Poster LP44

Comparison of manual vs machine learning approaches to liver biopsy scoring for NASH and fibrosis: a post hoc analysis of the FALCON 1 study

Diane E. Shevell,[1] Elizabeth Brown,[1] Anne Minnich,[1] Shuyan Du,[1] Janani Iyer,[2] Katy Wack,[2] Vipul Baxi,[1] Dimple Pandya,[1] Peter Schafer,[1] Zachary D. Goodman,[3] Edgar D. Charles[1]

[1]Bristol Myers Squibb, Princeton, NJ; [2]PathAI, Boston, MA; [3]Inova Fairfax Hospital, Falls Church, VA

Introduction

- While manual histological evaluation of liver biopsy tissue is the gold-standard method for fibrosis and disease staging in nonalcoholic steatohepatitis (NASH),[1] it is limited by inter- and intra-reader variability
- Machine learning models that have been trained to analyze and interpret liver histopathology may help improve reproducibility of NASH grading and staging[2]
- In liver biopsy tissue, fibrosis staging and nonalcoholic fatty liver disease activity score (NAS) results determined by PathAI, a machine learning-based approach, have been shown to correlate with those obtained from manual interpretation[2]
- This exploratory post hoc analysis compared manual (single central reader) and PathAI pathology scoring of liver biopsy samples from patients with NASH and stage 3 fibrosis in the phase 2b FALCON 1 study

Methods

Study design and participants

- FALCON 1 (NCT03486899) was a phase 2b, randomized, multicenter, placebo-controlled study assessing the efficacy and safety of pegbelfermin (PGBF)[3]
- Eligible adults were 18–75 years of age with a liver biopsy tissue specimen collected within 6 months prior to or during screening that was consistent with NASH with a score of ≥ 1 for each NAS component and stage 3 liver fibrosis according to the NASH CRN classification[4]
- During the 48-week, double-blind, treatment period, patients received 10, 20, or 40 mg PGBF or placebo subcutaneously once weekly
- The primary histological endpoint was ≥ 1 stage improvement in fibrosis without NASH worsening or NASH improvement with no worsening of fibrosis at week 24, as determined by a single central reader
- See oral presentation LO5 for additional FALCON 1 study details

Assessments

- Liver biopsies were performed within 6 months of screening and at week 24; patients who completed week 24 and had paired, evaluable, biopsy specimens at both timepoints were included in the analysis
- Biopsy tissue was manually scored according to NASH CRN fibrosis criteria and NAS components by a central pathologist (Z.D.G.) who was blinded to treatment assignment and specimen sequence
- The PathAI machine learning algorithm used in this study was trained using scored liver biopsy specimens from clinical trial patients with NASH, primary sclerosing cholangitis, or hepatitis B
  - For NASH specimens, fibrosis scoring according to NASH CRN fibrosis criteria and NASH disease activity using NAS were performed by 5 pathologists, and feature annotations were provided by 59 pathologists; all pathologists were board certified and had demonstrated prior experience scoring NASH cases
  - The non-NASH specimens were used to collect feature annotations from pathologists to train the algorithm to more specifically identify NASH-specific histological features
- For this study, the same baseline and week 24 liver biopsy tissue slides from patients enrolled in FALCON 1 were also scored using the machine learning algorithm to blindly evaluate the primary endpoint (ordinal scoring) and NASH CRN fibrosis criteria and NAS components (ordinal and continuous scoring)

