Abstract #2123

# Liver Biopsy Graph Neural Networks for Automated Histologic Scoring Using the NASH CRN System

## INTRODUCTION

The prevalence of nonalcoholic fatty liver disease (NAFLD) is rising rapidly, resulting in a concurrent increase in its progressive form, nonalcoholic steatohepatitis (NASH), which can lead to cirrhosis[1]. There are no currently approved therapeutics for NASH, with promising candidate drugs failing to meet surrogate clinical trial endpoints approved by the FDA that require pathologic review of liver biopsies[1]. Inter- and intra-pathologist variability in grading and staging disease based on histological features leads to inconsistency which may impact results[2,3]. "Black box" machine learning approaches using conventional neural networks (CNNs) can interpret NASH histology on digitized slides, but their application is limited by lack of interpretability. Graph Neural Networks (GNNs) are an emerging deep learning method that represent and characterize histologic features using graph representations and are well-suited to data types that can be modeled by a graph structure, such as fibrosis architecture[4].

The GNN described here has been incorporated in to a PathAI Drug Development tool (DDT) which provides AI-based histological measurements of nonalcoholic steatohepatitis (AIM-NASH). This AIM-NASH DDT has been accepted by the FDA into the Biomarker Qualification Program to be evaluated for use in trial enrollment and to determine histologic-based endpoints.

