Automatic quantification of lobular inflammation and hepatocyte ballooning in NAFLD liver biopsies.

Scott Vanderbeck MS[1], Joseph Bockhorst PhD[1], David Kleiner MD, PhD[2], Richard Komorowski MD[3], Naga Chalasani MD[4], Samer Gawrieh MD[4]

[1]Department of Electrical Engineering and Computer Science, University of Wisconsin, Milwaukee, WI.

[2] Laboratory of Pathology, National Cancer Institute, Bethesda, MD

[3] Department. of Pathology, Medical College of Wisconsin, Milwaukee, WI

[4] Division of Gastroenterology and Hepatology, Indiana University School of Medicine

Running Title: Automatic Classification of lobular inflammation and ballooning in NAFLD liver biopsy.

Keywords: fatty liver, lobular inflammation, hepatocyte ballooning, machine learning, NAFLD activity score, digital image analysis.

This is the author's manuscript of the article published in final edited form as:
Vanderbeck, S., Bockhorst, J., Kleiner, D., Komorowski, R., Chalasani, N., & Gawrieh, S. (2015). Automatic quantification of lobular inflammation and hepatocyte ballooning in nonalcoholic fatty liver disease liver biopsies. Human pathology, 46(5), 767-775. http://dx.doi.org/10.1016/j.humpath.2015.01.019

Correspondence Author:

Samer Gawrieh, MD

Division of Gastroenterology and Hepatology

Indiana University School of Medicine

702 Rotary Circle

Indianapolis, IN 46202

Email: sgawrieh@iu.edu

Phone (317) 278 9326

Fax: (317) 278 6870

Funding statement: This research was supported by the University of Wisconsin-Milwaukee Research Foundation (JB and SG) grant number PRJ-39IB and in part by the Intramural Research Program of the NIH, National Cancer Institute (DK)

Scott Vanderbeck is owner and founder of Organic Research Corp. (ORC), a startup funded in part by a grant from the University of Wisconsin-Extension Ideadvance Seed Fund through its partnership with the WI Economic Development Corporation and the University of Wisconsin System. The work contained herein was completed in its entirety while Vanderbeck was a student at University of Wisconsin - Milwaukee, prior to founding ORC.

Abstract

Automatic quantification of cardinal histological features of non-alcoholic fatty liver disease (NAFLD) may reduce human variability and allow continuous rather than semi-quantitative assessment of injury. We recently developed an automated classifier that can detect and quantify macro-steatosis with ≥ 95% precision and recall (sensitivity). Here we report our early results on the classifier's performance in detecting lobular inflammation and hepatocellular ballooning. Automatic quantification of lobular inflammation and ballooning was performed on digital images of H&E stained slides of liver biopsy samples from 59 individuals with normal liver histology and varying severity of NAFLD. Two expert hepatopathologists scored liver biopsies according the Nonalcoholic Steatohepatitis Clinical Research Network scoring system and provided annotations of lobular inflammation and hepatocyte ballooning on the digital images. The classifier had precision and recall of 70% and 49% for lobular inflammation, and 91% and 54% for hepatocyte, ballooning. In addition, the classifier had an AUROC of 95% for lobular inflammation and 98% for hepatocyte ballooning. The Spearman rank correlation coefficient for comparison with pathologist grades was 45.2% for lobular inflammation and 46% for hepatocyte ballooning. Our novel observations demonstrate that automatic quantification of cardinal NAFLD histological lesions is feasible and offer promise for further development of automatic quantification as a potential aid to pathologists evaluating NAFLD biopsies in clinical practice and clinical trials.

1. Introduction

Non-alcoholic fatty liver disease (NAFLD) is the most common liver disease in the U.S. affecting 1 in 3 adults, and 1 in 8 children (1, 2). The most severe phenotype of the disease, non-alcoholic steatohepatitis (NASH), is estimated to affect 3-5% of the U.S. population (3, 4). The spectrum of NAFLD begins with a mild phenotype, simple steatosis where only steatosis is present in the liver, and extends to NASH where steatosis is present with hepatic necro-inflammation and fibrosis (5). Liver biopsy is the current "gold standard" diagnostic test for phenotyping NAFLD (5). Accurate phenotyping of NAFLD is critical because simple steatosis rarely progresses, whereas NASH can progress to cirrhosis, liver failure, and hepatocellular carcinoma (6-9).

The NAFLD Activity Score (NAS), the state-of-the-art scoring system for liver biopsies, is based on the sum of three numerical grades determined by manual pathologist assessment and semi-quantification of steatosis, lobular inflammation, and hepatocyte ballooning (10). These three lesions were selected for inclusion in the NAS based on a multiple logistic regression analysis that showed these lesions were independently associated with diagnosis of NASH (10). Because these lesions are potentially reversible in the short term, unlike fibrosis, they were chosen as end-points for therapeutic trials for NASH. A recent expert panel report recommended use of liver biopsy to define histological outcomes in phase 2 and 3 clinical trials in NASH and also recommended the use of NAS to define and quantify NAFLD activity (11).

Semi-quantitative assessment of steatosis, lobular inflammation and hepatocyte ballooning has a couple important limitations stemming from the very nature of semi-quantitative grading that forces continuous measures to be threshold into discrete grading bins. The first limitation is semi-quantitative grades may fail to accurately show improvements. For example, a steatosis grade 0 implies 0-4% steatosis, whereas grade 1 implies 5-33% steatosis. Consider patient-A

which enters a study with 7% steatosis, and improves by 3%. With this, patient-A improves from steatosis grade 1 to 0. Now consider patient-B who enters a study with 30% steatosis and improves 24%. Patient-B is a steatosis grade 1 before and at the end of the study. Looking at study results one may be led to believe patient-A's 3% improvement was more significant than patient-B's 24% improvement. The second limitation is semi-quantitative grading scale inevitably leads to inter and intra-rater variability (10, 12-17). Rater variability will be amplified for cases that lie near a grading cutoff (threshold), and may worsen based on the skills and/or training of the rater.

We hypothesize that automated decision support tools for pathologists, by offering a continuous rather than semi-quantitative method for grading the histological lesions of NAFLD, could increase the precision and accuracy of grading histological activity. The aim of this initial study is to determine if an automated tool utilizing supervised machine learning could be trained by pathologists to detect lobular inflammation and hepatocyte ballooning. Our group has previously published research demonstrating the feasibility of the accurate categorization of the white regions in liver biopsy images including macro-steatosis, central veins, portal veins, portal arteries, sinusoids and bile ducts (18). To date, no previous work has set out to automatically quantify lobular inflammation and hepatocyte ballooning.

2. Materials and Methods

The analysis discussed herein, is based on a dataset of 59 unique liver biopsy scans. Of the 59 patients in the study, pathologist semi-quantitative grading was available for 47 patients with the remaining image scans being used solely for annotations and machine learning. Two study pathologists (DK, RK) provided semi-quantitative grades for each of the key histological lesions comprising the NAS (steatosis, lobular inflammation and hepatocyte ballooning). The patients in

the study represented the full range of phenotypes from patients not having NAFLD, to those with various stages of NAFLD.

High resolution digital images of the hematoxylin and eosin (H&E)–stained slides of liver biopsy images in the study were generated using the NanoZoomer scanner manufactured by Hamamatsu and housed in the Medical College of Wisconsin (MCW) pathology department. The images were scanned at 20X magnification and saved as RGB images in the lossless tiff file format. To create files small enough to efficiently work with given available hardware and computing resources, the files were reduced to 50% their original size. The smaller files are saved as RGB jpeg images with an 80% compression factor. The size reduction was performed with bicubic interpolation and antialiasing to preserve as much of the original image detail as possible. This resulted in a resolution of 0.92 microns per pixel with respect to actual tissue size. The research protocol was reviewed and approved by the Internal Review Board of the Medical College of Wisconsin.

2.1 Quantification of lobular inflammation and hepatocyte ballooning in biopsy sections:
Biopsy images are tiled into individual 25 pixel square sections. Tile size was selected by trial and error and by visual inspection of tile size appropriate with respect to lesion size. While there may be a more optimal tile size, our intent here is to show feasibility rather than develop a model intended for any specific use. Once an image is tiled, each tile is then assigned a probability of containing lobular inflammation and then automatically classified using a probability threshold as either containing lobular inflammation or not. The area of tiles classified as lobular inflammation versus the total area of the biopsy section is then computed to approximate the overall percentage of lobular inflammation. An identical process is also performed for quantifying the incidence of hepatocyte ballooning. Figure 1 pictorially demonstrates the process of classifying tiles for hepatocyte ballooning.

Our study pathologists used a custom built web-based Java Applet to manually annotate 138 areas of lobular inflammation and 48 areas of hepatocyte ballooning on biopsy images. In addition, study pathologists annotated 291 regions of fibrosis, 128 regions of portal inflammation, and 1969 types of white regions inclusive of macro-steatosis, central veins, portal veins, portal arteries, sinusoids and bile ducts. Lobular inflammation and hepatocyte ballooning annotations are available as bounded polygons. These regions serve as the positive class for learning data in each of their respective learning tasks. It was also necessary to establish a negative class for learning. To accomplish this, two different types of negative regions were developed:

1. Regions, excluding the positive class of interest (i.e. lobular inflammation or hepatocyte ballooning depending on the task). This includes macro-steatosis, central veins, portal veins, portal arteries, sinusoids and bile ducts, portal inflammation, fibrosis and a generic "other" class.
2. Ten randomly selected tiles from each image. While there is no guarantee a randomly selected tile does not contain lobular inflammation (or hepatocyte ballooning), even in cases with high incidence of these lesions, there is a high probability a randomly selected tile does not contain the lesion.

