Clinical Gastroenterology and Hepatology 2017;15:1453–1460

Diagnostic Accuracy of Noninvasive Fibrosis Scores in a Population of Individuals With a Low Prevalence of Fibrosis

Suzanne E. Mahady,[*,‡] Petra Macaskill,[*] Jonathan C. Craig,[*,§] Grace L. H. Wong,[‖] Winnie C. W. Chu,[¶] Henry L. Y. Chan,[‖] Jacob George,[‡,a] and Vincent W. S. Wong[‖,#,a]

*Sydney School of Public Health, University of Sydney, Sydney, Australia; [‡]Storr Liver Centre, Westmead Millennium Institute for Medical Research and Westmead Hospital, University of Sydney, Sydney, Australia; [§]Centre for Kidney Research, Kids Research Institute, Children's Hospital at Westmead, Westmead, Australia; [‖]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong; [¶]Department of Imaging and Interventional Radiology, The Chinese University of Hong Kong, Hong Kong; and [#]State Key Laboratory of Digestive Disease, The Chinese University of Hong Kong, Hong Kong

This article has an accompanying continuing medical education activity, also eligible for MOC credit, on page e148. Learning Objective–Upon completion of this activity, successful learners will be able to compare the difference in prevalence of advanced liver fibrosis in different healthcare settings, explain the most appropriate use of noninvasive fibrosis scoring systems in primary care, and identify the most important predictor of progression to end stage liver disease in NAFLD/NASH patients.

BACKGROUND & AIMS: **Noninvasive scoring systems for fibrosis are increasingly used in the clinic and in research because of their ease of use, accessibility, and low cost. However, test performance characteristics were established in groups of patients with a high prevalence of advanced fibrosis; little is known about diagnostic accuracy in low-risk populations.**

METHODS: **In a cross-sectional study, 922 members of a general ambulatory population in Hong Kong (randomly selected; 18–70 years old) underwent clinical assessment from May 2008 through December 2010. All participants completed a standard questionnaire that collected information on age, sex, and history of smoking and alcohol use. Results of fasting blood tests and transient elastography were used as the reference standard to identify patients with advanced fibrosis. We assessed performance characteristics of 3 noninvasive fibrosis scoring systems: the nonalcoholic fatty liver disease fibrosis scoring system, the Fibrosis-4 scoring system, and aspartate transaminase to platelet ratio index, using standard thresholds. To calculate diagnostic test characteristics, we constructed a 2-by-2 table with the presence or absence of advanced fibrosis according to the transient elastography reading against the presence or absence of advanced fibrosis according to the scoring systems. Area under the receiver operating curve was calculated to assess overall diagnostic accuracy.**

RESULTS: **Of the 922 individuals evaluated by transient elastography, 749 had a valid reading and 15 had advanced fibrosis (2%). The specificity of noninvasive scores in detection of advanced fibrosis approximated 100% (95% confidence interval [CI], 99%–100%), with a negative predictive value of 98% (95% CI, 97%–99%) for all systems. However, the scoring systems detected fibrosis with a low level of sensitivity, ranging from 7% (95% CI, 0%–32%) to 13% (95% CI, 2%–40%). Positive predictive values ranged from 50% (95% CI, 7%–93%) to 67% (95% CI, 9%–99%). Their negative likelihood ratios ranged from 0.87 (95% CI, 0.71%–1.06%) to 0.93 (95% CI, 0.82%–1.07%); positive likelihood ratios were uninformative because of the small number of people with positive scores.**

CONCLUSIONS: **In low-risk populations, negative results from noninvasive scoring systems reliably exclude advanced fibrosis, without requirements for further tests. Positive test results are often a false-positive result and should prompt further testing.**

Keywords: Sensitivity; Specificity; Fatty Liver; Liver Fibrosis; Noninvasive Diagnosis.

[a]Authors share co-senior authorship.

Abbreviations used in this paper: APRI, aspartate transaminase to platelet ratio index; CI, confidence interval; FIB-4, Fibrosis-4; NAFLD, nonalcoholic fatty liver disease; NFS, NAFLD fibrosis score.

