TGH TRANSLATIONAL GASTROENTEROLOGY AND HEPATOLOGY
A RISING JOURNAL FOR HIGH-QUALITY RESEARCH IN GASTROENTEROLOGY AND HEPATOLOGY

Peer Review File

Article information: https://dx.doi.org/10.21037/tgh-22-27

Reviewer A

This retrospective study of 193 patients diagnosed with NAFLD who had MRE-assessed liver stiffness, FIB-4, and NFS measurements within six months. The goal is to evaluate serum biomarkers of FIB-4 and NFS in stratifying fibrosis stages concerning the reference standard of MRE-assessed liver stiffness. The authors concluded that FIB-4 and NFS are good at screening advanced fibrosis at >80% NPV, but suboptimal in identifying advanced fibrosis with poor PPV of around 70%.

COMMENT 1: The statistical analyses need expertise review. In my opinion, the dichotomized outcome and associated correlation are OK to report but not the best to evaluate the accuracy of FIB-4 and NFS which can provide 3 ordinal classes against the reference standard that gives 6 stages of fibrosis. In addition, there is no consensus on the cut-off thresholds of these noninvasive biomarkers for NAFLD yet. Therefore, concordance evaluations might be more appropriate to validate the agreement between these tests and the potential of FIB-4 and NFS in screening mild to moderate fibrosis.

REPLY 1: Our statistics were performed by a Senior Data Analyst and Biostatistician at our institution. FIB-4 and NFS divide liver disease into 3 fibrosis categories while MRE is more detailed and differentiates 6 different categories. To compare FIB-4 and NFS to MRE, we grouped the 6 MRE categories into 3 sub-categories ('Stage ≥1 to ≤2' fibrosis or lower, 'Stage >2 to <3', 'Stage ≥3 to <4' fibrosis or higher). When it comes to FIB-4 and NFS classes, we used cutoff values that are widely used and presumed to be standards. There is debate as to whether these cutoff values should be adjusted based on underlying medical conditions (for example https://www.journal-of-hepatology.eu/article/S0168-8278(20)30445-1/fulltext), however we excluded patients with underlying medical conditions that might impact our results (see Methods lines 92-94). The NFS scoring system was made specifically for NAFLD and we used cutoff values that are well established (https://www.mdcalc.com/nafld-non-alcoholic-fatty-liver-disease-fibrosis-score#evidence). FIB-4 was initially created for viral hepatitis patients but has become widely used for NAFLD. With regards to proper cutoff values, some use 3.25 for the detection of advanced fibrosis however this is mainly used for evaluating viral hepatitis. In our study, 2.67 was used as the cutoff as this is a known and accepted standard cutoff for patients

with NAFLD (https://www.mdcalc.com/fibrosis-4-fib-4-index-liver-fibrosis#evidence - click 'Evidence' to see details with regards to FIB-4 use in NAFLD). The 'Serum biomarkers and associated algorithms' section of the following paper also discusses this (https://journals.lww.com/ajg/Fulltext/2021/02000/Role_of_Noninvasive_Tests_in_Clinical.13.aspx) and we added this information and citation to our paper. The cutoff values used are well known and have proven to be quite accurate (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3079239/). Based on our study design, the greatest statistical value that we could obtain is displayed by our NPV values of .84 for FIB-4 and .89 for NFS as this directly represents the high degree of correspondence between MRE and the biomarker tests in early stage disease.

CHANGES IN THE TEXT 1: We added a sentence about the accuracy of FIB-4 using the citation mentioned above (lines 67-68)

COMMENT 2: High-risk NAFLD with clinically significant fibrosis (F2 and greater) is crucial in therapeutical trials and patient management. Advanced fibrosis (F3 and greater) requires HCC surveillance imaging anyway. The study goal can be stronger to rule out or identify clinically significant fibrosis.

REPLY 2: Our results suggests good correspondence in early stage disease while in late stage disease the data is not as strong. In our Methods section, lines 122-131 discuss how the 3 different categories in FIB-4 and NFS were compared to the MRE classification sub-categories. We used the Further Investigation Needed and Indeterminate groups as a grey zones given that the terminology used ("Further Investigation Needed" and "Indeterminate") suggests that the data in these categories is uneliable. Our data for Advanced fibrosis (F3 and greater) has already shown to not be very convincing. For identifying advanced fibrosis/cirrhosis, FIB-4 was dichotomized as Advanced Fibrosis Likely vs. other (Further Investigation Needed and Advanced Fibrosis Excluded groups) and NFS was dichotomized as F3-4 vs. other (Indeterminate and F0-F2 groups). These were compared to MRE, which was dichotomized as 'Stage ≥3 to <4' fibrosis or higher vs. 'Stage >2 to <3' fibrosis or lower. If we change our study to instead revolve around identifying clinically significant fibrosis (F2 and greater) then our new groups would have FIB-4 dichotomized as Advanced Fibrosis Likely and Further Investigation Needed vs. other (only Advanced Fibrosis Excluded) while NFS would dichotomize as F3-4 and Indeterminate vs. other (only F0-F2). These would be compared to MRE 'Stage >2 to