Because lobular inflammation and hepatocyte ballooning were annotated as bounded polygons, it was necessary to convert the polygon to a tile similar to what is used to quantify total incidence of lobular inflammation or hepatocyte ballooning. Figure 2 shows the feature process with shaded areas representing the tile image features are extracted for. The first is a tile centered on the polygon's centroid (Figure 2b). The second is a tile randomly offset from the polygon centroid (2c). The motivation behind splitting polygon annotations into two different

unique tiles is to first capture what a tile looks like should it fall directly on the lesion, and second to capture what it looks like if only part of the feature falls within a tile.

For each positively and negatively labeled region, the types of features used by the classifier for learning are:

- **Texture** - Texture and histogram statistics are computed for the gray scale region at each sigma level (19).
- **Gray level Co-Occurrence Matrix (GLCM)** - The co-occurrence matrix is computed for pixels in the each region (20).
- **GLCM Statistics** – Statistical measures related to the GLCM, such as contrast and correlation.
- **N-jet** - For each sigma level in our scale representation we compute the 2-jet of the region and extract related statistics (21).
- **Nuclear Density** - For each region, the mean, min, max and standard deviation of nuclear density are used as features (see discussion below).

Accuracy of the classifier is measured in 2 ways. First, a data set consisting of positive and negative learning tiles is analyzed using a ten-fold cross validation to gauge overall accuracy of the classifier (22). Cross fold validation entails taking our dataset of positive and negative tiles and splitting the dataset into ten subsets, and then running experiments where a model is learned from data in nine of the subsets and tested against data in the tenth. To reduce variability the experiment is repeated ten times with each subset serving as the “tenth” test subset exactly once. Results are then aggregated across all ten experiments. Second, entire images are tiled into individual sections and the total area of tiles classified as lobular inflammation or hepatocyte ballooning versus the total area of the biopsy section is computed and correlated with pathologists’ semi-quantitative grades.

2.2 Nuclear Density

Lobular inflammation is most visible by the presence of the nuclei of inflammatory cells. As inflammatory cells are smaller than other nearby cells, the number of nuclei in inflamed areas is higher. To quantify this, a measure was established called Nuclear Density. The nuclear density metric for a given pixel P is measured as the number of pixels within a fixed radius of P that are also part of a nucleus. We hypothesized this metric would serve as a good proxy for quantifying inflammation as a higher concentration of inflammatory cells should yield a higher number of nuclei in a region, and consequently a higher nuclear density.

The first step towards calculating nuclear density was to develop a process for isolating cell nuclei. Based on the hemotoxyphilic staining characteristics of nuclei, steps are taken to threshold nuclei from the biopsy images (23). Once the nuclei are extracted, it is possible to compute nuclear density. Nuclear Density for a pixel (x,y) and a surrounding radius r is defined as:

$$NuclearDensity(x, y, r) = \sum_{(i,r)\in R} f(i, j)$$

where R represents the set of pixels within radius r of (x, y) and f (i, j) = 1 if a pixel is identified as nuclei, and 0 otherwise.

Using this equation for nuclear density, it is possible to calculate the nuclear density metric for all pixels in an image. Figure 3 shows a pictorial representation of the nuclear density calculations for an image. The nuclear density image clearly shows a high concentration of nuclei in this case caused by portal inflammation, stromal cells and other portal structures. While nuclear density statistics for an entire tissue section correlate with pathologist lobular inflammation grades, better concordance is obtained using the nuclear density measure as a feature for supervised learning experiments.

2.3 Precision and Recall

The model's performance for detecting lobular inflammation and hepatocyte ballooning is measured by calculating the precision and recall (specificity) rates. Precision (also known as "Positive Prediction Rate") is a measure of the model's positive predictive ability. Specifically, we are measuring what percent of tiles classified as containing a given lesion type are correct. For both lobular inflammation and hepatocyte ballooning, precision is measured as:

$$Precision = \frac{true\ positive\ tile\ classifications}{true\ + false\ positive\ tile\ classifications}$$

Recall is the fraction of all positive tiles that are detected for each lesion type. For lobular inflammation, recall is the percent of all tiles that actually contain lobular inflammation that are correctly identified. Mathematically, recall is measured as:

$$Recall = \frac{true\ positive\ tile\ classifications}{true\ positive + false\ negative\ tile\ classifications}$$

3. RESULTS

3.1 Histological characteristics of subjects

Pathologist scored H&E liver biopsy slides from 47 (20 with normal liver histology and 27 with NAFLD of varying severity) of the 59 total patients in our dataset according to the NAS scoring system (10). The remaining 12 image scans were used only for annotations and machine learning. In the NAFLD group, 19 subjects had simple steatosis and 8 had NASH. Figure 4 shows the pathologist grading distribution of lobular inflammation and hepatocyte ballooning

amongst the 47 patients. Across our data set, a total 138 areas of lobular inflammation and 48 areas of hepatocyte ballooning were annotated on biopsy images.

As shown in Figure 4, our data set is skewed towards cases with minimal findings of lobular inflammation and hepatocyte ballooning. This does not present an immediate problem for our analysis as our focus is on the precision and recall (sensitivity) of our model to correctly classify individual image tiles. Additional data would, however, be needed across the full spectrum of lobular inflammation and hepatocyte ballooning grades to draw more meaningful conclusions about the impact of false positives and false negatives on quantifying an entire tissue sample.

3.2 Automatic Quantification of Lobular Inflammation

Evaluation of lobular inflammation classification is carried out using a 10-fold cross validation experiment

It is important to point out that our dataset of tiles is heavily skewed towards negative examples (tiles without lobular inflammation). Experiments were intentionally designed with a large skew towards tiles without lobular inflammation as even in patients with high incidence of lobular inflammation, the number of tiles without lobular inflammation would significantly outnumber those with lobular inflammation. The model classifies lobular inflammation with a 0.70 precision and 0.49 recall (sensitivity).

In the cross validation experiment, 95.6% of tiles were classified correctly. This represents a statistically significant improvement over the baseline accuracy of 94.0% (p-value <0.001) obtainable by always predicting not lobular inflammation. While the accuracy metric is not a clinically meaningful, it demonstrates the improvements over naïve baseline methods based on the predictive power of the model.

The recall-precision curve is shown in Figure 5 along with the ROC curve for the experiment. Examination of both of these curves to evaluate classifier performance is important as our data set is largely skewed towards negative non-lobular inflammation examples (24). Namely, with a ROC curve the goal is to be towards the “upper left” of the curve, while with precision-recall curves the goal is to be near the “upper right”. The ROC curve has a large area under the curve (AUROC) of 0.946 indicating the model has a strong ability to discriminate between tiles with lobular inflammation and those without.

While the immediate focus of our research was the accurate identification of individual tiles of lobular inflammation, we also sought to gauge model performance by measuring how it compared to pathologist grades. For each patient, the overall percentage of tissue with lobular inflammation was calculated by taking the total area of tiles classified as having lobular inflammation and dividing by total tissue area. The motivation for this metric in our analysis is tiles containing lobular inflammation would approximately represent 1 focus of lobular inflammation, and the metric would therefore be proportional to the number of lobular inflammation foci per unit area of tissue. To obtain percentage lobular inflammation, a different model was created for each patient using only training data from other patients, otherwise known as a leave-one-sample out approach.

There is a general concordance between the modeled percent lobular inflammation and pathologist grade. The four patients that received the highest pathologist grades all rank near the top of the model. Conversely, those with the lowest grade score near the bottom (Figure.6). Case FLE038 stands out as an outlier in the analysis. Examination of the case revealed this tissue sample is considerably smaller by surface area compared to the other samples in our study. In fact the FLE038 sample was ~ 75% smaller than the mean size of all samples. As

evident in this case, a smaller tissue section would be far more susceptible to the impact of false positive tiles. The overall spearman rank correlation for the comparison between the model and average of the pathologists score is 0.452 with a p-value of 0.002.

3.3 Automatic Quantification of Hepatocyte Ballooning

Similar to lobular inflammation, evaluation of the hepatocyte ballooning classifier was carried out using a 10-fold cross validation experiment. The model classified hepatocyte ballooning with 0.91 precision and 0.54 recall. As with lobular inflammation, the dataset is skewed heavily towards examples without hepatocyte ballooning. Based on this a baseline accuracy of 97.9% would be obtainable simply by always predicting not hepatocyte ballooning for every tile. In our model, 98.9% of regions were correctly classified. This is a statistically significant (p-value <0.001) improvement over the baseline method. As with lobular inflammation the high level of accuracy has no clinical meaning, but it does demonstrate the predictive ability of the model and show a gain over a naïve baseline method.

With the dataset skewed heavily towards examples without hepatocyte ballooning, classifier performance must be evaluated through examination of both the recall-precision curve and the ROC curve (Figure.7). The ROC curve shows the model has a strong ability (AUROC of 0.983) to discriminate between tiles as having or not having a ballooned hepatocytes. Similarly, the precision-recall curve shows a very high precision rate of 90% is obtainable while still recalling more than 50% of all tiles containing ballooned cells. This confirms that the model performs well despite the skew towards negative instances.

The next step in the analysis was to examine the concordance of a continuous metric derived from model predictions of hepatocyte ballooning with scores provided by our expert pathologists. For each patient, we took the surface area of tiles classified as containing

hepatocyte ballooning and divided by total tissue area. This metric should be approximately equal to the percentage of tissue area containing hepatocyte ballooning if all tiles are correctly classified. The model obtained a spearman rank correlation of 0.460 and a p-value = 0.001 with our pathologist grades.