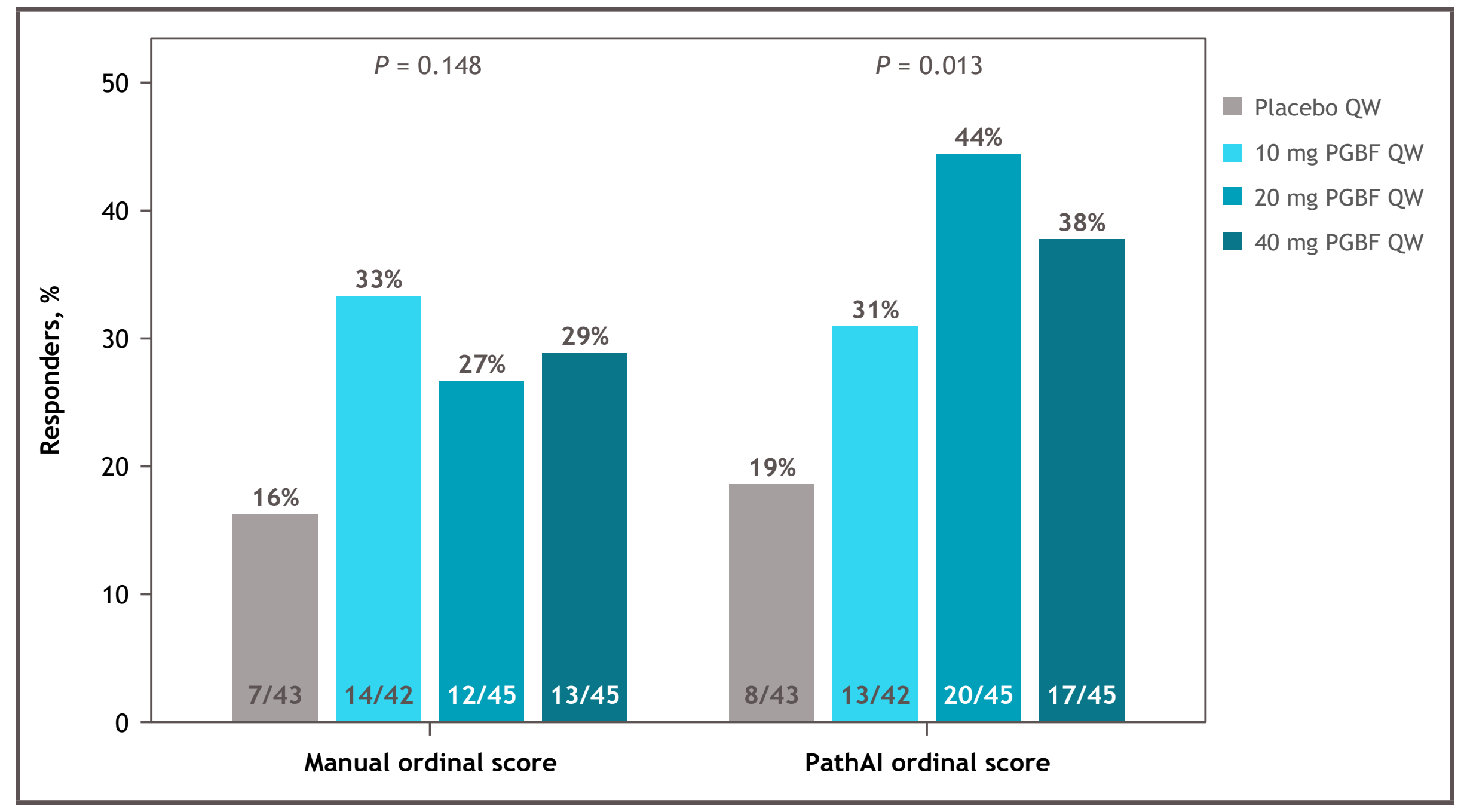
Statistical analyses

- The Cochran-Armitage test for trend was used to assess differences in the proportion of responders or patients with improvements in PGBF vs placebo arms
- Primary endpoint responders were patients with ≥ 1 stage NASH CRN fibrosis improvement without NASH worsening (≥ 1 point increase in NAS) or NASH improvement (≥ 2 point decrease in NAS with contribution from > 1 component) with no worsening of fibrosis at week 24 according to histopathological analysis
- Pairwise canonical correlations were calculated for manual and PathAI scores, and biopsy-based and imaging metrics; correlations are reported as absolute values for those that passed the Benjamini-Hochberg adjusted *P* value of 0.1 after correction for multiple testing
- Linear mixed-effect models were fit for each continuous PathAI score; measurements were regressed on time and treatment arm, including an interaction between time and treatment, and a random effect for each patient

