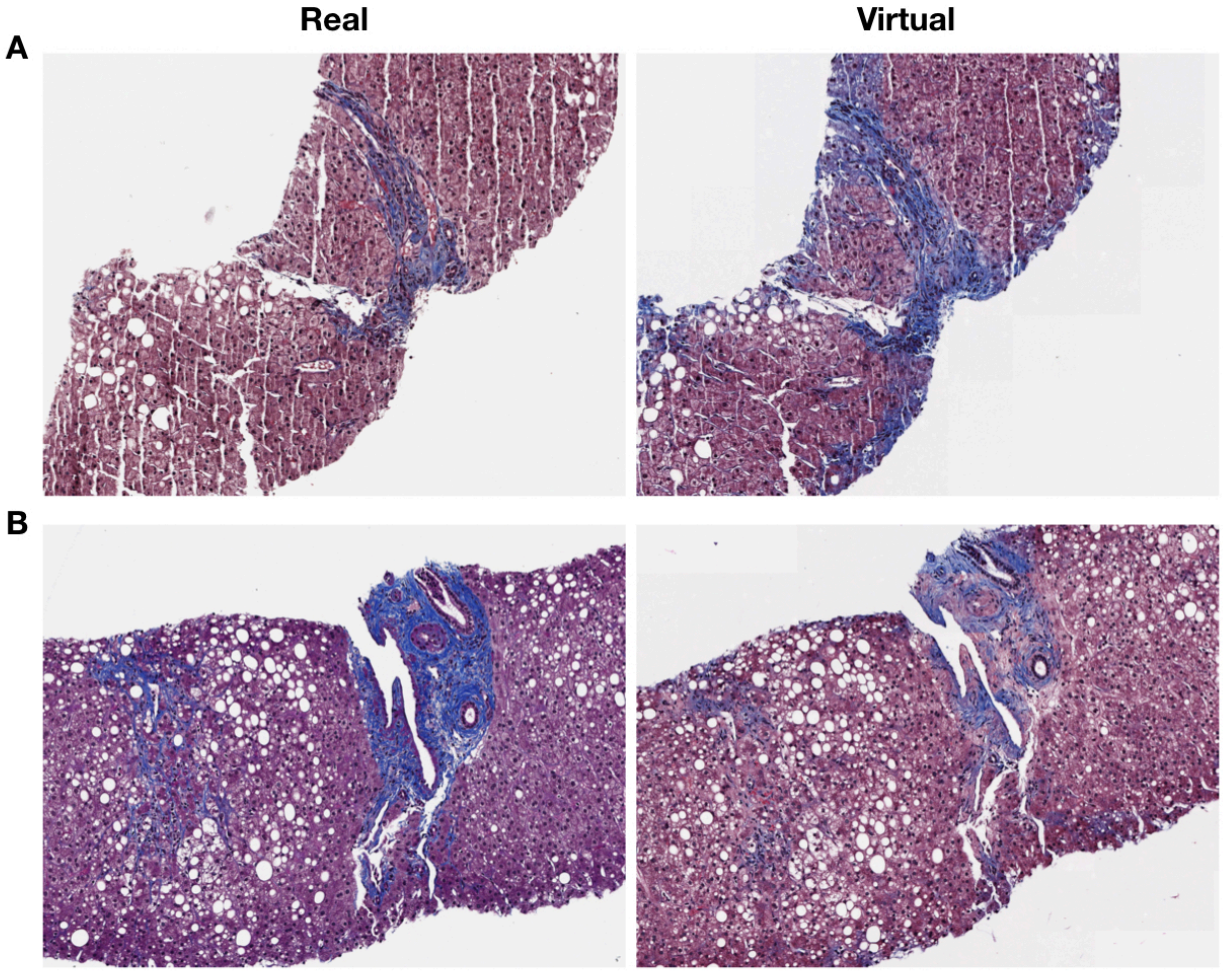

Figure 3: Examples of Virtual Staining Artifacts: a) Overstaining; b) Understaining

Histological Inspection

When comparing pairs of real and virtual trichrome images there was generally excellent agreement and for these cases it was difficult to appreciate a difference at low magnification (Supplementary Figures 1-5). For stage 0 fibrosis (figure 2a) the virtual trichrome demonstrates hazy, serum-like staining in the periphery of the tissue and slight understaining of the wall of the central vein. In stage 1 fibrosis (figure 2b), the virtual stain shows patterns consistent with chicken wire fibrosis (due to perisinusoidal scarring) although to a lesser extent than in the analogous field in the real trichrome stained slides. In stage 2 fibrosis (figure 2c) the virtual stain shows similar levels of portal and chicken-wire fibrosis with some aberrant edge staining. In stage 3 fibrosis (figure 2d) the virtual stain demonstrates a nearly identical distribution of fibrosis as compared to the real image with slight edge-staining. In stage 4 fibrosis (figure 2e) the virtual fibrosis staining pattern is virtually indistinguishable from the real trichrome. During pathologist review of the virtual trichrome WSI, the most frequent artifacts encountered were aberrant edge staining in the absence of capsule, and patchy overstaining / understaining. In some cases, the synthetic image was subjectively overstained (figure 3a). In this example, a low magnification inspection demonstrates

areas of both overstaining and understaining (large portal area). At higher magnification there is obvious portal, parenchymal and edge overstaining in the virtual stain. An example of understaining of the synthetic image (figure 3b) is apparent in both the portal and parenchymal areas.

Pathologist Staging and Observations

Four pathologist raters were first given the entire set of virtual trichrome WSI to stage. Then two weeks after staging the virtual stains, they were asked to stage the real trichrome WSI. We solicited comments from the pathologists on the staining quality for each slide to improve future iterations of virtual staining procedures. We compiled a table of all pathologist comments on the real and virtual stains (Table 1)

Table 1: Collection/Summary of pathologist comments on slides; a lot of the same comments mentioned for the real stains (e.g. fragmented), were also noted in the virtual stains, but there were additional artifacts introduced from virtual staining. In comparison to real stains (two slides did not receive conclusive stage), all virtual stains were able to be staged.