Figure 8 shows the results of each individual case. The chart shows a good relationship between average pathologist grade and the computed percentage of ballooning with cases that received a higher pathologist grade typically receiving a higher computed percentage ballooning. This is particularly prevalent on the two cases that received the highest grades from study pathologists. FLE008 received a score of 2 from both RK and DK, and FLE029 received a score of 0 and 2 from RK and DK, respectively. These two patients received the 2nd and 3rd highest overall percentage ballooning from the model. Case FLE021 presents as an obvious outlier. Examination of the case more closely revealed the model was misclassifying glycogenosis as hepatocyte ballooning. There are several cases that received a grade 0 from both study pathologists but where the model detected some hepatocyte ballooning. These cases are all the result of 1-3 tiles in the entire image being false positives. As the model had a far higher precision (positive predictive rate) than recall (sensitivity) of tiles, the impact of just a few false positives is seen in these cases.

4. DISCUSSION

The results of this study demonstrate that it is feasible to develop a system using supervised machine learning to automatically quantify two of the cardinal features needed to phenotype NAFLD, lobular inflammation and hepatocyte ballooning. These findings are significant as more accurate continuous measurements are more desirable than semi-quantitative scores to measure NAFLD activity and quantify patient’s response to therapeutics used in trials or patient care.

While the overall test statistics for correlation with pathologist grade are not as high as those our research group has shown for steatosis grading (12, 18), they show a general concordance between the model scores and pathologist grades and demonstrate the feasibility of such an approach. While our ultimate goals are to replace discrete grades with continuous measures of lesions, we thought it important to show the general relationship between continuous measures and pathologist grade. It is important to note that the continuous metrics we used in our analysis are different than the metrics used by pathologists for semi-quantitative grading. Both our continuous metric and pathologist grade should, however increase with lesion prevalence. Correlation coefficients may also be of limited meaning as we are correlating 47 continuous values to four discrete bins of average pathologist grade. Further, our data set is skewed towards patients with minimal incidence of lobular inflammation and hepatocyte ballooning so additional research is needed to examine model performance on more severe cases.

Due to the large skew in the dataset of tiles without lobular inflammation or hepatocyte ballooning, a large increase in the number of false positives (i.e. incorrect predictions of lobular inflammation or hepatocyte ballooning) would have a minimal impact on the false positive rate. This is because the metric is calculated as (false positives) / (false positives + true negatives), and the true negatives (correct predictions of not lobular inflammation or hepatocyte ballooning) in the denominator will dominate the metric. Based on this, one must also look to the recall-precision curve to evaluate the classifier. Specifically, the precision rate is not susceptible to the large skew of negative examples in the data set. Examination of the recall-precision curve shows predictions of lobular inflammation may be made with approximately 65% precision, while still recalling 50%+ of all lobular inflammation. A decrease in recall rate should be an acceptable tradeoff for increased precision in this experiment, provided recall is consistent across patients from different labs, cutting and staining procedures, etc. In other words, if

patient A has more lobular inflammation than patient B, identifying 50% of patients A's lobular inflammation will still quantify to a higher score than recalling 50% of patient B's.

Examination of the classifier results showed areas with glycogen and/or many fat droplets confusing the detection of hepatocyte ballooning. Example 2 of Figure 1 shows such a case. In Example 2, a ballooned cell adjacent to small fat droplets(upper-left of Figure 1, Example 2) is correctly identified. However, another ballooned cell located in between the three steatotic cells with large fat droplets at the bottom right of the example is missed. The model assigned this tile 23% chance of containing hepatocyte ballooning demonstrating that the model detected some ballooning activity, however, this fails to meet the threshold for being considered a positive detection of hepatocyte ballooning. Examination of the tile probabilities of Example 2 also shows an overall increase in the probability of ballooning activity in tiles where no ballooning exists presumably do to the appearance of the cytoplasm. Future models should include more training examples of both the positive and negative ballooning class in areas of glycogen or high steatotic cells incidence to afford the classifier a richer training set to make correct predictions in these areas. In practice, a tool could be developed that pre-sorts tiles by probability allowing an interactive step whereby a pathologist makes the final determination.

The continuous measures used herein are based on surface area of tiles with lesions divided by the total surface area of tissue. This metric can likely be improved by modifying the denominator. Specifically, rather than total surface area of tissue, it may be desirable to use total surface area of tissue excluding portal regions. This may more accurately reflect disease activity in the regions of primary interest for a given lesion.

No previous efforts have been published attempting to automatically quantify lobular inflammation and hepatocyte ballooning in images of scanned NAFLD liver biopsy sections.

Automatic quantification of lobular inflammation and hepatocyte ballooning may provide a means for reducing the inherent human variability in semi-quantitative assessment of NAFLD histology and provide pathologists a reliable tool for measuring NAFLD lesions on a continuum. Furthermore, continuous accurate measurement of NAFLD cardinal histological features, such as steatosis, lobular inflammation, hepatocyte ballooning and portal inflammation, is highly desirable in assessing NAFLD disease activity and monitoring response to therapeutic interventions in clinical trials.

In summary, this is the first study showing that automatic quantification of lobular inflammation and hepatocyte ballooning is feasible in digital images of liver biopsies from patients with NAFLD. We are currently conducting studies to optimize and validate the performance of our automated classifier and are also developing algorithms for automated quantification of hepatic fibrosis. These studies will include exhaustive annotation of unseen test cases so the performance of our model may be reported on the ability to identify all lesions on a given tissue sample, rather than proxy metrics based on a subset of annotated tiles. These early findings offer promise for further development of automatic quantification as a potential aid to pathologists evaluating NAFLD biopsies in clinical practice and clinical trials.

REFERENCES

1. Browning JD, Szczepaniak LS, Dobbins R, Nuremberg P, Horton JD, Cohen JC, et al. Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. Hepatology. 2004;40(6):1387-95.
2. Schwimmer JB, Deutsch R, Kahen T, Lavine JE, Stanley C, Behling C. Prevalence of fatty liver in children and adolescents. Pediatrics. 2006;118(4):1388-93.
3. Vernon G, Baranova A, Younossi ZM. Systematic review: the epidemiology and natural history of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in adults. Alimentary Pharmacology & Therapeutics. 2011;34(3):274-85.
4. Chalasani N, Younossi Z, Lavine JE, Diehl AM, Brunt EM, Cusi K, et al. The diagnosis and management of non-alcoholic fatty liver disease: practice guideline by the American Gastroenterological Association, American Association for the Study of Liver Diseases, and American College of Gastroenterology. Gastroenterology. 2012;142(7):1592-609.

5. Brunt EM. Pathology of nonalcoholic fatty liver disease. Nature reviewsGastroenterology & hepatology. 2010;7(4):195-203.
6. Matteoni CA, Younossi ZM, Gramlich T, Boparai N, Liu YC, McCullough AJ. Nonalcoholic fatty liver disease: a spectrum of clinical and pathological severity. Gastroenterology. 1999;116(6):1413-9.
7. Hui JM, Kench JG, Chitturi S, Sud A, Farrell GC, Byth K, et al. Long-term outcomes of cirrhosis in nonalcoholic steatohepatitis compared with hepatitis C. Hepatology. 2003;38(2):420-7.
8. Ekstedt M, Franzen LE, Mathiesen UL, Thorelius L, Holmqvist M, Bodemar G, et al. Long-term follow-up of patients with NAFLD and elevated liver enzymes. Hepatology. 2006;44(4):865-73.
9. Sanyal AJ, Banas C, Sargeant C, Luketic VA, Sterling RK, Stravitz RT, et al. Similarities and differences in outcomes of cirrhosis due to nonalcoholic steatohepatitis and hepatitis C. Hepatology. 2006;43(4):682-9.
10. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology. 2005;41(6):1313-21.
11. Sanyal AJ, Brunt EM, Kleiner DE, Kowdley KV, Chalasani N, Lavine JE, et al. Endpoints and clinical trial design for nonalcoholic steatohepatitis. Hepatology. 2011;54(1):344-53.
12. Gawrieh S, Knoedler DM, Saeian K, Wallace JR, Komorowski RA. Effects of interventions on intra- and interobserver agreement on interpretation of nonalcoholic fatty liver disease histology. Annals of Diagnostic Pathology. 2011;15(1):19-24.
13. Juluri R, Vuppalanchi R, Olson J, Unalp A, Van Natta ML, Cummings OW, et al. Generalizability of the nonalcoholic steatohepatitis Clinical Research Network histologic scoring system for nonalcoholic fatty liver disease. Journal of clinical gastroenterology. 2011;45(1):55-8.
14. Ratziu V, Charlotte F, Heurtier A, Gombert S, Giral P, Bruckert E, et al. Sampling variability of liver biopsy in nonalcoholic fatty liver disease. Gastroenterology. 2005;128(7):1898-906.
15. Younossi ZM, Stepanova M, Rafiq N, Makhlouf H, Younoszai Z, Agrawal R, et al. Pathologic criteria for nonalcoholic steatohepatitis: interprotocol agreement and ability to predict liver-related mortality. Hepatology (Baltimore, Md). 2011;53(6):1874-82.
16. Fukusato T, Fukushima J, Shiga J, Takahashi Y, Nakano T, Maeyama S, et al. Interobserver variation in the histopathological assessment of nonalcoholic steatohepatitis. Hepatology research : the official journal of the Japan Society of Hepatology. 2005;33(2):122-7.
17. Merriman RB, Ferrell LD, Patti MG, Weston SR, Pabst MS, Aouizerat BE, et al. Correlation of paired liver biopsies in morbidly obese patients with suspected nonalcoholic fatty liver disease. Hepatology. 2006;44(4):874-80.
18. 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(4):785-92.
19. Lindeberg T. Scale-space theory: A basic tool for analysing structures at different scales. J of Applied Statistics. 1994;21((2)):224-70.
20. Tou JY, Tay, Y. H., & Lau, P. Y. Gabor filters and grey-level co-occurrence matrices in texture classification. MMU International Symposium on Information and Communications Technologies 2007:197-202.
21. Lindeberg T. Scale-Space. Wiley Encyclopedia of Computer Science and Engineering.2008. p. 2495.-504.
22. Picard RR, Cook, R.D. Cross-validation of regression models. Journal of the American Statistical Association. 1984;79(387):575-83.
23. Vanderbeck S. Automatic Quantification of the Histological Features in Liver Biopsy Images to Aid in the Diagnosis of Non-alcoholic Fatty Liver Disease. Diss University of Wisconsin--Milwaukee. 2011.

24. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learn-ing, ICML '06 New York, NY, USA ACM. 2006:233-40.

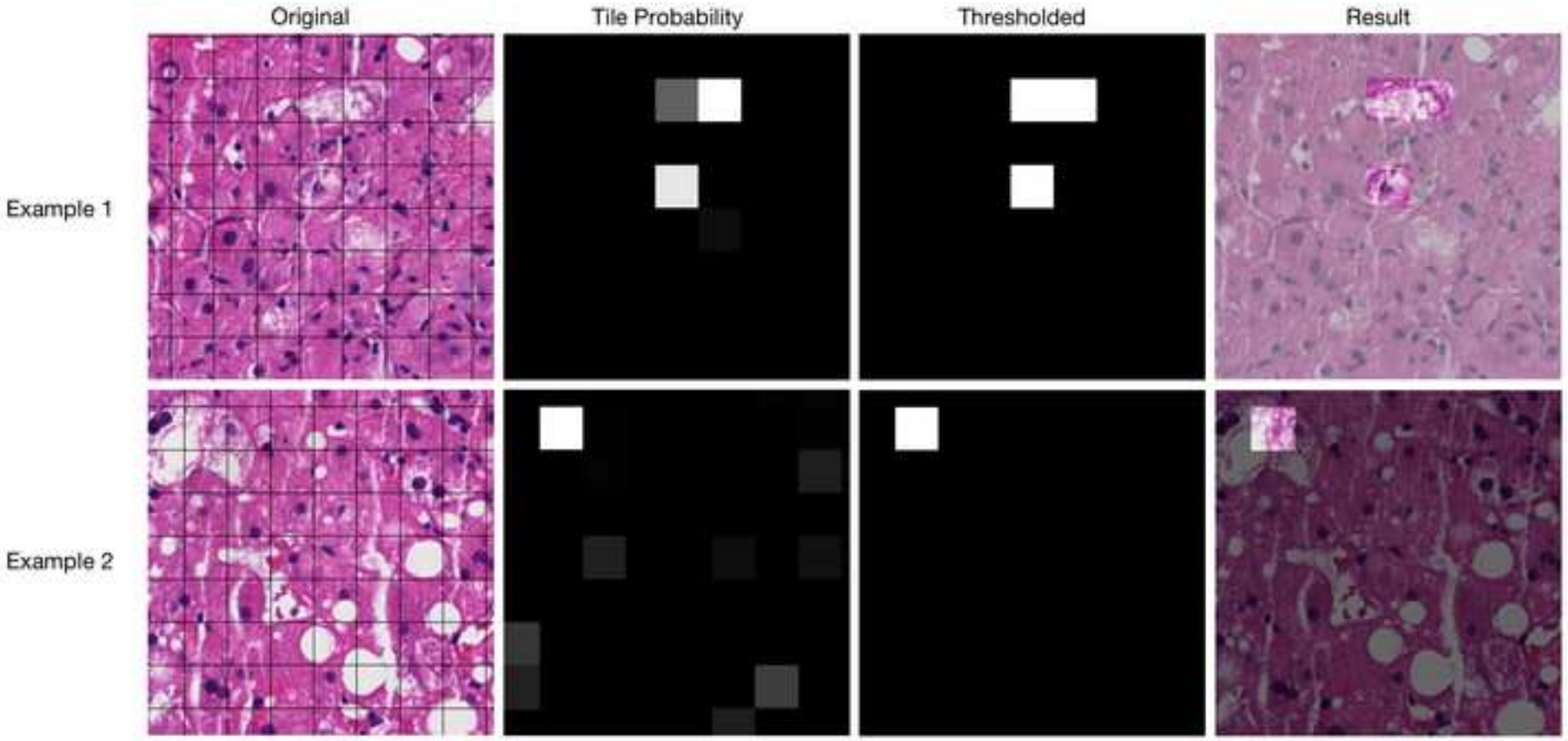
LIST OF FIGURES

Highlights

- We examine automated detection of lobular inflammation and hepatocyte ballooning on digital imaged of liver biopsy
- The models had precision and recall (sensitivity) of 70% and 49% for lobular inflammation
- The model had precision and recall (sensitivity) and 91% and 54% for hepatocyte, ballooning.
- We demonstrate that automatic quantification of cardinal NAFLD histological lesions is feasible.

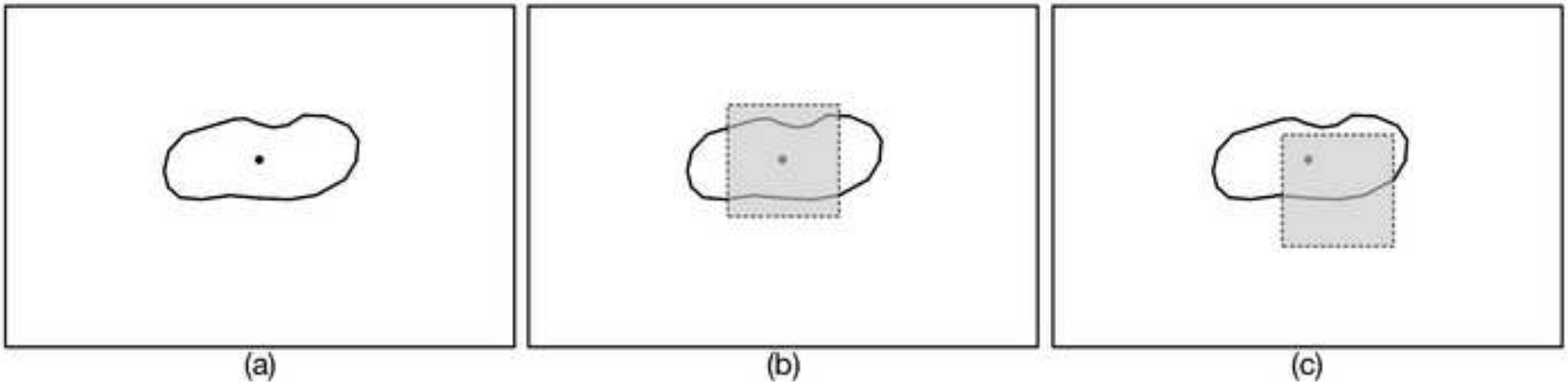
Figure(s)

Click here to download high resolution image



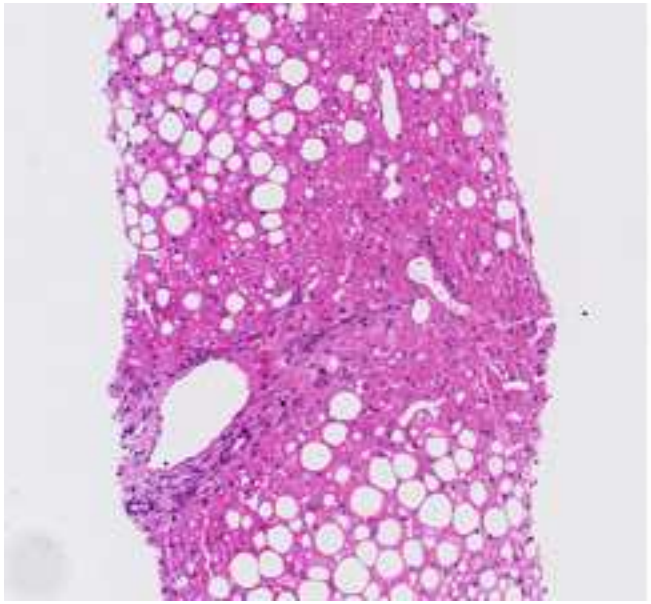
Figure(s)

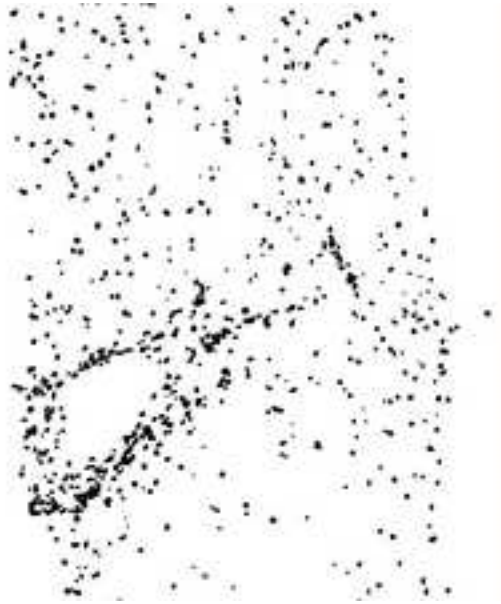
Click here to download high resolution image