Most current article

1542-3565/$36.00
http://dx.doi.org/10.1016/j.cgh.2017.02.031

See editorial on page 1355.

Long-term cohort studies have identified advanced fibrosis as the strongest predictor for progression to cirrhosis in nonalcoholic fatty liver disease (NAFLD).[1–5] These individuals have an increased risk of liver failure and its sequelae that is proportional to the degree of fibrosis, and is independent of established risk factors, such as age and diabetic status.[1] The ability to accurately identify those with advanced fibrosis is of considerable clinical and public health importance, because NAFLD is the most prevalent liver disease worldwide.[6,7] Fibrosis has traditionally been assessed using liver biopsy as the reference standard, but this is not feasible for a disease with high prevalence because of logistics, safety concerns, and cost.[8] This has prompted the development of noninvasive scoring systems, such as the NAFLD Fibrosis Score (NFS),[9] aspartate transaminase to platelet ratio index (APRI),[10] and Fibrosis (FIB)-4.[11] These scores use a panel of readily available anthropometric and biochemical variables and were developed in populations with a high prevalence of advanced fibrosis up to 30%, with excellent diagnostic accuracy (positive predictive value of 90% and negative predictive values of 85% for NFS).[9]

Because of their ease of use, noninvasive scoring systems have been increasingly applied to a diverse array of clinical and research populations where the prevalence of advanced fibrosis varies. However, given that test performance is influenced by the spectrum of disease in a given population,[12] the validity of these approaches is unclear. Noninvasive scores have been proposed as a screening tool for fibrosis in primary care populations,[13,14] population-based diabetes clinics,[15] and as part of the diagnostic algorithm for liver disease in low-risk settings.[16] They have also been proposed as a tool to identify participants for clinical trials in NAFLD[17] and to isolate persons with advanced fibrosis in epidemiologic cohorts, to evaluate associations of fibrosis and mortality.[18] In addition, there is discordance on the best use of these scoring systems in current clinical practice guidelines. Guidelines from the American Association for the Study of Liver Diseases recommend that noninvasive scoring systems should be used to identify patients with fibrosis irrespective of population prevalence,[7] whereas the European Association for the Study of the Liver states that noninvasive scoring systems should be used to identify those with advanced fibrosis who are at low risk.[19] Noninvasive scoring systems have also been advocated as first-line testing in children.[19] However, because diagnostic accuracy has not been studied in low-risk populations, evidence to support many of these recommendations is lacking.

In this study, we evaluated the test performance of noninvasive scores in a general ambulatory population, and provide data on diagnostic accuracy that can be used to inform the applicability of noninvasive scores in low-risk settings.

Methods

Study Design and Conduct

This cross-sectional study is reported following the STARD guidelines for diagnostic accuracy research[20] and relevant sections of the Liver-FibroSTARD guidelines.[21] As reported previously,[22] between May 2008 and December 2010, members of the general population listed on the government census in Hong Kong and aged 18–70 were randomly invited by mail or telephone to participate. Patients were excluded if they had known active malignancy; positive viral hepatitis serology (hepatitis B surface antigen or hepatitis C antibody); existing decompensated liver disease; or a secondary cause for fatty liver, such as tamoxifen or amiodarone. Baseline data were collected prospectively before the index and reference tests were performed. All participants provided written, informed consent and ethical approval was granted by the Clinical Research Ethics Committee of the Chinese University of Hong Kong.

Data Collection

All participants underwent clinical history using a standardized questionnaire that collected information on age, gender, smoking, and alcohol history. A threshold of daily alcohol intake of 20 g for men and 10 g for women was used to exclude fatty liver caused by excess alcohol intake. Impaired fasting glucose was defined as serum glucose between 5.6 mmol/L and 6.9 mmol/L. Diabetes was defined as fasting glucose $\geq$7.0 mmol/L or previous diagnosis of diabetes. Hypertension was defined as blood pressure $>$140/90 mm Hg, with the average of 2 measurements taken in the seated position reported. Weight was recorded on digital scales, and body mass index was calculated as weight/height^2 (kg/m^2). Waist circumference was taken at the midpoint between the lowest rib and top of iliac crest and measured in centimeters. Fasting blood samples were taken by qualified phlebotomists in the morning for all participants, and analyses for liver biochemistry, full blood count, serum triglycerides, iron studies, and fasting glucose were performed.