<3' fibrosis or higher vs. 'Stage ≥1 to ≤2' fibrosis or lower. Including the data from the Further Investigation Needed and Indeterminate groups in our target group might result in our data being less clinically significant as FIB-4 and NFS scores in these ranges have proven to be inaccurate and unpredictable. In addition, our results analyzing FIB-4 and NFS as compared to MRE in advanced fibrosis have already shown questionable PPV-based correlation. If we saw a significant jump in PPV from 0.63 for FIB-4 and 0.72 for NFS (Results lines 153-155) after expanding our data groups to include the Further Investigation Needed group for FIB-4, Indeterminate group for NFS and Stage >2 to <3 for MRE, this wouldn't indicate a strong correlation for all data within the clinically significant fibrosis (F2 and greater) group as we have already proven that the data is not strong for a large portion of that group, the portion falling within the advanced fibrosis (F3 and greater) range. Higher PPV values would just signify that data for MRE corresponds closely with that of the two biomarker groups in this grey zone in particular, given that our data for advanced fibrosis is already calculated and known.

CHANGES IN THE TEXT 2: None

COMMENT 3: Please provide confusion matrix for Table 2. Did you perform iterative cross-validation to avoid bias?

REPLY 3: Please see Table 4 and Table 5 which were added. We did not perform iterative cross-validation in this study. To avoid bias we performed a retrospective chart review where raw data was obtained from patient charts. Then, statistical results based on this patient data were calculated and analyzed in a blinded fashion.

CHANGES IN THE TEXT 3: Table 4 and Table 5 were created and a sentence was added to the manuscript (lines 149-150) to explain this.

Reviewer B

COMMENT 1: This study set out to examine to association between MRE and established NITs for staging liver fibrosis (FIB-4 and NFS) to assess suitability in the clinical pathway for patients with NAFLD. Using fibrosis gradings from MRE as the gold standard, the analysis reported the NPV and PPV for FIB-4 and the NFS for pre

selected thresholds for both biomarkers. Results reported acceptable NPV for ruling out significant fibrosis (≥0.8) but only fair PPV for ruling in advanced fibrosis (<0.8). Whilst I would agree that the use of expensive imaging biomarkers may not have a universal place for screening all patients, especially from a health economics perspective, I think the conclusion is overstating the use of the FIB-4 and NFS for staging fibrosis in the early stages (The LITMUS consortium recommend that minimum acceptable performance level of a biomarker is 80% for both sensitivity and specificity for a given cut-off). That said I would agree this is useful work to enter the scientific literature but have a few recommendations to be addressed in the text.

REPLY 1: Given the concern about our conclusion being an overstatement, we edited our Discussion section so that we would be sure to clarify that biopsy remains the gold standard. We want to make it known that MRE and biomarker tests are alternatives to liver biopsy and that the purpose of our study is to see how well these alternative options correspond, not to suggest that biopsy should be replaced by these options entirely.

CHANGES IN THE TEXT 1: We added two sentences to the discussion (lines 177-181).

COMMENT 2: In general I think this work strengths is that it is a retrospective analysis of real world data from patients evaluated within the hospital system and thus I think more of an emphasis on the current guidelines at this hospital (including a figure of the pathway) would help put in context when and when the biomarkers could complement each other and provide useful data for inputting into health economic models for health technology appraisal. This should include the opportunity to capture and assess patients who fall within indeterminate or 'further investigation needed' categories.

REPLY 2: It is worth noting that there is no uniform fibrosis screening and monitoring approach used by physicians within our healthcare system. These decisions are made independently. However, based on our study we created Figure 2, a new diagram which shows our proposed screening algorithm.

CHANGES IN THE TEXT 2: We added a sentence to the Discussion based on the creation of Figure 2 (lines 241-242)

COMMENT 3: I am also not sure of the novelty, this is an area being widely discussed in the academic literature, with posters presented at EASL by the LITMUS consortium (https://litmus-project.eu/wp-content/uploads/2020/11/FRI_23_ILC2020.pdf) and a recent meta-analysis by Xiao et al (2017 - https://pubmed.ncbi.nlm.nih.gov/28586172/). Also MRE is being considered as a composite metric with FIB-4 to assess the earlier stages (https://pubmed.ncbi.nlm.nih.gov/33214165/) to overcome some of the shortfalls of both biomarkers in the earlier stages of disease - an analysis using MRE in combination with FIB-4 to examine and validate this pathway would be helpful contribution to the literature.