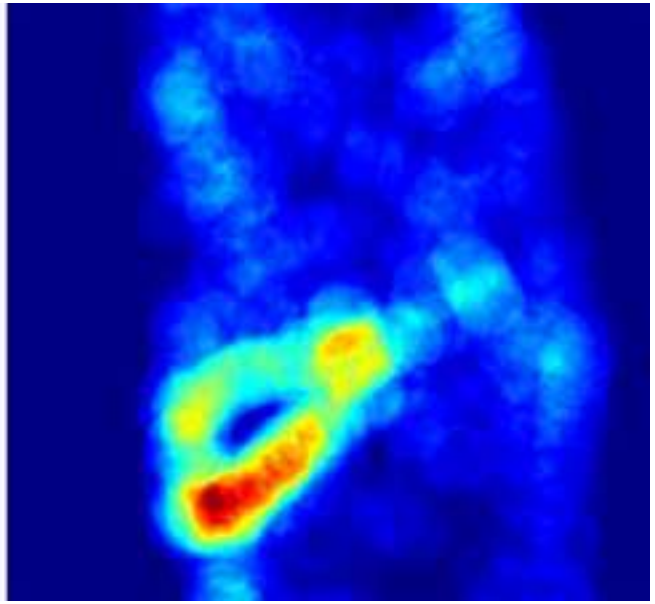


Figure(s)

Click here to download high resolution image

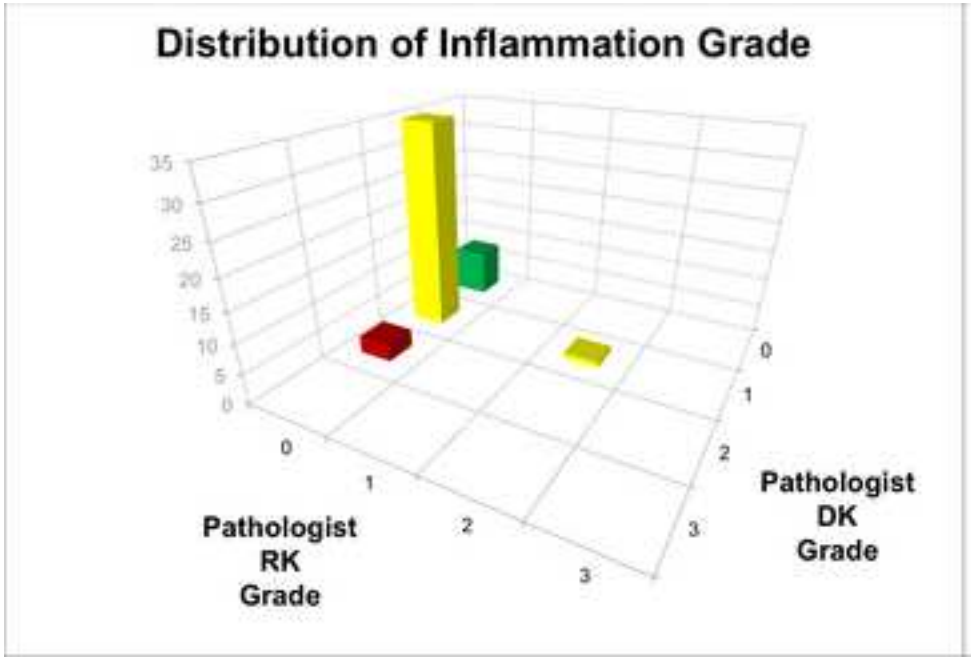


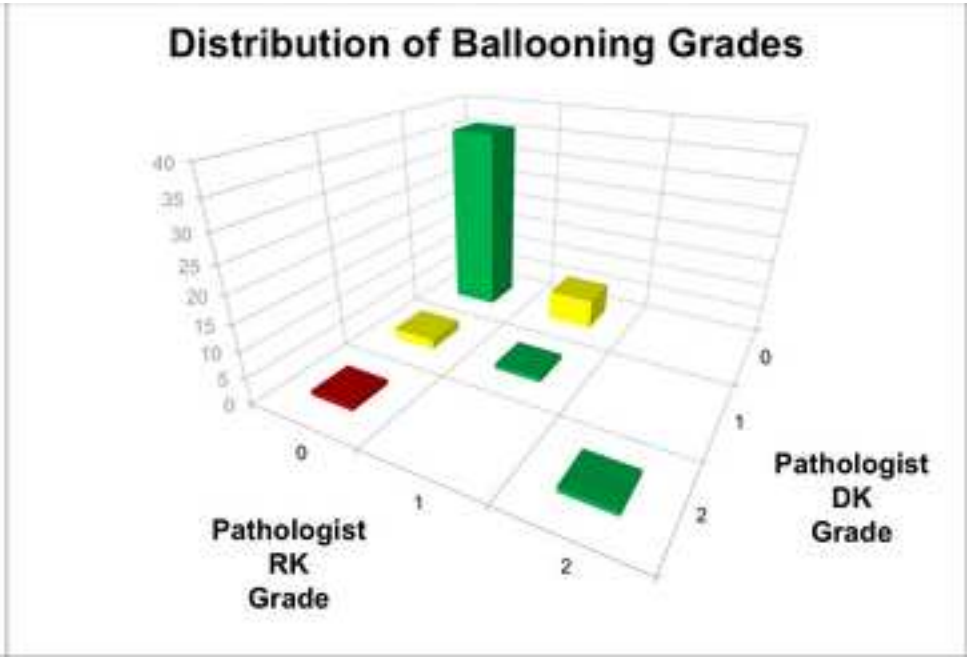




Figure(s)

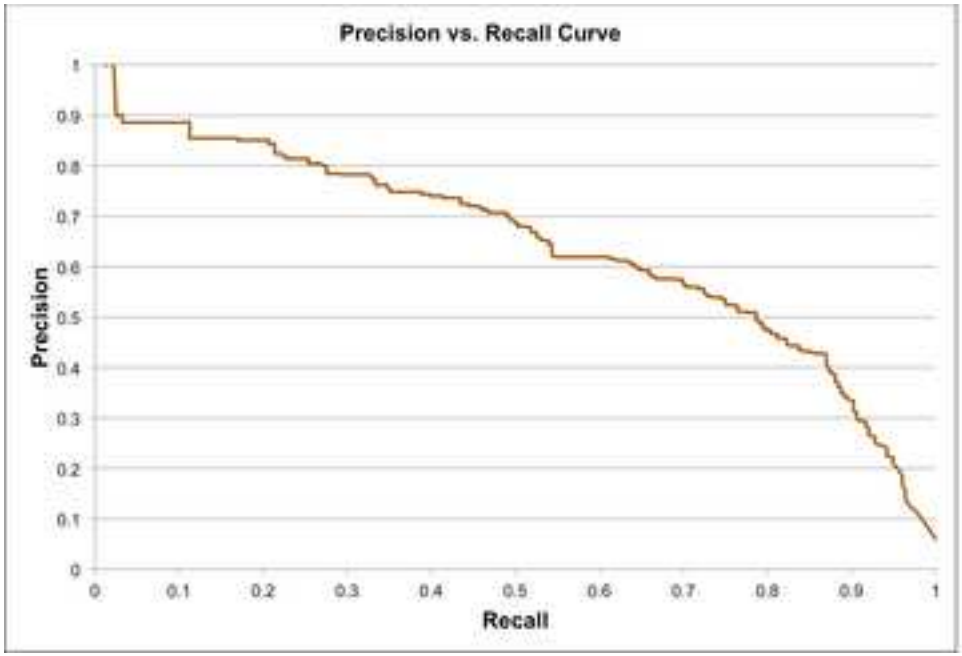
Click here to download high resolution image

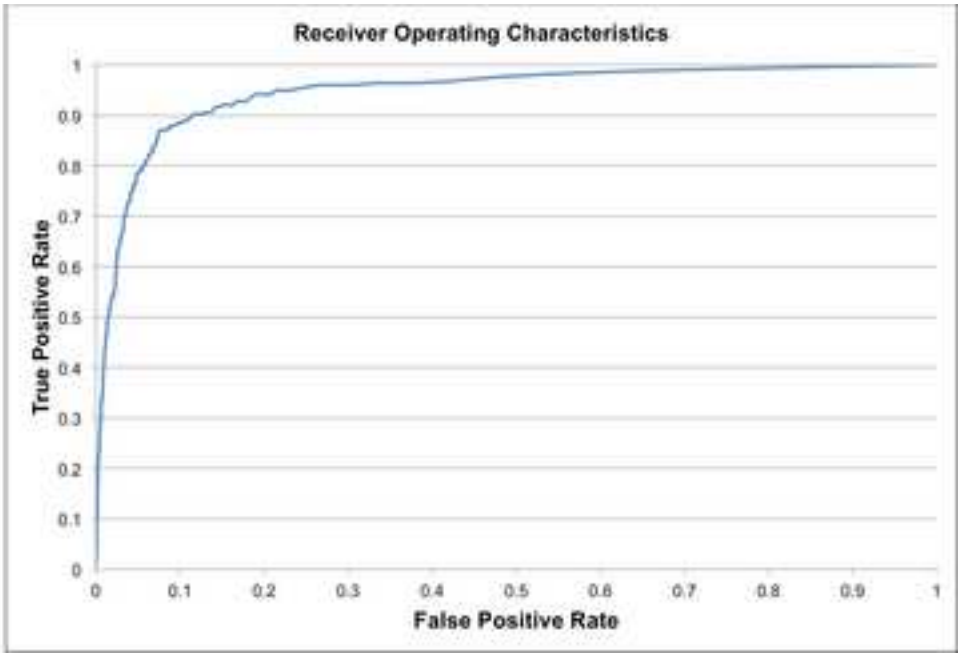




Figure(s)