Comments on Real Stains
Wedge/Needle biopsy
Fragmented biopsy (affected by biopsy procurement)
Rare Portal tract (suboptimal number of portal tracts for analysis)
Tangential section to capsule
Unable to Stage (n=2)

Additional Comments on Virtual Stains
Understaining
Understaining of capsules, with visible nodules
Overstaining of periphery regions
Blue hepatocytes

While there were additional understaining/overstaining artifacts in addition to artifacts related to biopsy procurement, the quality of the majority of the virtual stains was excellent. We additionally felt that the common artifacts would have negligible impact on the ability for the pathologist to stage the slide. Thus, we sought to quantify the concordance between the clinical staging of the real and virtual slides to assess the true viability of staging virtual stains while accounting for rater subjectivity.

Interrater Concordance

First, we compared the distribution of pathologist staging for both real and virtual trichrome stains to assess interrater reliability issues via an ANOVA test, Tukey's post-hoc honest significant different test, and a correlation matrix of scores. We found that only 2 pairs of all possible combinations of pairing pathologists agreed with each other within real-staged images; likewise, only two pairs of all possible pairings of pathologists agreed with each other within virtual-staged images. Running pairwise Kolmogorov

Smirnov tests (with Bonferroni correction) for differences in the distribution of assigned stage between pathologists demonstrated agreement between only one pair of pathologists. An assessment of pairwise correlation demonstrated that the average correlation in fibrosis stages assigned between pairs of pathologists was 0.705 for the real stains, 0.704 for the virtual stains, and within the same pathologist for real versus virtual, was 0.700, indicating that the rough concordance within pathologist for real versus virtual was not outside of what would be expected when comparing two pathologists to each other (Figure 4d). We also compared the median absolute deviation between pathologists. The disagreement between pathologists tended to be highest in the middle stages (F2) for both real and virtual trichromes (supplementary table 2, spearman correlation coefficient for real: $\rho = -0.11, \mathrm{p} < 0.001$; Spearman correlation coefficient for real: $\rho = -0.33, \mathrm{p} < 0.001$) and disagreement in scores for virtual stains was statistically significantly higher than disagreement over real stages ($\beta = 0.96, p = 0.0007$).

Correlation between Real and Virtual Stains

Next, we sought to estimate the concordance between mean pathologist staging of the real and virtual trichrome stains. The initial correlation between mean virtual and mean real stage yielded a conditional R^2 value of 0.952 ($\beta = 0.82, \mathrm{CI}\ 0.75 - 0.89, \mathrm{p} < 0.001$) after accounting for explainable heterogeneity between cases (ICC=0.85, n=287, within cluster variance 0.05). However, when correlating individual pathologist's virtual and real stages, this concordance dropped to an R^2 of 0.657 ($\beta = 0.52, \mathrm{CI}\ 0.47 - 0.56, \mathrm{p} < 0.001$) after accounting for variability between raters and cases, as the stages for virtual slides were pushed away from the extreme values (0 and 4) (ICC=0.51, n=1146, within cluster variance 0.42). Finally, when accounting for physician bias, the conditional R^2 increased to 0.769 ($\beta = 0.89, \mathrm{CI}\ 0.81 - 0.98, \mathrm{p} < 0.001$; ICC=0.38, n=1146, within cluster variance 0.33) when correlating to virtual fibrosis stage (Supplementary Table 3; Figure 4a,c).