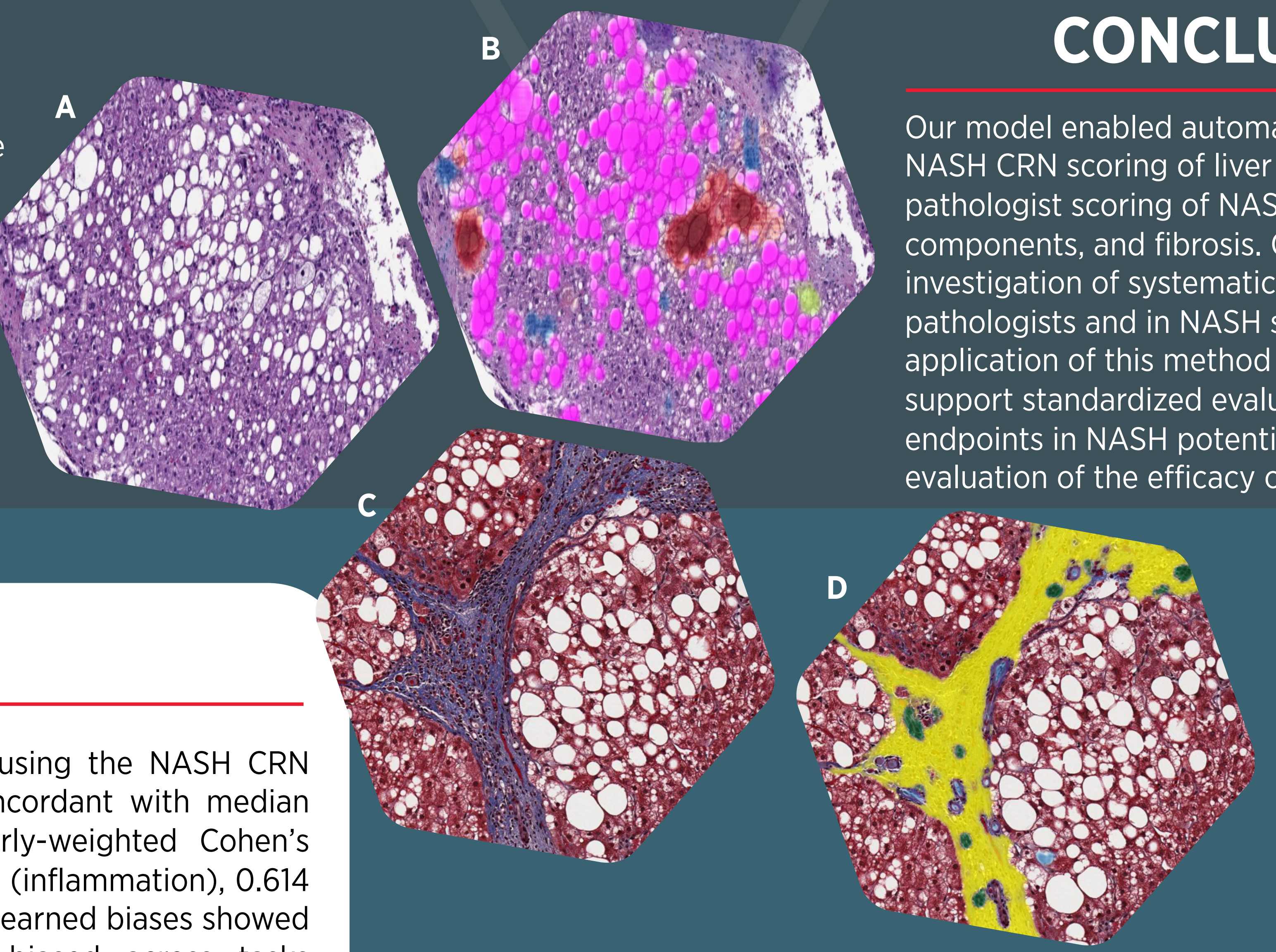


**Figure 2. GNN Model Input NASH Liver Tissue** (A) hematoxylin and eosin (H&E) stain; (B) ML model detection of H&E tissue classes including NAS components: steatosis (pink), lobular inflammation (blue), and ballooning (red); (C) NASH Liver Tissue with Masson trichrome stain; (D) ML model detection of trichrome tissue classes including fibrosis (yellow) and bile duct (green).

## CONCLUSIONS

Our model enabled automated and reproducible NASH CRN scoring of liver biopsies, recapitulating pathologist scoring of NAS, the three NAS components, and fibrosis. Our approach enabled investigation of systematic biases among pathologists and in NASH scoring tasks. Future application of this method to clinical trials could support standardized evaluation of histologic endpoints in NASH potentially improving evaluation of the efficacy of NASH therapeutics.