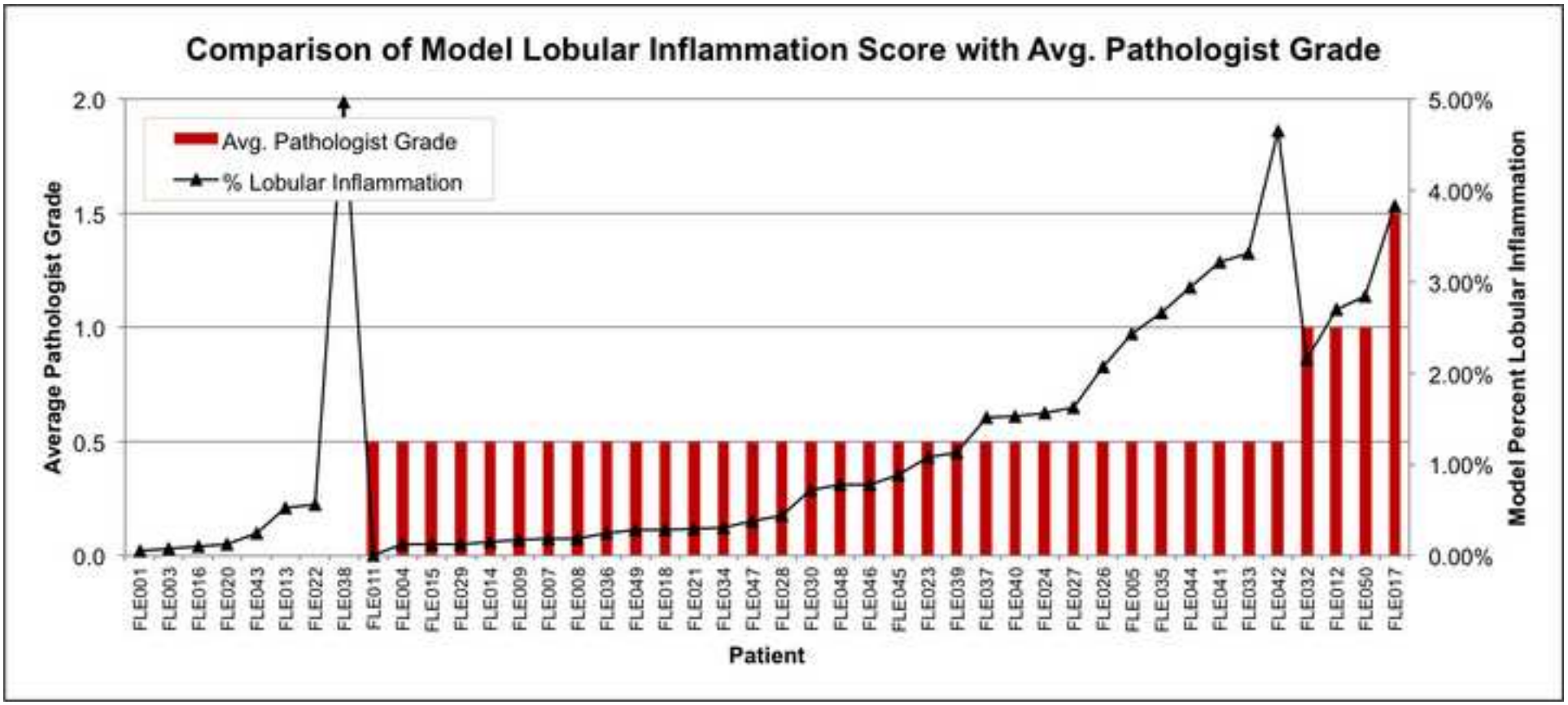
Click here to download high resolution image





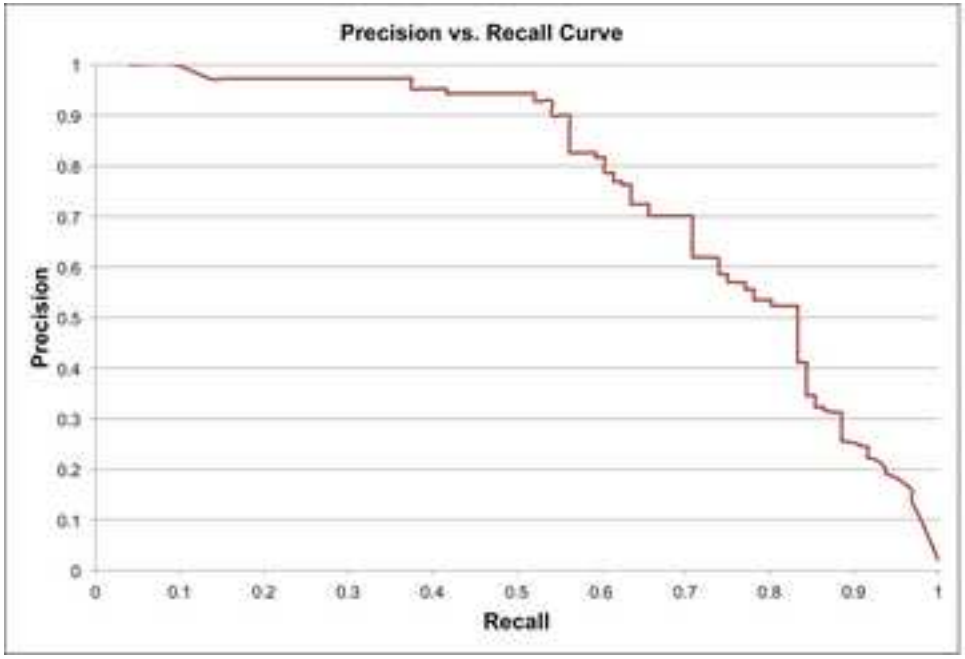
Figure(s)

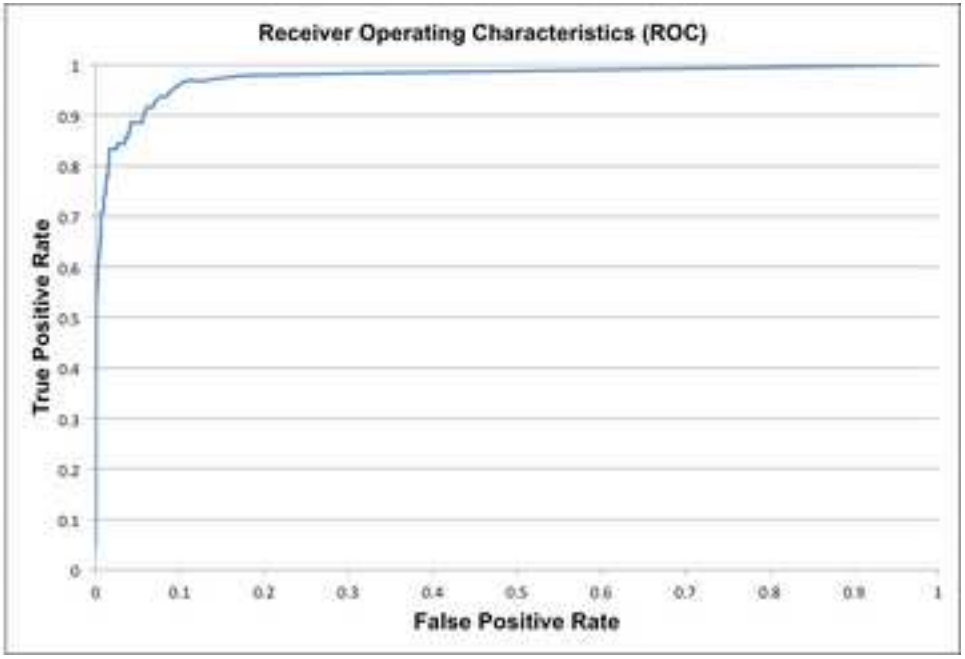
Click here to download high resolution image



Figure(s)