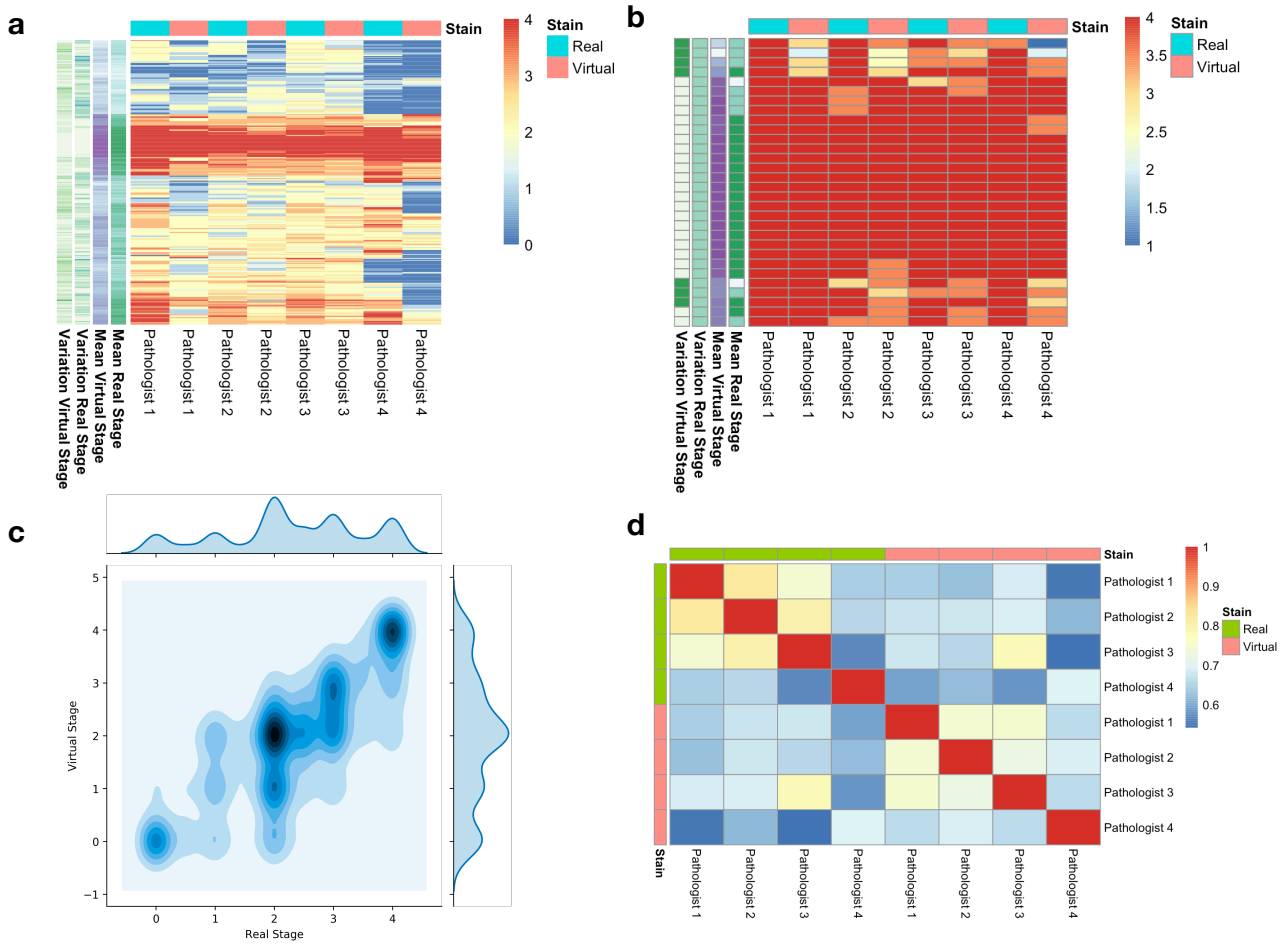

Figure 4: Concordance Plots Comparing Staging of Real to Virtual Stains; a) Heatmap of raw scores assigned by pathologists for real (red column color) and virtual (blue column color); each row represents a slide, each column represents a pathologist and whether they were staging real/virtual slide; row color tracks indicate (from left to right): variation in staging virtual trichromes between pathologists, variation in staging real trichromes between pathologists, mean virtual fibrosis stage assigned, mean real fibrosis stage assigned; darker colors indicate higher values for color tracks; b) Similar Heatmap of Agreement between pathologists but for slides deemed to have real advanced stage fibrosis (mean real stage ≥ 3.5, median absolute deviation ≤ 0.38); c) 2-D Density plot of all scores assigned between real and virtual stains for the same slide; darker areas indicate higher concentration of score combinations between real and virtual; off-diagonal elements have different staging between real and virtual images; d) correlation matrix of staging between pathologists; column and row color track indicates whether staging real or virtual slides; blue indicates lower correlation, red indicates higher correlation; block diagonal elements indicate agreement between pathologists for real stains and virtual stains (within staining type) respectively; off-block diagonal indicate correlations between real and virtual stains between pathologists (between staining types).

Advanced Stage Fibrosis and Sensitivity Analysis

It is of paramount importance to detect advanced fibrosis with a high degree of sensitivity. Thresholding by a median absolute deviation of 0.38 for mean real fibrosis scores at least 3.5 ($F \geq 3.5$), staging of virtual slides recapitulated the advanced fibrosis ($F \geq 3.5$) staging 89.2% of the time (0.867, 0.867, 0.867, 0.966 for the pathologists respectively) (Figure 4b). A sensitivity analysis for advanced fibrosis ($F \geq 3.5$) across the entire image set yielded an AUROC of 0.9603 (95% CI 0.9361-0.9845; Figure 5).

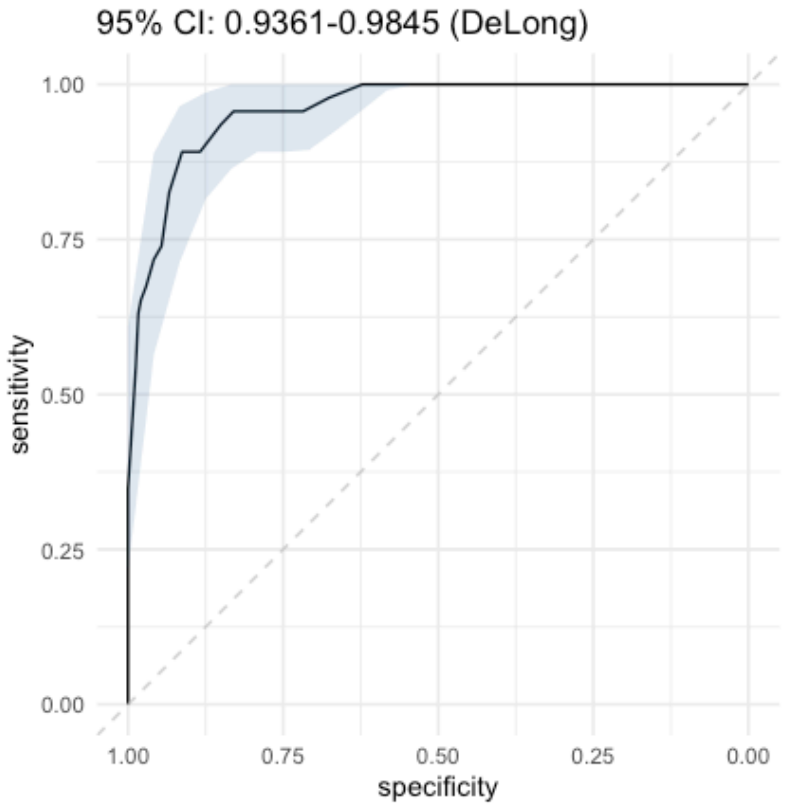

Figure 5: Receiver operating curve for virtual staining of advanced fibrosis ($F \geq 3.5$) when comparing mean virtual stain to mean real stain after dichotomizing real stains by an average fibrosis stage of at least 3.5

Serological Analysis

While the results above demonstrate strong agreement in fibrosis stage between virtual and real trichrome staining, we hoped to correlate the assigned stages with serological markers of fibrosis and long-term effects of end stage liver disease to determine if the virtual stain correlates with disease severity. There were significant positive associations with markers of hepatocyte injury AST and ALT ($p < 0.001$) as well as hepatic synthetic function such as INR ($p < 0.001$). Additionally, there was a strong association between virtual fibrosis stage and downstream effects of portal hypertension such as ascites ($p=0.001$) and sodium levels ($p=0.005$). Lastly, real and virtual fibrosis stage were tested for correlation with non-invasive scoring systems MELD and Fib-4. Interestingly, although we observed a trend toward correlation for both real and virtual stains with the MELD score, only the virtual stage met statistical significance criteria ($p = 0.01$). Both virtual and real stages were significantly associated with Fib-4 score (p <

2.2E-16). These results further support the validity of utilizing virtual trichrome stains for fibrosis staging (Supplementary Table 4).