REPLY 3: Studies analyzing the accuracy of MRE and biomarker tests tend to use biopsy as the reference standard. Our study is novel in that its purpose is to see how well the results obtained using MRE compare to FIB-4 and NFS. These screening tools are used when people are either unable or unwilling to undergo biopsy. Our study goal is determine whether a costly and timely test such as MRE is truly a necessary next step in these cases or if biomarker testing may pose as an effective alternative, either in certain cases or as a whole. The first link above is a meta analysis based upon the NASH CRN histological scoring system. It utilizes MRE performed within 6 months of biopsy with the biopsy results serving as the reference standard. Our study differs in that we don't compare MRE and biomarker results to biopsy, but instead to each other. The second article is similar to the first and states "pathological examination was used as the reference for assessing fibrosis." In addition, this article only analyzes the detection of significant fibrosis (SF), advanced fibrosis (AF), and cirrhosis. It doesn't assess the same fibrosis groups as ours, including our 'rule out advanced fibrosis' group from which the main statistical value of our study is derived. With regards to the final point, combining MRE with biomarker testing certainly has the potential to result in further improvements in diagnostic accuracy. Our study serves as the first step in this process. Future studies can determine whether there is any benefit in combining the staging models together. This is a very complex question that might even lead to the creation of new equations based on combination models that use similar lab values as FIB-4 and NFS but combine them with MRE results.

CHANGES IN THE TEXT 3: We added a few sentences to the introduction (lines 53-55 and 82-84).

COMMENT 4: Whilst MRE is recognized as an excellent marker of fibrosis, I am unaware of any society guidelines where it is accepted as an alternative to liver biopsy, and also as mentioned, MRE suffers criticism from is the lack of pre-specified thresholds in the literature. It is good to see in the paper the thresholds were seemingly taken from those used in clinical practice from the clinical care setting in which it is being used, but I think validation against histology for these thresholds would be a useful addition to the paper - in a sub group would suffice if this data is available.

REPLY 4: Our study is to assess how well MRE compares to FIB-4 and NSF, not to compare these to biopsy. We agree that biopsy remains the gold standard in NAFLD and made sure not to refer to MRE as a current gold standard in our manuscript. We did however, discuss how it might potentially become a gold standard in the future as technological advancements and further statistical analysis occur. We included citations for articles supporting the accuracy and utility of MRE (for example https://pubmed.ncbi.nlm.nih.gov/30582669/). We also discussed how MRE is an alternative in cases where liver biopsy is unable to be performed. Our study used lab values drawn within 6 months or MRE so that they could be accurately compared. We would be unable to accurately compare our MRE, FIB-4 and NSF results to biopsy for a few reasons. The first is that biopsy was not performed on many patients in our study. We could potentially still look at those who did have a biopsy, however a very large percentage of these biopsies weren't done in an appropriate time frame which would be acceptable for comparison to the MRE and lab draws that we utilized. In our retrospective study, we had practically no patients with biopsy and MRE performed in the same year so comparing results from years apart would lack statistical significance.

CHANGES IN THE TEXT 4: None.

COMMENT 5: In the introduction it is rightly mentioned that there are no approved drugs for NAFLD however I think it is accepted that catching those in the earlier stages of the disease when regression is more likely is imperative, especially given the growing disease prevalence. I think there is again opportunity to explore the best clinical pathway to stage and monitor disease progression/regression as we know lifestyle changes do work when adhered to, but motivating patients to make changes is very hard. The optimal pathway needs to be sensitive enough to detect worrying

and/or positive changes in NASH and not just in fibrosis in order to ensure the right advice is given. Imaging biomarkers may have more of a role in the monitoring of those in the earlier stages of disease to detect patterns of changes across the entire organ and also to have a visual representation to show the patient to help with motivation. An analysis of the agreement of staging those excluding significant (F2) fibrosis rather that only advanced (F3) would be good, especially as this is the likely target group to received drugs when they do come to market.

REPLY 5: We use the terminology and groupings found in our study as they are based upon the terminology used by those who developed the fibrosis scoring models. The middle categories in FIB-4 and NFS are considered grey zones with the terminology used ("Further Investigation Needed" and "Indeterminate") suggesting that the data in these categories is not reliable. The value of these biomarker scoring models is based upon their abilities to either identify or rule out advanced fibrosis (https://onlinelibrary.wiley.com/doi/10.1002/edm2.127). We agree that being able to do research built around excluding significant fibrosis as opposed to advanced fibrosis might open the door for further analysis and potential advancements in this field. However, we do not feel that it would be just to group the Advanced Fibrosis Excluded group with the Further Investigation Needed and Indeterminate groups.as the data for results within the Further Investigation Needed and Indeterminate ranges has proven to be very unreliable. As a result, we left these groups as a grey zone not included in our target groups is more appropriate.

CHANGES IN THE TEXT 5: None.