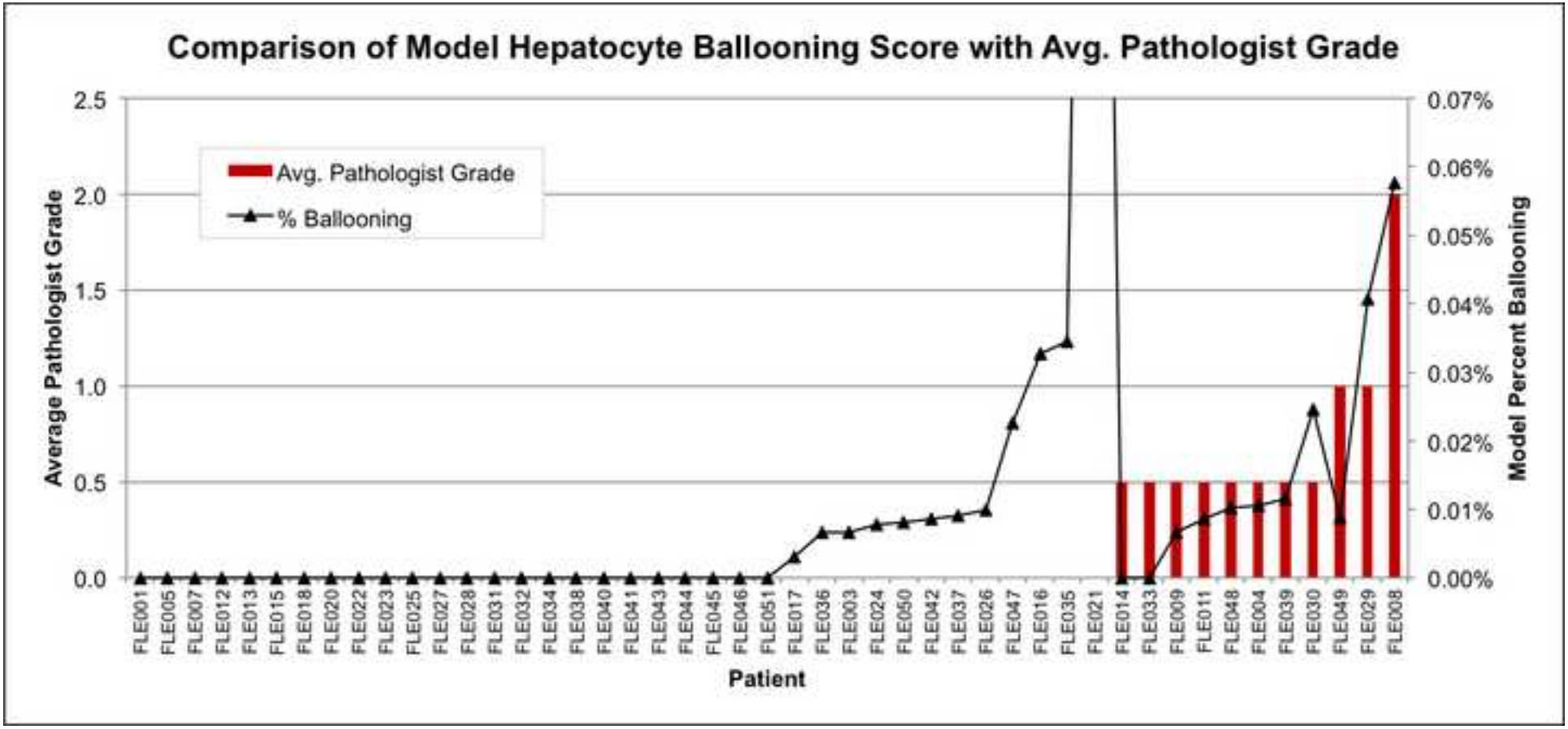
Click here to download high resolution image



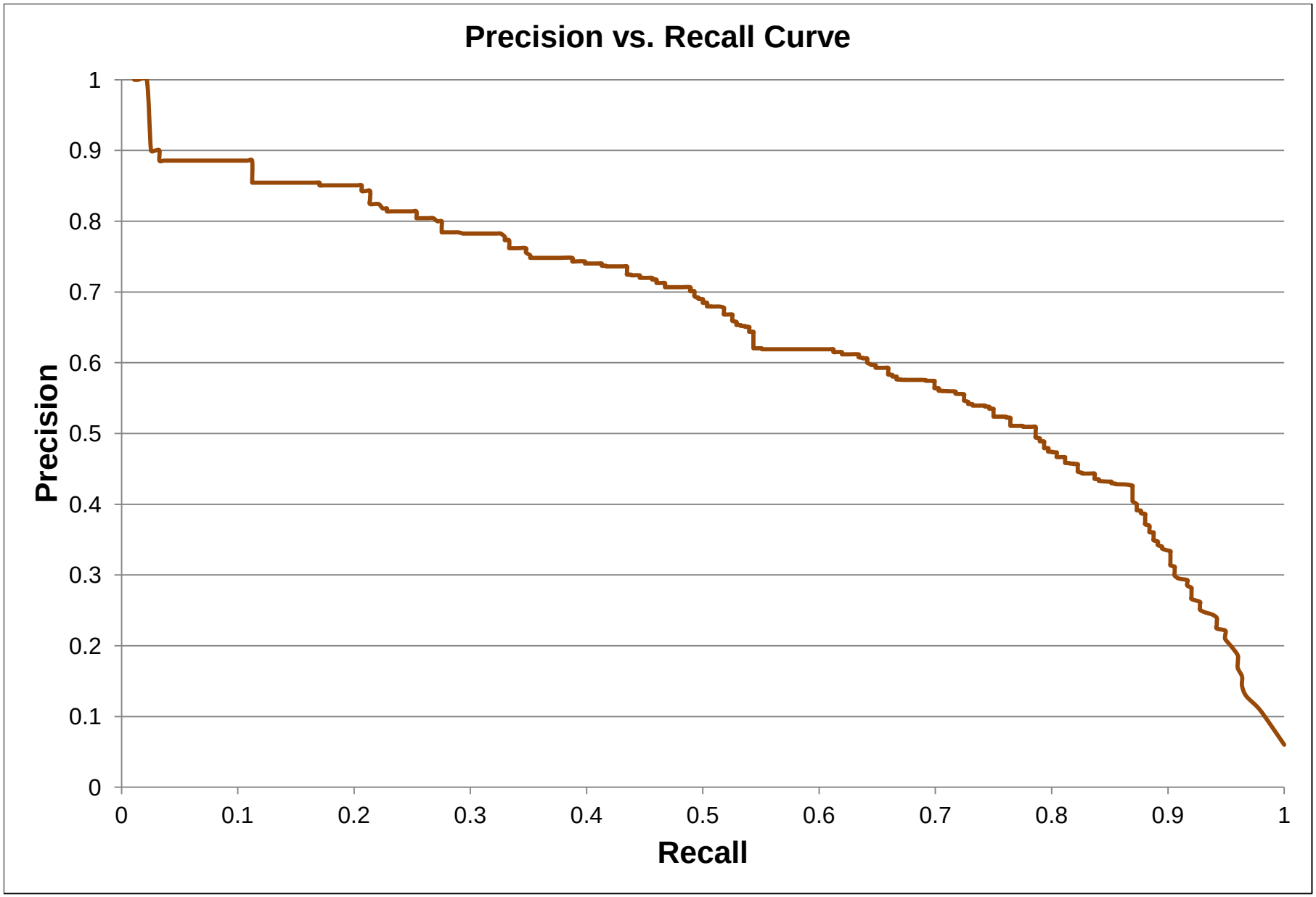


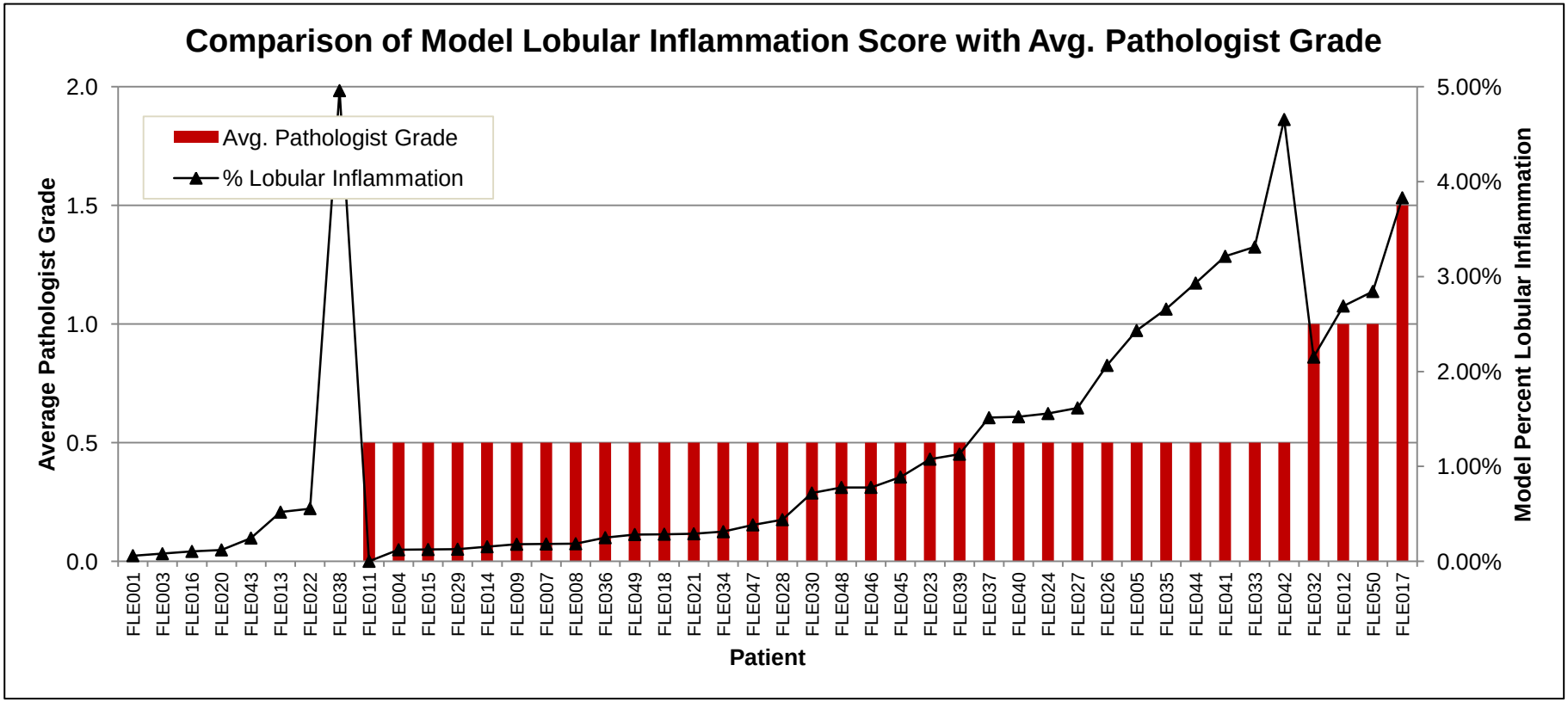
Figure(s)