Reference Standard: Transient Elastography

Transient elastography (Fibroscan, Echosens, Paris, France) was used as the reference standard for fibrosis. Although liver biopsy is the traditional reference standard for fibrosis, transient elastography measures liver stiffness over an area that is approximately 100 times greater than liver biopsy, is noninvasive, has no significant side effects,[23] and is increasingly used in clinical practice as the reference standard for fibrosis to avoid liver biopsy. Assessment of liver stiffness was performed in fasting state on the same morning as the blood tests and clinical evaluation for all participants. Measurements were

performed according to the manufacturer's instructions and have been described in detail previously.[22] Briefly, measurements were taken on the right lobe of the liver through intercostal spaces with the right arm in abduction, by 1 of 2 operators who had performed more than 500 examinations each before the study. They were reported as median values in kilopascal and were considered acceptable if the following criteria were met: success rate was >60% for each person (defined as number of successful readings divided by the total number of readings for that person), 10 successful acquisitions were obtained, and the interquartile range to median ratio of the 10 acquisitions was <0.3. A threshold of 9.6 kPa was used to define advanced fibrosis, consistent with previous validation studies that have shown a specificity of >90% for advanced fibrosis when compared against a reference standard of liver biopsy.[24]

Index Tests: Nonalcoholic Fatty Liver Disease Fibrosis Score, Fibrosis-4, and Aspartate Transaminase to Platelet Ratio Index

NFS was originally developed in a cohort of predominantly white patients attending tertiary level liver clinics, with a high prevalence of advanced fibrosis.[9] The score has since been validated in other ethnicities including Chinese[25] and other subgroups, such as those with morbid obesity.[26] NFS was calculated for each person using the published formula: -1.675 + 0.037 × age (years) + 0.094 × body mass index (kg/m^2) +1.13 × impaired fasting glucose/diabetes (yes = 1, no = 0) + 0.99 × aspartate transaminase/alanine transaminase ratio - 0.013 × platelet count (10^9/L) - 0.66 × albumin (g/dL). A cutoff of NFS of >0.676 was used as the upper threshold, and < -1.455 as the lower threshold, because these thresholds have been previously established to have maximal diagnostic accuracy according to area under the receiver operating characteristic curve.[9] The FIB-4 score is a noninvasive fibrosis score consisting of 4 variables (age, alanine transaminase, aspartate transaminase, and platelet count) and was originally developed to assess fibrosis in hepatitis C and human immunodeficiency virus coinfection[11] with subsequent validation in people with NAFLD.[27] An upper threshold of 3.25 to rule in advanced fibrosis, and a lower threshold of 1.3 to exclude fibrosis were applied.[27] In addition, we used the APRI, developed in patients with hepatitis C[28] and validated in NAFLD,[29,30] with a conservative threshold of 1.0 as the cutoff for advanced fibrosis.[29] To ensure accuracy of calculation of noninvasive scores, random results were verified with online calculators.[31]

Statistical Analysis

For descriptive statistics, the means and 95% confidence interval (CI) were estimated for normally distributed continuous variables, and the median and interquartile ranges were calculated for nonnormally distributed variables. For categorical variables, frequency tables were constructed and values reported in percentages. To calculate diagnostic test characteristics, a 2-by-2 table was constructed with presence or absence of advanced fibrosis according to transient elastography reading, against presence or absence of advanced fibrosis according to noninvasive scores using the established thresholds. Sensitivity, specificity, positive and negative predictive values, and exact (Clopper-Pearson) 95% CI were calculated. Although sensitivity and specificity inform how well a test differentiates true-positive and true-negative people within a population, predictive values inform on how likely a positive or negative test result is to be correct given the disease prevalence within that population. Likelihood ratios with 95% CI were computed to allow interpretation of test results in individual patients, because tests with a positive likelihood ratio >10 or a negative likelihood ratio <0.1 are considered clinically useful because they substantially change the posttest probability of disease.[32] The overall diagnostic accuracy of noninvasive scores was estimated by plotting a receiver operating curve of sensitivity versus 1- specificity, and estimating the area under the curve, using designated thresholds. We also aimed to assess diagnostic accuracy using a combination of scores to determine if this approach offered incremental diagnostic value; however, this was not possible because of the small number of positive scores. All statistical analyses were undertaken in SAS version 9.4 (SAS institute, Cary, NC).