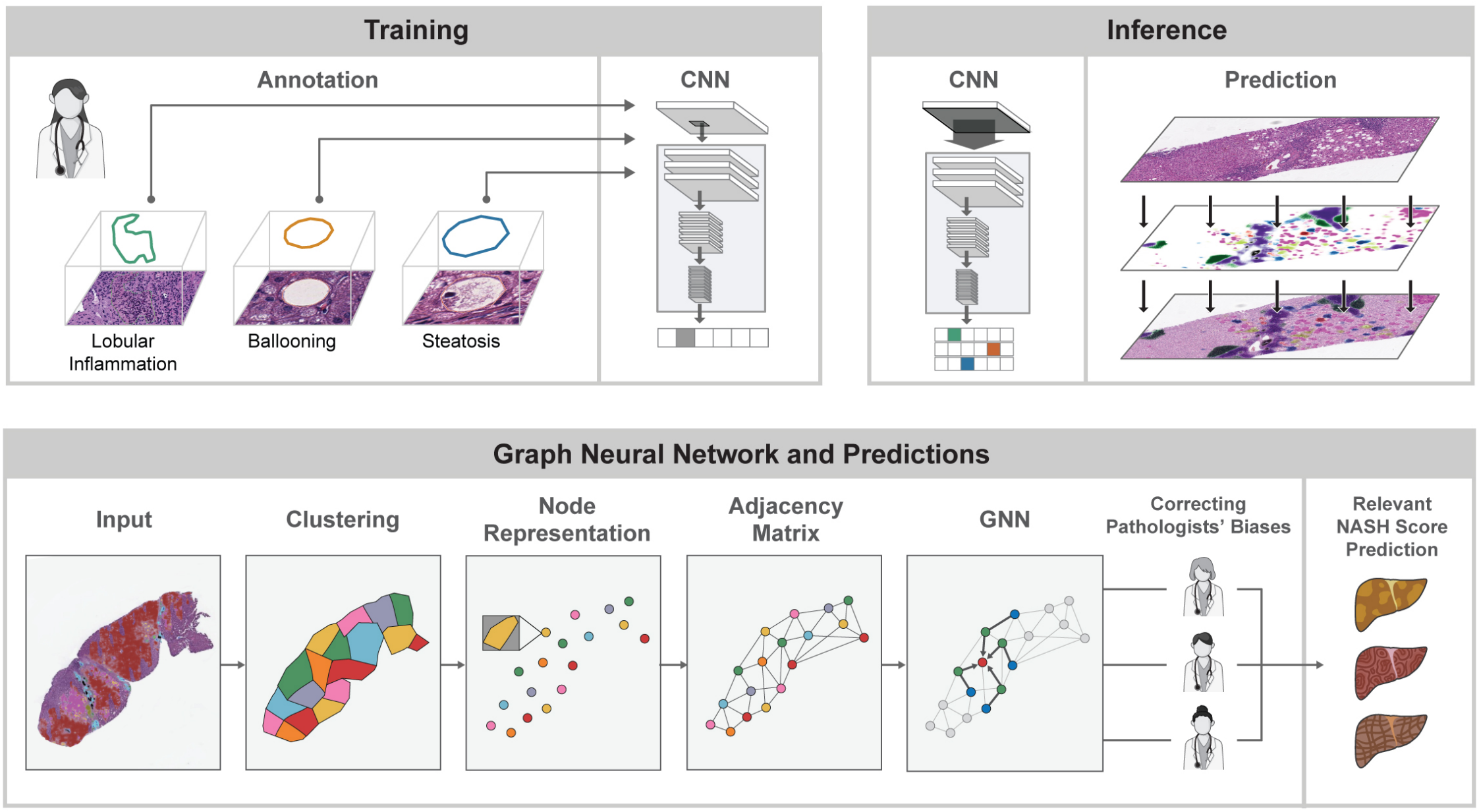
## METHODS

**Feature Overlay Generation on H&E and Trichrome WSIs**

Whole-slide images (WSI) of 639 H&E and 633 trichrome NASH liver biopsies from clinical trial participants (EMMINENCE, NCT02784444) were scored by 3 pathologists for NAFLD Activity (NAS 0-8) and its 3 components - inflammation (0-3), steatosis (0-3), and ballooning (0-2), and for fibrosis (0-4). Images were split into train, validation and test sets.

A pathologist network annotated WSIs for tissue regions. Using the annotations on training set images, CNNs were trained to generate pixel-level predictions of 13 H&E and 5 trichrome classes (*e.g.,* steatosis, bile duct, etc.) (**Figures 1-2**). WSIs were then converted into directed graphs using these CNN generated overlays as follows.

**Figure 1.** Workflow for CNN model training and GNN Development



**Pixel Clustering**

The CNN predictions for each WSI were clustered into "super-pixels" to construct the nodes in the graph. To increase the computational efficiency of clustering, pixels are randomly sampled from each WSI and clustered based on their spatial coordinates via the Birch clustering method. All CNN predictions were then assigned to the cluster of their nearest neighbor from the clustered subset of ~5000. This process reduced hundreds of thousands of pixel-level predictions into thousands of super-pixel clusters. WSI regions predicted as background or normal tissue were excluded during clustering.

## METHODS

**Graph Construction and Node Featurization**

Directed edges were placed between each node and their 5 nearest neighboring nodes (via the K Nearest Neighbor algorithm). Self-loops were also incorporated. Each graph node is represented by three classes of features generated from previously trained CNN predictions pre-defined to be biological classes of known clinical relevance: Spatial features included the mean and standard deviation of (x, y) coordinates. Topological features included area, perimeter, convexity of the cluster. Logit-related features included the mean and standard deviation of logits for each of the classes of CNN generated overlays (**Figure 1, bottom panel**).

**GNN Architecture**

Using an architecture equipped with graph convolution and graph pooling modules the GNN performed a graph-level ordinal classification of NAS components (from H&E WSI ) and CRN scores (from trichrome WSI). Graph convolution aggregates features from nodes' local neighbors and generalizes the operation of convolution from grid data (in standard CNN) to graph data.[5] Hidden layers transform each input graph into another graph with updated node features. The final graph pooling layer enables GNNs to update the graph structure in addition to the node features [6,7].

**Accounting for Pathologist Bias**

Historically, pathologist concordance for scoring NAS is moderate to weak (**Figure 3**). Individual pathologists may interpret guidelines inconsistently, leading to systematic biases. Our GNN approach models and accounts for these biases to produce unbiased estimates for each WSI.

**Figure 3.** Assessment of pathologists' agreement (Cohen's kappa) Across NAS Components and CRN Score

| Annotation Task | Pathologists 1, 2 | Pathologists 2, 3 | Pathologists 1, 3 | Average Kappa |
|---|---|---|---|---|
| Steatosis score (0 -3) | 0.632 | 0.621 | 0.535 | 0.596 |
| Lobular inflammation score (0-3) | 0.351 | 0.307 | 0.319 | 0.325 |
| Ballooning score (0-2) | 0.474 | 0.516 | 0.424 | 0.481 |
| CRN score (0-4) | 0.484 | 0.409 | 0.609 | 0.501 |

## RESULTS

Here, GNNs were applied to NASH scoring using the NASH CRN system[1] producing predictions that were concordant with median expert pathologists scores (test set): linearly-weighted Cohen's kappa of 0.613 (NAS), 0.758 (steatosis), 0.584 (inflammation), 0.614 (ballooning), and 0.507 (fibrosis) (**Figure 4**). Learned biases showed that pathologist scoring was consistently biased across tasks (Pathologist 1 scored lower than consensus, mean bias -0.459; Pathologist 3 scored higher, 0.495) (**Figure 5**). In addition, we observed task-specific biases among pathologists, with inflammation showing minimal bias while fibrosis being highly biased (**Figure 5**).