Staging differences between virtual and real slides was also associated with key serological markers (Supplementary Table 4).

Survival Analysis

Perhaps the most important indicator of whether the virtual stains can be meaningfully employed in the clinic is whether staging of these tissues are equally predictive of risk of progression to end stage liver disease.

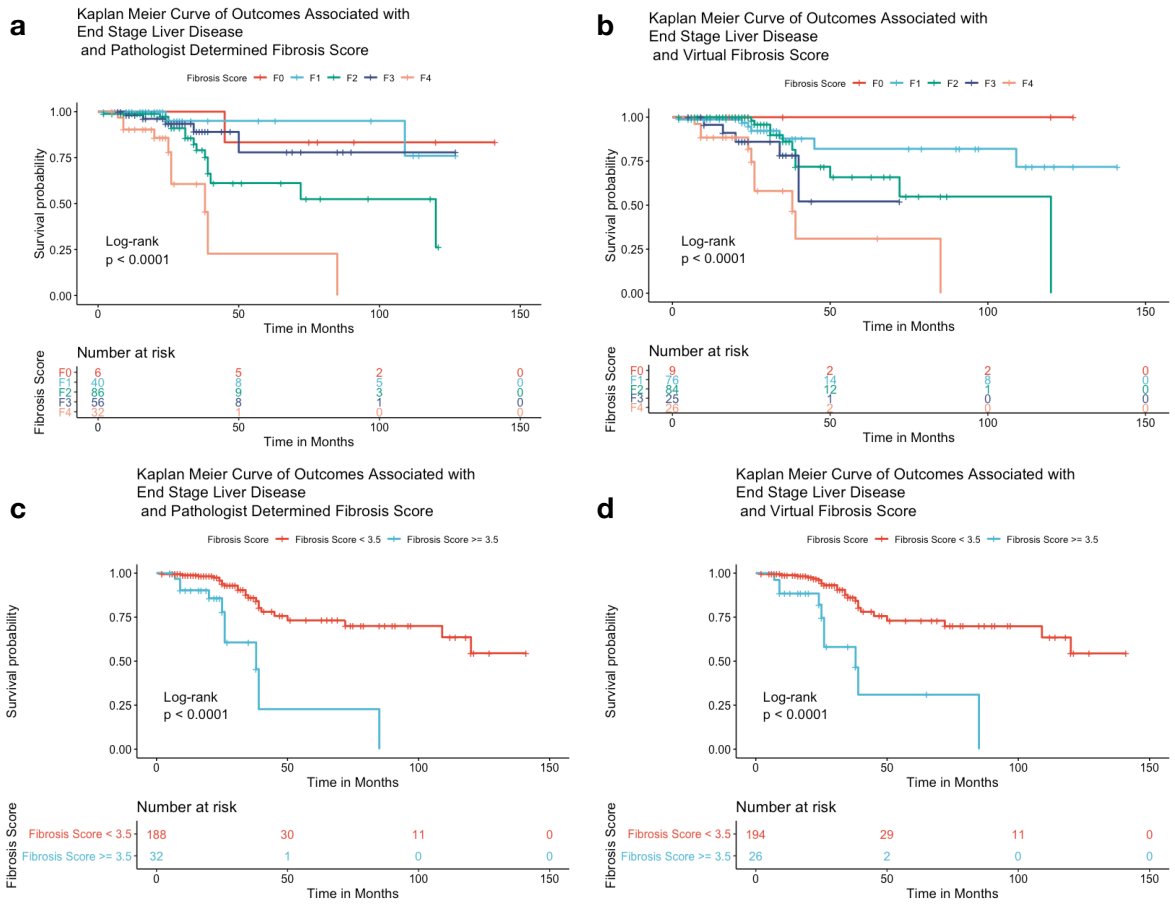

Figure 6: Kaplan-Meier Plots for the following conditions: (a,c) real stain staging; (b,d) virtual stain staging; (a,b) utilizing the mean stage assigned by pathologists for that slide rounded to nearest integer; (c,d) dichotomizing staging by advanced fibrosis ($F \geq 3.5$)

We found that for both virtual and real stained tissue the NASH CRN fibrosis stages were equally predictive of occurrence of symptoms of end stage liver disease (ESLD) via Kaplan-Meier (KM) analysis (virtual trichrome log-rank p < 0.001, real trichrome log-rank p < 0.001, Figure 6). All KM estimates for treating staging as a continuous measure (mean across pathologists, rounded to nearest integer) and dichotomous measure ($F \geq 3.5$) yielded p-values less than 0.0001 via log-rank statistic (Figure 6). The KM plots were consistent for dichotomized staging between real and virtual stains. Interestingly,

the virtual trichrome fibrosis stage, considered as a continuous measure, seemed to reflect the risk of progression to End Stage Liver Disease (ESLD) more accurately for the various fibrosis stages. Namely, the order of time to progression to ESLD for the real trichrome fibrosis stages is F4 < F2 < F3 < F0 < F1 while in the virtual trichrome fibrosis stages the order is F4 < F3 < F2 < F1 < F0. (Figure 6b).

A Cox-Proportional Hazards (CoxPH) model for time-to-event outcomes adjusting for age and BMI was also applied which demonstrated an increased hazard of experiencing symptoms of end stage liver disease with increasing fibrosis stage in both real (HR = 2.06, CI 95% 1.38-3.08, P < 0.001) and virtually stained tissues (HR = 2.16, CI 95% 1.52-3.05, p < 0.001). The concordance statistic, which is a measure of fit of the survival curves to the data, was statistically equivalent for models fit on staging of real and virtual trichrome staging, irrespective of whether the staging was dichotomized. The computed hazard ratios were similar in effect and significance (Supplementary Tables 5-6). These trends, as elucidated via the KM and CoxPH analyses, were more pronounced when scores were dichotomized to advanced fibrosis (Fibrosis score greater than 3.5) or early fibrosis (score < 3.5). The aforementioned results suggest that virtual trichrome has the same ability to capture the potential for progression to ESLD as the real trichrome.