Results

Study Cohort

The selection of the study cohort is shown in Figure 1. Of 3000 people invited to participate, 1069 responded (36% response rate), of which 147 participants were excluded because of hepatitis B or C infection or contraindication to imaging.[22] Overall, 922 people underwent clinical assessment, blood tests, and transient elastography. A further 163 people were excluded because of a lack of valid transient elastography measurements, similar to success rates in other populations,[33] either because 10 successful acquisitions were not obtained (n = 60) or the interquartile range to median ratio was inadequate for a valid reading (n = 103). A total of 759 people completed a successful scan. A further 10 people did not have the full set of variables required for calculation of scores, leaving a total of 749 people with both valid transient elastography and score calculation as the study cohort. Despite these exclusions, the remaining population seemed to be typical of the population at large.[34]

Patient Characteristics

The demographic and clinical characteristics of the study population are shown in Table 1. The mean age

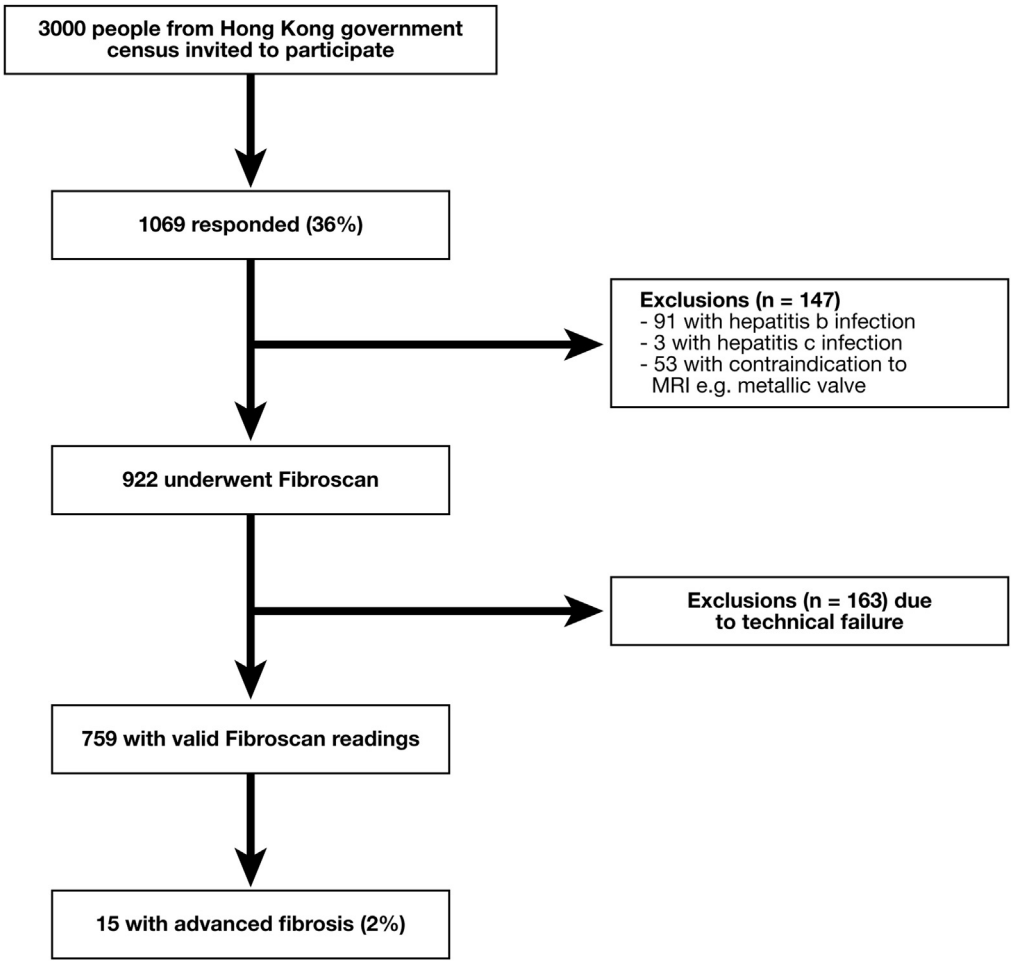


Figure 1. Study cohort. MRI, magnetic resonance imaging.

was approximately 48 years old, with slightly more males than females (55% vs 45%). The mean body mass index was 22.6 kg/m^2, with prevalence of diabetes of approximately 1 in 20 persons (4.5%). Median liver enzymes were in the normal range (alanine transaminase, 22; interquartile range, 17–31). The median Fibroscan score in the overall population was 4.2 kPa (interquartile range, 3.6–5.2).

Receiver Operating Curves For Noninvasive Fibrosis Scores

A receiver operating curve was constructed to demonstrate the trade-off between sensitivity and specificity at differing thresholds of positive noninvasive fibrosis scores (Supplementary Figures 1–3). The area under the receiver operating curve was estimated at 0.63 for NFS (95% CI, 0.46–0.81), 0.64 for FIB-4 (95% CI, 0.45–0.84), and 0.71 for APRI (95% CI, 0.56–0.88).