**Figure 4.** GNN Models Recapitulate NASH Scoring Classification

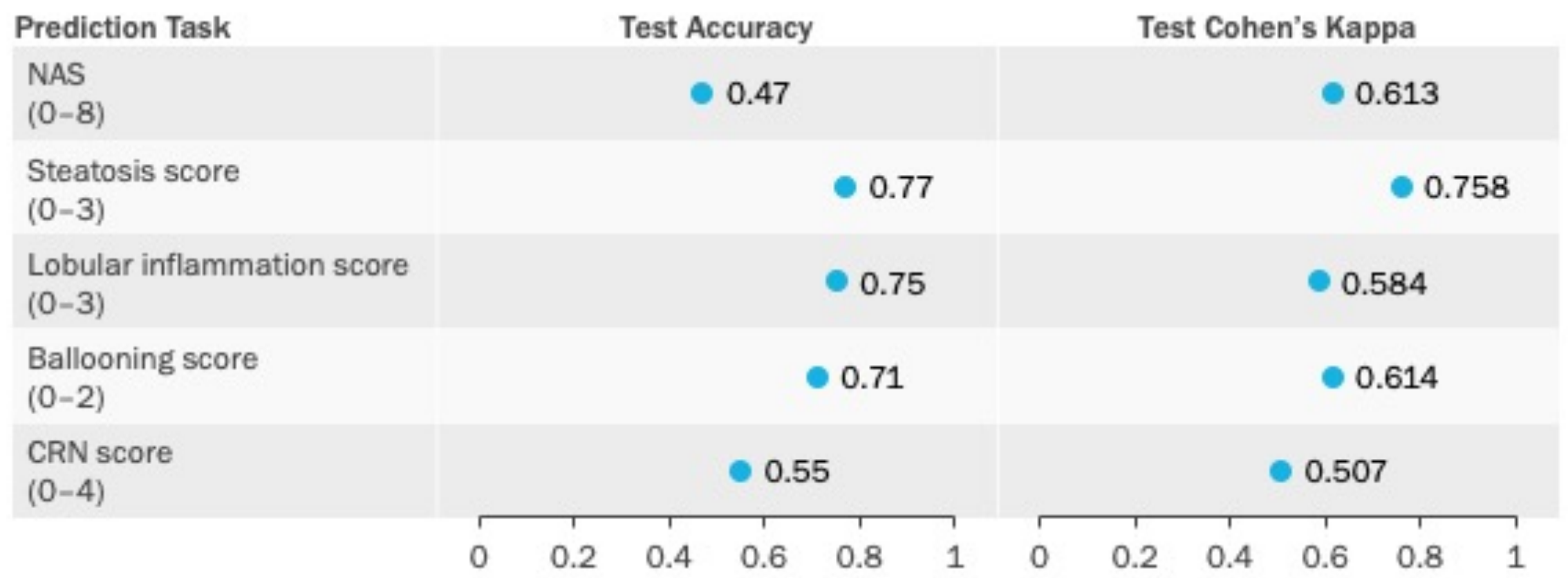
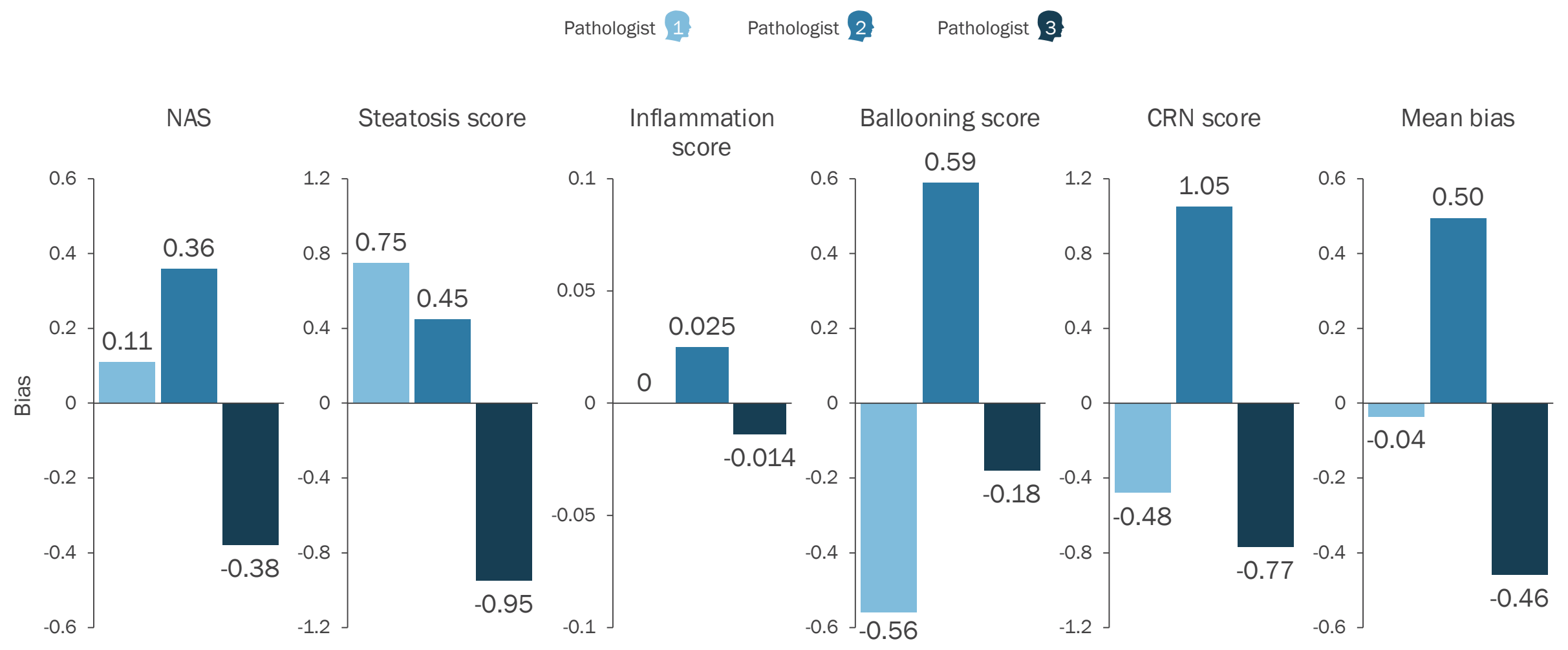