There were no statistically significant differences in dichotomized Fibrosis stage for both KM and CoxPH analyses of the real (KM p-value: 0.077, CoxPH p-value: 0.499) and virtual (KM p-value: 0.068, CoxPH p-value: 0.518) stain staging for the outcome of death.

We have summarized the results for concordance between pathologists staging scores for fibrosis. Overall, there was moderately strong reported agreement in scores from the real trichrome fibrosis stage to the stage derived from inspection of the virtual images. This indicates that virtual trichrome conversion can generate stains that can be readily deployed in clinical practice.

Discussion

We investigated technologies that could potentially transition trichrome staining to a digital pathology test rather than a chemical stain, reducing the cost and personnel required to stage NASH biopsies. Our virtually stained tissue were visually consistent with real trichrome stained tissue as assessed by pathologists, and our large-scale internal validation cohort quantitatively demonstrates strong correlation in staging between the real and virtual stains as assessed by pathologists. Virtually stained tissue was able to recapitulate continuous and dichotomized real fibrosis scores with similar correlations to biomarkers and was as predictive of progression to ESLD as their real trichrome stained counterparts. Our research presents a less labor and resource intensive solution to trichrome staining with equivalent predictive value as traditional staining protocols. The ideal use of the technology as it currently stands is likely reflexive virtual trichrome staining for liver biopsies with real trichrome available by request (significantly reducing the volume of real trichrome stains performed).

Compared to other medical GAN studies, the strength in our design is the evaluation of clinical endpoints of the examined disease process. Other studies present "Turing Tests" that, while useful for evaluating the realism of generated images, have previously been utilized in the histopathology setting to assess small pockets of tissue, which do not reflect whether the whole slide "looks real" [38,56–58]. These measures of quality do not assess the viability of the generated slide for the potential to capture actionable clinical endpoints and place the pathologist in the position of evaluating a tiny subimage, rather than the entire slide. Accurately diagnosing and staging the presence of hepatic fibrosis is critical to the management of NASH. The presence of fibrosis, especially advanced fibrosis, has significant implications for medical management, frequency of disease assessment and prognosis. Additionally, accurate diagnostic information may serve as the impetus patients need to initiate lifestyle modifications and may indicate to the clinician the need for medication to manage their NASH risk factors [59]. For these reasons it was extremely important for us to demonstrate, not only that our model could produce subjectively acceptable virtual trichrome subimages, but that it could create convincing WSI which were of sufficient quality to be accurately staged by pathologists. Given that this is an unproven technology that aims to replace or augment part of an already risky diagnostic procedure, we also thought it was necessary to take the additional steps of demonstrating that virtual trichrome fibrosis stage was strongly correlated with serological and clinical markers of liver injury and ESLD in a similar manner to real trichrome fibrosis stages. We were unable to make any substantive claims in our preliminary mortality-based survival analysis due to the relatively few patients who died during the study period [2]. These factors merit further study of virtual staining in a cohort with greater representation of F4 stage tissue and sufficient mortality.

In general, the virtual trichrome images were deemed by the reviewers to be subjectively acceptable. The most common comment recorded for the virtual stains was the presence of aberrant edge staining on needle core biopsies. This is likely the due to the inclusion of wedge biopsies in the training set. Because wedge tissue has significant areas of capsule, the model may have begun to associate any edge with trichrome positivity, even in the absence of capsular tissue [60]. If this is the reason for the edge staining effect, then elimination of wedge biopsies with capsules in the final fine-tuning of our modeling approach may alleviate edge overstaining. Nonetheless, the raters in our study universally indicated that these areas of aberrant staining did not impede fibrosis scoring. Furthermore, quantitative results demonstrated that pathologists were able to recapitulate real late stage fibrosis using the virtual trichrome set.

Limitations

There are several limitations to our work. The experiment was not completely blinded. Each pathologist was given a set of virtual trichromes first and then, two weeks later, staged the real set of trichromes. In order to minimize the chance of a pathologist recognizing a correspondent case between the real and virtual trichrome sets we enforced a "washout" period between viewing the two sets of images and took the additional step of scrambling the case numbers and slide order. We acknowledge that

this experimental design risks introducing bias, given that the raters might seek to judge the synthetic images overly harshly or overly leniently and there is the small risk that the rater might remember their diagnosis from a synthetic case while reviewing the analogous real case. However, given the limited time that raters were able to grant us from their incredibly busy schedules, the goals we were seeking to accomplish (collecting subjective information about the virtual trichromes and rating fibrosis scores between real and virtual trichromes), and the limited number of pathologists with liver expertise available, we deemed these limitations necessary and the results indicate that these potential biases did not play a significant role.

There was a high degree of inter-rater variability amongst assigned fibrosis stages from pathologists, regardless of whether the slide was real or virtual, as assessed from distributional comparisons of tissue staging, correlations, and calculations of reliability. The nature of the variability in scores was not random, as the final concordance modeling approach demonstrated the predisposition of some pathologists to statistically significantly assign a lower/higher virtual/real stage depending on the "true/agreed-upon" stage of the patient. These behaviors may be related to the background of the clinician, age, time constraints, training, blinding, fatigue, personal preference, etc. That pathologists (and other visual experts) can render significantly different diagnoses for the same case (or even reviewing the same case at a later timepoint) has been well demonstrated in the literature [61,62] and represents a known limitation of qualitative diagnostics. In addition to this inter and intra rater variability, oftentimes the pathologist posted a range of stages (e.g. F2-F3) that represented their uncertainty in staging. While we have begun to assess for these points of bias and uncertainty in our modeling approaches in the study, statistical methodologies that more naturally lend themselves towards uncertainty in the observations and outcomes (Bayesian modeling procedures) are warranted to lend additional insight as to the assessment and validation of these technologies and will be pursued in the future [63,64]. Due to physician workload, we were unable to assess intra-rater reliability and were thus unable to see whether their staging of real versus virtual tissue was within what was expected from intra-rater reliability. However, versus other medical GAN study designs which usually average out the scores between clinicians, and consequently reduce the number of observations and inflate the usefulness of the technology, our study utilized all observations through a linear mixed model evaluation of the concordance measurements.