Results

- In FALCON 1, a total of 197 patients were randomized to the 4 study arms; patients with evaluable biopsy samples were included in this analysis (43 patients in the placebo arm, 42 patients in the 10 mg PGBF arm, and 45 patients each in the 20 mg and 40 mg PGBF arms)
- Baseline demographics and patient characteristics were similar across study arms
  - The majority of patients were female (59%) and White (85%), and had type 2 diabetes (74%); the mean age and mean body mass index were 57 years and 36 kg/m^2, respectively
  - See oral presentation LO5 for additional FALCON 1 baseline data
- Precise agreement between manual and PathAI ordinal scores was relatively low for all NAS components; kappa estimates (95% CIs) were 0.49 (0.39–0.58) for ballooning, −0.06 (−0.11 to −0.01) for lobular inflammation, 0.11 (0.03–0.19) for steatosis, and 0.42 (0.30–0.53) for NASH CRN fibrosis score
- Both ordinal scoring methods indicated that the percentage of primary endpoint responders was nearly double in the PGBF arms compared with the placebo arm (**Figure 1**)
  - A significantly greater number of primary endpoint responders was detected in the PGBF vs placebo arms by PathAI ordinal scoring (*P* = 0.013) but not by manual ordinal scoring (*P* = 0.148)

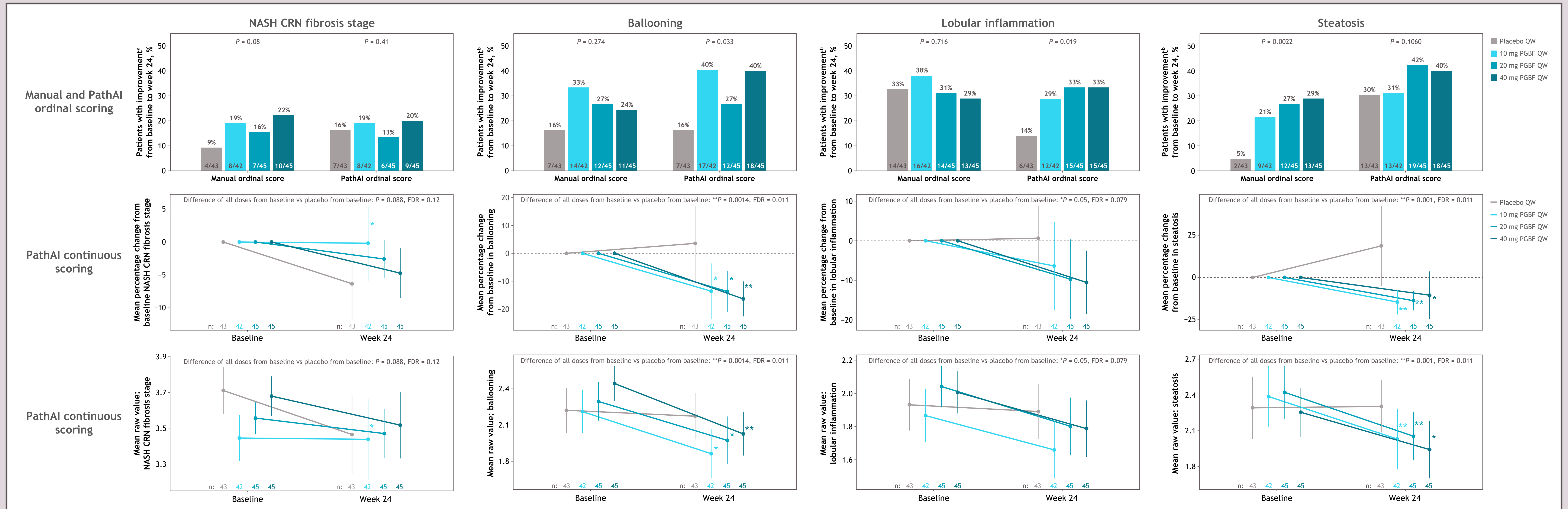
Figure 1. Manual and PathAI ordinal scoring of primary endpoint responders[a]



[a]Primary endpoint responders were patients with ≥ 1 stage NASH CRN fibrosis improvement without NASH worsening or NASH improvement with no worsening of fibrosis at week 24. Cochran-Armitage test for trend was used to compare PGBF vs placebo. NASH, nonalcoholic steatohepatitis; PGBF, pegbelfermin; QW, once weekly.