**Figure 5.** Pathologist Biases Across NAS Components and CRN Score



To learn more about model performance and our AIM-NASH DDT attend our oral presentation at ILC 2021, *AI-based histologic measurement of NASH (AIM-NASH): A drug development tool for assessing clinical trial endpoints* (*OS-1611*).

## AUTHORS

Jason K Wang[1], Maryam Pouryahya[1], Kenneth Leidal[1], Harsha Pokkalla[1], Dinkar Juyal[1], Zahil Shanis[1], Aryan Pedawi[1], Quang Huy Le[1], Victoria Mountain[1], Sara Hoffman[1], Jackie Honerlaw[1], Murray Resnick[1], Michael Montalto[1], Andy Beck[1], Katy Wack[1], Ilan Wapinski[1], Oscar M. Carrasco-Zevallos[1], Amaro Taylor-Weiner[1]

[1]PathAI, Boston, MA

## CONTACT INFORMATION

Amaro Taylor-Weiner, amaro.taylor@pathai.com

## REFERENCES

1. US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. Noncirrhotic nonalcoholic steatohepatitis with liver fibrosis: developing drugs for treatment draft guidance for industry. https://www.fda.gov/media/119044/download. Published December 2018. Accessed 06/04/2021
2. Kleiner D.E., et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology. 2005; 41(6):1313-1321.
3. Kleiner D.E., et al. Association of Histologic Disease Activity With Progression of Nonalcoholic Fatty Liver Disease. JAMA Netw. Open. 2019 Oct 2;2(10):e1912565.
4. Hamilton et al., Representation Learning on Graphs: Methods and Applications arXiv:1709.05584v3
5. Morris, C., et al. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. Proceedings of the AAAI Conference on Artificial Intelligence,. 2019. 33(01), 4602-4609.
6. Diehl, F. Edge Contraction Pooling for Graph Neural Networks. 2019. arXiv:1905.10990
7. Lee, J., et al. Self-Attention Graph Pooling.2019. arXiv:1904.08082v4

**ACKNOWLEDGMENTS** This poster template was developed by SciStories LLC. https://scistories.com/