Click here to download high resolution image

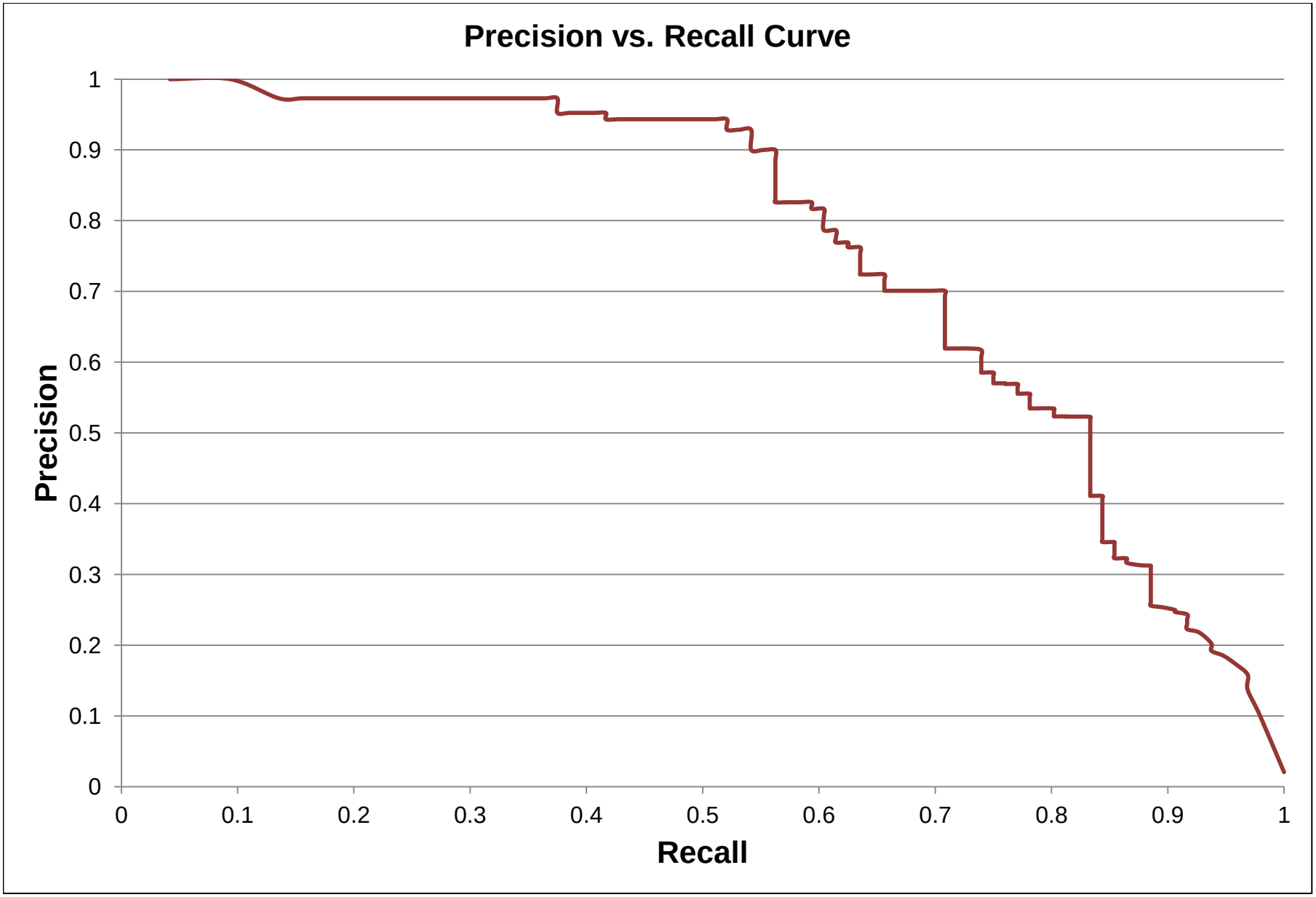


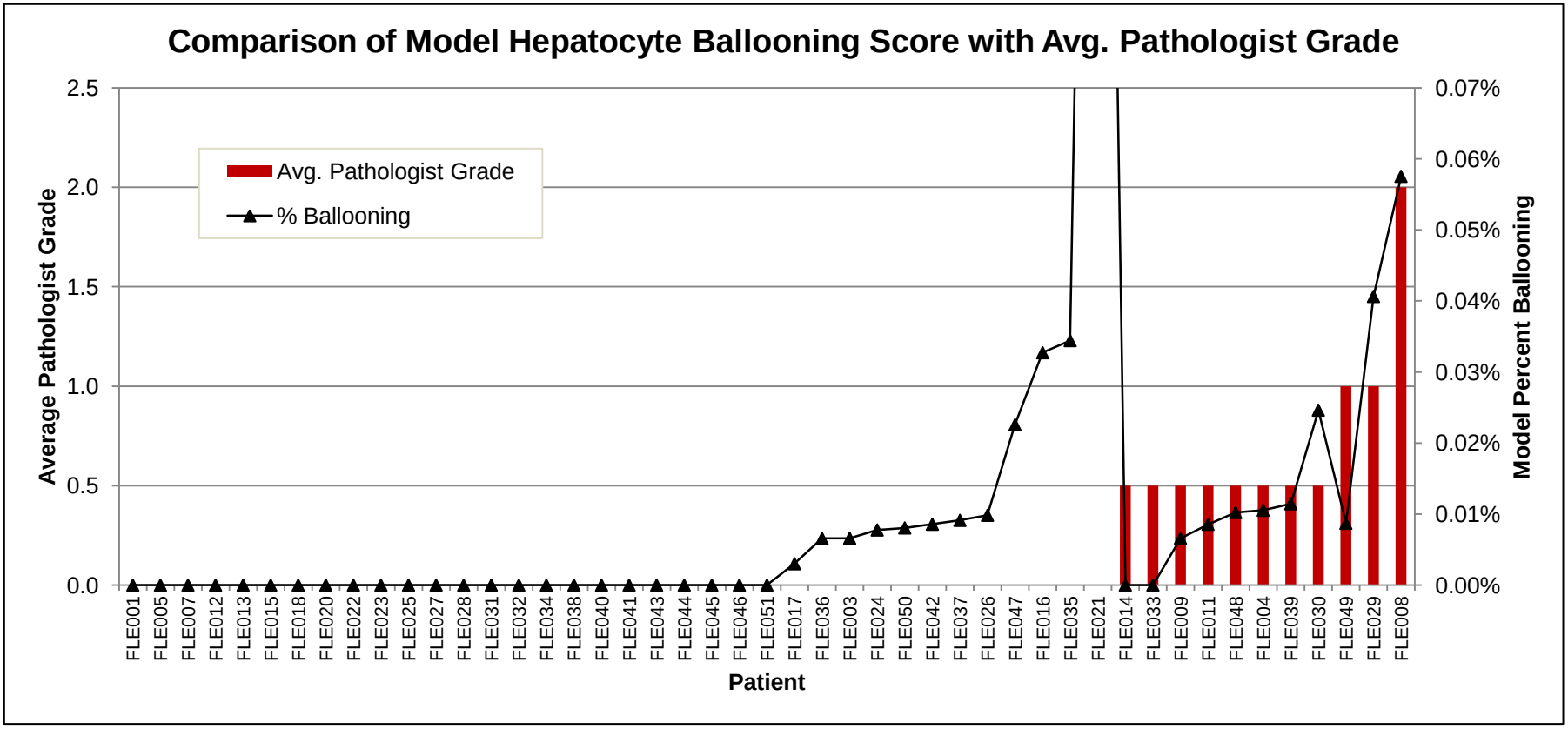
Figure(s)


Precision vs. Recall Curve
Precision
Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Comparison of Model Lobular Inflammation Score with Avg. Pathologist Grade
Avg. Pathologist Grade
% Lobular Inflammation
Average Pathologist Grade
Model Percent Lobular Inflammation
Patient
2.0
1.5
1.0
0.5
0.0
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
FLE001
FLE003
FLE016
FLE020
FLE043
FLE013
FLE022
FLE038
FLE011
FLE004
FLE015
FLE029
FLE014
FLE009
FLE007
FLE008
FLE036
FLE049
FLE018
FLE021
FLE034
FLE047
FLE028
FLE030
FLE048
FLE046
FLE045
FLE023
FLE039
FLE037
FLE040
FLE024
FLE027
FLE026
FLE005
FLE035
FLE044
FLE041
FLE033
FLE042
FLE032
FLE012
FLE050
FLE017

Figure(s)


Precision vs. Recall Curve
Precision
Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Comparison of Model Hepatocyte Ballooning Score with Avg. Pathologist Grade
Avg. Pathologist Grade
% Ballooning
Average Pathologist Grade
Model Percent Ballooning
Patient
2.5
2.0
1.5
1.0
0.5
0.0
0.07%
0.06%
0.05%
0.04%
0.03%
0.02%
0.01%
0.00%
FLE001
FLE005
FLE007
FLE012
FLE013
FLE015
FLE018
FLE020
FLE022
FLE023
FLE025
FLE027
FLE028
FLE031
FLE032
FLE034
FLE038
FLE040
FLE041
FLE043
FLE044
FLE045
FLE046
FLE051
FLE017
FLE036
FLE003
FLE024
FLE050
FLE042
FLE037
FLE026
FLE047
FLE016
FLE035
FLE021
FLE014
FLE033
FLE009
FLE011
FLE048
FLE004
FLE039
FLE030
FLE049
FLE029
FLE008