Future Opportunities

Here, we discuss further opportunities to improve upon the results from this study. While these analyses were conducted on NASH cases, fibrosis can result from any chronic liver injury. Aside from NASH, chronic viral hepatitis, alcoholic steatohepatitis, autoimmune hepatitis, chronic biliary cirrhosis, primary sclerosing cholangitis and outflow obstruction are all relatively common causes of chronic liver injury. While fibrosis may be recognized equally well in these disease etiologies, the model needs additional testing on other manifestations of fibrosis unrelated to NASH which often have distinct scoring systems[65].

Furthermore, there are opportunities to improve on the modeling approach through registration and a paired analysis, as hinted at earlier in the paper (e.g. Pix2Pix). Even when tissue is sectioned and captured to a slide perfectly (e.g. 5 microns between H&E and trichrome stain) there are significant alterations to the microstructure of the tissue (e.g. nuclei, cells and other small objects are caught in a different sectioning plane or are completely absent). Achieving perfect registration of an H&E and a trichrome stain for subsequent sections is therefore very challenging and rarely possible. Even if the H&E slide were destained and restained with trichrome, the chemical processing and handling involved (especially repeated cover slipping) introduces significant tissue distortions. We were not surprised when a preliminary evaluation of training a Pix2Pix model after registration of H&E and subsequent trichome sections demonstrated suboptimal results (principally significant structural laxity in the virtual images, so called 'hallucination') where the generated image was subjectively convincing but often bore no structural similarity to the source H&E. These difficulties may be alleviated through use of GAN techniques that allow imperfect registration, perhaps combining the best elements of the CycleGAN and Pix2Pix algorithms to build a more sophisticated model or incorporating new components into the loss term to enforce both structure preservation and capture overall similar tissue features.

There are obvious avenues to improving our current CycleGAN modeling approach including conducting a paired translation analysis, augmenting the number of capsule containing and capsule devoid training images (to address the edge staining artifact), and more examples capturing the center of large fibrotic nodules and portal areas (to address understaining seen in some large structures). In addition, the GAN generator was subjectively assessed (by comparing generated trichromes to their real counterpart) only every 5000 training iterations and there were often large differences between different iterations. Saving more frequently and analyzing additional virtual WSI may therefore allow us to select an even better model for clinical use.

We have containerized our model and pre-processing pipeline into an opensource, deployable software package. These software elements can be readily incorporated into automated clinical workflows through the use of pipelining software such as Docker and Common Workflow Language (CWL) [66,67], and can be scaled to meet any clinical load with no additional coding (given the appropriate high-performance computing (HPC) resources). With additional fine-tuning the model presented herein could be readily deployed as a diagnostic aid. Pathologists could use this tool as a diagnostic aid to flip rapidly between or compare images of H&E and trichrome slides (and other staining modalities), and if they are not satisfied with the staining quality of the virtual slide or find the interpretation ambiguous, can order a chemical stain. These steps can facilitate faster clinical decision making with less resources (both departmental and tissue). Future analyses may explore the prediction of Fibrosis stage from an H&E, obviating the need for clinical inspection of a virtual stain [68]. However, virtual staining is a technology that clinicians may trust more than a computer-generated risk score and therefore presents an immediately viable decision aid technology.

Conclusion

Our CycleGAN-based H&E to trichrome conversion model is interpretable and accurate and with appropriate validation could be incorporated into the clinical workflow as a diagnostic aid. The application of these techniques has the potential to save money, save time and improve patient outcomes by allowing for faster and more efficient fibrosis staging. In addition, methods like virtual trichrome staining offer a strong financial incentive for the adoption of digital pathology. We will investigate techniques to further refine this process while seeking implementation in the clinical workflow at DHMC.

Abbreviations

ANN – Artificial Neural Network
ANOVA – One-Way Analysis of Variance
AUROC – Area Under the Receiver Operating Curve
CNN – Convolutional Neural Network
CoxPH – Cox Proportional Hazards
CWL – Common Workflow Language
DHMC – Dartmouth Hitchcock Medical Center
ESLD – End Stage Liver Disease
GAN – Generative Adversarial Network
H&E - Hematoxylin and Eosin
HPC – High Performance Computing
HR – Hazards Ratio
KM – Kaplan Meier
LMM – Linear Mixed Effects Model
NAFLD - Non-Alcoholic Fatty Liver Disease
NASH – Non-alcoholic Steatohepatitis
ROC – Receiver Operating Curve
UNOS – United Network Organ Sharing
WSI – Whole Slide Images

Disclosures / Conflicts of Interest

The authors declare that they have no financial or non-financial conflicts of interest.

Funding

This work was supported by:

- NIH grants R01CA216265, R01DE022772, and P20GM104416 to BCC
- Two Dartmouth College Neukom Institute for Computational Science CompX awards to BCC, LJV and JJL.
- JJL is supported through the Burroughs Wellcome Fund Big Data in the Life Sciences training grant at Dartmouth.
- Norris Cotton Cancer Center, DPLM Clinical Genomics and Advanced Technologies EDIT program

The funding bodies above did not have any role in the study design, data collection, analysis and interpretation, or writing of the manuscript.

Authors' Contributions

The conception and design of the study were contributed by JJL, BCC and LJV. Implementation, programming, data acquisition, and analyses were by JJL. Survival analysis, collection and analysis of serological markers and demographics were by NA. AS, XL, ML, BR, and MA contributed to data collection and fibrosis staging. All authors contributed to writing and editing of the manuscript.

Acknowledgements

Preliminary results for a small subset of samples featured in this study were presented at the COMP2CLINIC workshop at BIOSTEC 2020 [23].