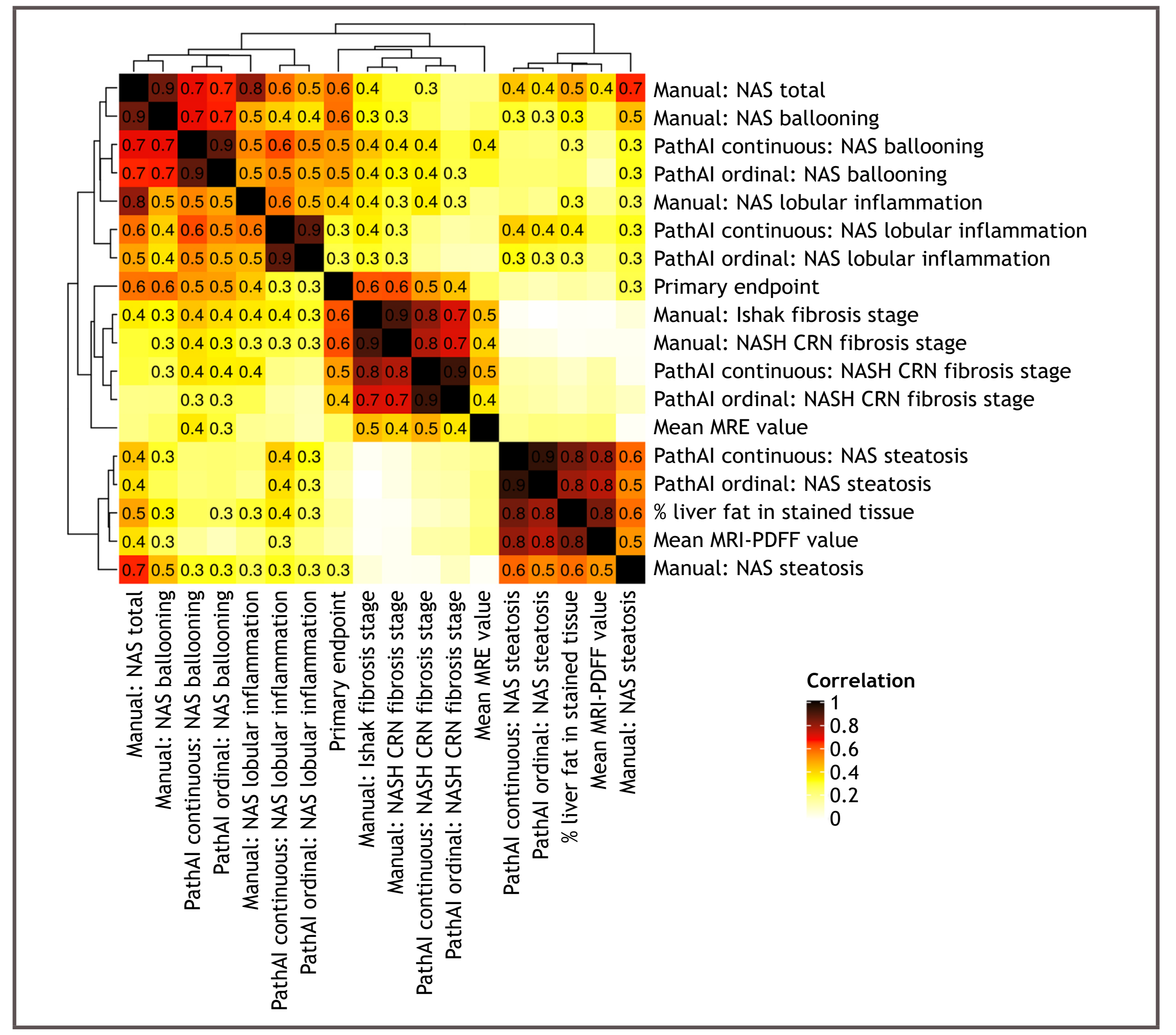
- Fibrosis stage was not significantly improved with PGBF vs placebo with any scoring method (manual ordinal: *P* = 0.08; PathAI ordinal: *P* = 0.41; PathAI continuous: *P* = 0.088; **Figure 2**)
- PathAI ordinal scoring, but not manual scoring, detected a significant difference in the number of patients in PGBF vs placebo arms who had improvements in ballooning (PathAI ordinal: *P* = 0.033; manual ordinal: *P* = 0.274) and lobular inflammation (PathAI ordinal: *P* = 0.019; manual ordinal: *P* = 0.716)
  - The opposite was true for steatosis; manual ordinal scoring (*P* = 0.0022) but not PathAI ordinal scoring (*P* = 0.1060) identified a significant difference in the number of patients with improvement in the PGBF arms compared with the placebo arm
- PathAI continuous scoring demonstrated statistically significant improvement from baseline for PGBF compared with placebo for all 3 NAS components (ballooning: *P* = 0.0014; lobular inflammation: *P* = 0.05; steatosis: *P* = 0.001)
- Correlations between manual and PathAI scores, and other biopsy-based and imaging metrics were further investigated; as shown in **Figure 3**, the following clusters were observed at week 24:
  - Ballooning and lobular inflammation measured by both manual and PathAI (ordinal and continuous) scoring using NAS
  - Fibrosis measured by Ishak stage and NASH CRN fibrosis stage (manual and PathAI continuous scoring), and magnetic resonance elastography
  - Steatosis measured by manual and PathAI (ordinal and continuous) scoring, % fat on biopsy, and magnetic resonance imaging-proton density fat fraction