References

1. Dietrich, P. & Hellerbrand, C. Non-alcoholic fatty liver disease, obesity and the metabolic syndrome. *Best Practice & Research Clinical Gastroenterology* **28**, 637–653 (2014).
2. Hagström, H. *et al.* Fibrosis stage but not NASH predicts mortality and time to development of severe liver disease in biopsy-proven NAFLD. *J. Hepatol.* **67**, 1265–1273 (2017).
3. Paik, J. M. *et al.* Mortality Related to Nonalcoholic Fatty Liver Disease Is Increasing in the United States. *Hepatology Communications* **3**, 1459–1471 (2019).
4. Mantovani, A. *et al.* Complications, morbidity and mortality of nonalcoholic fatty liver disease. *Metabolism* 154170 (2020) doi:10.1016/j.metabol.2020.154170.
5. Al-Busafi, S. A., McNabb-Baltar, J., Farag, A. & Hilzenrat, N. Clinical Manifestations of Portal Hypertension. *Int J Hepatol* **2012**, (2012).
6. Younossi, Z. M. *et al.* Changes in the prevalence of the most common causes of chronic liver diseases in the United States from 1988 to 2008. *Clin. Gastroenterol. Hepatol.* **9**, 524-530.e1; quiz e60 (2011).
7. Pais, R. *et al.* NAFLD and liver transplantation: Current burden and expected challenges. *J. Hepatol.* **65**, 1245–1257 (2016).

8. Fischer, A. H., Jacobson, K. A., Rose, J. & Zeller, R. Hematoxylin and Eosin Staining of Tissue and Cell Sections. *Cold Spring Harb Protoc* **2008**, pdb.prot4986 (2008).
9. Molavi, D. W. *The practice of surgical pathology: A beginner's guide to the diagnostic process*. (Springer New York, 2008). doi:10.1007/978-0-387-74486-5.
10. Kleiner, D. E. & Makhlouf, H. R. Histology of NAFLD and NASH in Adults and Children. *Clin Liver Dis* **20**, 293–312 (2016).
11. Matteoni, C. A. *et al.* Nonalcoholic fatty liver disease: a spectrum of clinical and pathological severity. *Gastroenterology* **116**, 1413–1419 (1999).
12. Raab, S. S. The Cost-Effectiveness of Immunohistochemistry. *Archives of Pathology & Laboratory Medicine* **124**, 1185–1191 (2000).
13. Rivenson, Y. *et al.* PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning. *Light Sci Appl* **8**, 1–11 (2019).
14. Gurcan, M. N. *et al.* Histopathological Image Analysis: A Review. *IEEE Rev Biomed Eng* **2**, 147–171 (2009).
15. Komura, D. & Ishikawa, S. Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal* **16**, 34–42 (2018).
16. Ching Travers *et al.* Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface* **15**, 20170387 (2018).
17. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
18. Levy, J. J., Salas, L. A., Christensen, B. C., Sriharan, A. & Vaickus, L. J. PathFlowAI: A High-Throughput Workflow for Preprocessing, Deep Learning and Interpretation in Digital Pathology. *Pac Symp Biocomput* **25**, 403–414 (2020).

19. Wei, J. W. *et al.* Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific Reports* **9**, 3358 (2019).
20. Tomita, N. *et al.* Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides. *JAMA Netw Open* **2**, e1914645–e1914645 (2019).
21. Wei, J. W. *et al.* Evaluation of a Deep Neural Network for Automated Classification of Colorectal Polyps on Histopathologic Slides. *JAMA Netw Open* **3**, e203398–e203398 (2020).
22. Levy, J. J., Jackson, C. R., Haudenschild, C. C., Christensen, B. C. & Vaickus, L. J. PathFlow-MixMatch for Whole Slide Image Registration: An Investigation of a Segment-Based Scalable Image Registration Method. *bioRxiv* 2020.03.22.002402 (2020) doi:10.1101/2020.03.22.002402.
23. Levy, J., Jackson, C., Sriharan, A., Christensen, B. & Vaickus, L. Preliminary Evaluation of the Utility of Deep Generative Histopathology Image Translation at a Mid-sized NCI Cancer Center. in 302–311 (2020).
24. Jackson, C. Sox-10 Virtual Immunohistochemistry: An Application of Artificial Intelligence Using a Convolutional Neural Network. (2019).
25. Jackson, C. R., Sriharan, A. & Vaickus, L. J. A machine learning algorithm for simulating immunohistochemistry: development of SOX10 virtual IHC and evaluation on primarily melanocytic neoplasms. *Modern Pathology* 1–11 (2020) doi:10.1038/s41379-020-0526-z.

26. Lahiri, A., Ayush, K., Biswas, P. K. & Mitra, P. Generative Adversarial Learning for Reducing Manual Annotation in Semantic Segmentation on Large Scale Miscroscopy Images: Automated Vessel Segmentation in Retinal Fundus Image as Test Case. in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 794–800 (2017). doi:10.1109/CVPRW.2017.110.
27. Lei, Y. *et al.* MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Med Phys* **46**, 3565–3581 (2019).
28. Yi, X., Walia, E. & Babyn, P. Generative adversarial network in medical imaging: A review. *Medical Image Analysis* **58**, 101552 (2019).
29. Rivenson, Y. *et al.* Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature Biomedical Engineering* **3**, (2019).
30. Bayramoglu, N., Kaakinen, M., Eklund, L. & Heikkila, J. Towards Virtual H&E Staining of Hyperspectral Lung Histology Images Using Conditional Generative Adversarial Networks. in 64–71 (2017). doi:10.1109/ICCVW.2017.15.
31. Borhani, N., Bower, A. J., Boppart, S. A. & Psaltis, D. Digital staining through the application of deep neural networks to multi-modal multi-photon microscopy. *Biomed. Opt. Express, BOE* **10**, 1339–1350 (2019).
32. Quiros, A. C., Murray-Smith, R. & Yuan, K. Pathology GAN: Learning deep representations of cancer tissue. *arXiv:1907.02644 [cs, eess, stat]* (2019).
33. Rana, A., Yauney, G., Lowe, A. & Shah, P. Computational Histological Staining and Destaining of Prostate Core Biopsy RGB Images with Generative Adversarial Neural Networks. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* 828–834 (2018) doi:10.1109/ICMLA.2018.00133.