Figure 2. Manual and PathAI scoring of NASH CRN fibrosis stage and NAS components



Improvement was defined as [a]≥ 1 stage improvement in NASH CRN fibrosis stage or [b]≥ 1 point improvement in NAS components (ballooning, lobular inflammation, and steatosis). PathAI continuous scoring data reflect mean (95% CI). Cochran-Armitage test for trend was used to compare PGBF vs placebo: *$P \leq 0.05$; **$P \leq 0.01$; ***$P \leq 0.001$; ****$P \leq 0.0001$. FDR, false discovery rate; NAS, nonalcoholic fatty liver disease activity score; NASH, nonalcoholic steatohepatitis; PGBF, pegbelfermin; QW, once weekly.

Figure 3. Pairwise correlations in week 24 biopsy tissue[a]



[a]Correlations are reported as absolute values for those that passed the adjusted *P* value of ≤ 0.1 after correction for multiple testing. MRE, magnetic resonance elastography; MRI-PDFF, magnetic resonance imaging-proton density fat fraction; NAS, nonalcoholic fatty liver disease activity score; NASH, nonalcoholic steatohepatitis.

Conclusions

- Agreement between machine learning and the single central pathologist was relatively low; however, both machine learning and manual scoring showed improvements in histological responses in PGBF arms compared with the placebo arm
- Significant moderate and strong correlations were observed between ballooning and inflammation, fibrosis, and steatosis measures for both manual and PathAI scoring
- PathAI continuous scoring demonstrated statistically significant improvement from baseline for PGBF compared with placebo for all 3 NAS components
- Determination of the clinical significance of these findings will require larger trials, more detailed evaluation of specific histological changes, and correlation with clinical outcomes

References

1. Chalasani N, et al. *Hepatology*. 2018;67(1):328–357.
2. Pokkalla H, et al. AASLD 2019. Abstract 187.
3. Abdelmalek MF, et al. *Contemp Clin Trials*. 2021;104:106335.
4. Kleiner DE, et al. *Hepatology*. 2005;41(6):1313–1321.

Acknowledgments & Disclosures

- DE Shevell, E Brown, S Du, V Baxi, D Pandya, P Schafer, and ED Charles were employees of Bristol Myers Squibb at the time this study was conducted and may own company stock. A Minnich is a consultant for Bristol Myers Squibb. J Iyer and K Wack are employees of PathAI. ZD Goodman has received grants from Alexion Pharmaceuticals, Allergan, Conatus Pharmaceuticals, Exalenz Bioscience, Galactin Therapeutics, Gilead Sciences, and Intercept Pharmaceuticals.
- This study was supported by Bristol Myers Squibb
- All authors contributed to and approved the presentation; writing assistance was provided by Kendall Foote, PhD, of Medical Expressions, funded by Bristol Myers Squibb

Presented at AASLD The Liver Meeting®, November 12-15, 2021, Virtual

Email: Edgar.Charles@bms.com