34. Xu, Z., Fernández Moro, C., Bozóky, B. & Zhang, Q. *GAN-based Virtual Re-Staining: A Promising Solution for Whole Slide Image Analysis*. (2019).
35. Bentaieb, A. & Hamarneh, G. Adversarial Stain Transfer for Histopathology Image Analysis. *IEEE Trans Med Imaging* **37**, 792–802 (2018).
36. Ghazvinian Zanjani, F., Zinger, S., Ehteshami Bejnordi, B., van der Laak, J. & With, P. Stain normalization of histopathology images using generative adversarial networks. in 573–577 (2018). doi:10.1109/ISBI.2018.8363641.
37. Pontalba, J. T. *et al.* Assessing the Impact of Color Normalization in Convolutional Neural Network-Based Nuclei Segmentation Frameworks. *Front. Bioeng. Biotechnol.* **7**, 300 (2019).
38. Wei, J. *et al.* Generative Image Translation for Data Augmentation in Colorectal Histopathology Images. *ArXiv* **abs/1910.05827**, (2019).
39. Bug, D. *et al.* Supervised and Unsupervised Cell-Nuclei Detection in Immunohistology. in (2019).
40. Mahmood, F. *et al.* Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images. *arXiv:1810.00236 [cs]* (2018).
41. Hollandi, R. *et al.* A deep learning framework for nucleus segmentation using image style transfer. *bioRxiv* 580605 (2019) doi:10.1101/580605.
42. Harnois, F. *et al.* Prevalence and Predictive Factors of Non-Alcoholic Steatohepatitis (NASH) in Morbidly Obese Patients Undergoing Bariatric Surgery. *OBES SURG* **16**, 183–188 (2006).
43. Chitturi, S. *et al.* NASH and insulin resistance: Insulin hypersecretion and specific association with the insulin resistance syndrome. *Hepatology* **35**, 373–379 (2002).

44. Angulo, P., Keach, J. C., Batts, K. P. & Lindor, K. D. Independent predictors of liver fibrosis in patients with nonalcoholic steatohepatitis. *Hepatology* **30**, 1356–1362 (1999).

45. Kim, W. R. *et al.* Hyponatremia and mortality among patients on the liver-transplant waiting list. *N. Engl. J. Med.* **359**, 1018–1026 (2008).

46. Vallet-Pichard, A. *et al.* FIB-4: an inexpensive and accurate marker of fibrosis in HCV infection. comparison with liver biopsy and fibrotest. *Hepatology* **46**, 32–36 (2007).

47. Shah, A. G. *et al.* Comparison of noninvasive markers of fibrosis in patients with nonalcoholic fatty liver disease. *Clin. Gastroenterol. Hepatol.* **7**, 1104–1112 (2009).

48. Sterling, R. K. *et al.* Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology* **43**, 1317–1325 (2006).

49. Freeman, R. B. *et al.* The new liver allocation system: moving toward evidence-based transplantation policy. *Liver Transpl.* **8**, 851–858 (2002).

50. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv:1611.07004 [cs]* (2018).

51. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv:1703.10593 [cs]* (2018).

52. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci Rep* **7**, 1–7 (2017).

53. OpenSeadragon. http://openseadragon.github.io/.

54. Kleiner, D. E. *et al.* Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* **41**, 1313–1321 (2005).

55. Goodman, Z. D. Grading and staging systems for inflammation and fibrosis in chronic liver diseases. *Journal of Hepatology* **47**, 598–607 (2007).

56. Han, C. *et al.* GAN-based synthetic brain MR image generation. in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* 734–738 (2018). doi:10.1109/ISBI.2018.8363678.

57. Chuquicusma, M. J. M., Hussein, S., Burt, J. & Bagci, U. How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* 240–244 (IEEE, 2018). doi:10.1109/ISBI.2018.8363564.

58. Gohorbani, A., Natarajan, V., Coz, D. D. & Liu, Y. DermGAN: Synthetic Generation of Clinical Skin Images with Pathology. in (2019).

59. Hallsworth, K. & Adams, L. A. Lifestyle modification in NAFLD/NASH: Facts and figures. *JHEP Rep* **1**, 468–479 (2019).

60. Ferrell, L. Liver Pathology: Cirrhosis, Hepatitis, and Primary Liver Tumors. Update and Diagnostic Problems. *Modern Pathology* **13**, 679–704 (2000).

61. Juluri, R. *et al.* Generalizability of the NASH CRN Histological Scoring System for Nonalcoholic Fatty Liver Disease. *J Clin Gastroenterol* **45**, 55–58 (2011).

62. Pournik, O. *et al.* Inter-observer and Intra-observer Agreement in Pathological Evaluation of Non-alcoholic Fatty Liver Disease Suspected Liver Biopsies. *Hepat Mon* **14**, (2014).

63. Barr, P. J. *et al.* The psychometric properties of Observer OPTION(5), an observer measure of shared decision making. *Patient Educ Couns* **98**, 970–976 (2015).

64. Bobak, C. A., Barr, P. J. & O'Malley, A. J. Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. *BMC Med Res Methodol* **18**, 93 (2018).

65. Poynard, T. *et al.* A comparison of fibrosis progression in chronic liver diseases. *J. Hepatol.* **38**, 257–265 (2003).

66. Boettiger, C. An Introduction to Docker for Reproducible Research. *SIGOPS Oper. Syst. Rev.* **49**, 71–79 (2015).

67. Amstutz, P. *et al.* Common Workflow Language, v1.0. (2016) doi:10.6084/m9.figshare.3115156.v2.

68. Heinemann, F., Birk, G. & Stierstorfer, B. Deep learning enables pathologist-like scoring of NASH models. *Scientific Reports* **9**, 18454 (2019).