Sensitivity, Specificity, and Predictive Values at Upper and Lower Thresholds

Of the 749 persons with valid transient elastography results, 15 had readings >9.6 kPa, resulting in a prevalence of advanced fibrosis of approximately 2% in the general population. Using the upper thresholds, only 1 person also had a positive NFS, and 2 people had a positive FIB-4 and APRI score (Table 2). For the NFS, the sensitivity for advanced fibrosis was estimated at 7% (95% CI, 0–32), with a specificity approximating 100% (95% CI, 99.2–100) (Table 3). The positive predictive value was estimated at 50% (95% CI, 1–99), indicating that for every positive result, another would be a false-positive result; however, the true value is uncertain because of the wide CI. The negative predictive value was 98% (95% CI, 97–99). For FIB-4 and APRI, the sensitivity was 13% (95% CI, 2–40) for both, with specificity for both that also approximated 100% (99–100) and a negative predictive value of 98% (97–99) (Table 3). When the lower thresholds for NFS and FIB-4 were applied, the sensitivities increased as expected; for NFS and FIB-4, sensitivity 33% (95% CI, 9–57) for both. The specificity was reduced; for NFS, 91% (88–92) and for FIB-4, 87% (85–90). The negative predictive value of the scores remained at ≥98% for both upper and lower thresholds (Table 3).

Likelihood Ratios

For test accuracy at the upper threshold, the negative likelihood ratios ranged from 0.87 to 0.93, suggesting

Table 1. Characteristics of Population With Valid Transient Elastography

Characteristic	All participants (n = 749)
Age, *y* (mean, SD)	47.8 (10.5)
Sex, males, %	55
Current smoker, %	9.7
Hypertension, %	14.1
Diabetes, yes, %	4.5
Presence of either diabetes or IFG, %	15.7
Anthropometry	
BMI, *kg/m²* (mean, SD)	22.6 (3.4)
Waist circumference, *cm* (mean, SD)	80.8 (11.8)
Biochemistry	
ALT (*IU/L*, median, IQR)	22 (17–31)
AST (*IU/L*, median, IQR)	19 (17–23)
Albumin, *g/L* (mean, SD)	45.2 (2.6)
Fasting glucose, *mmol/L* (mean, SD)	5.0 (3.9)
Platelets, *x 10⁹/L* (mean, SD)	242 (57)
Triglycerides, *mmol/L* (mean, SD)	1.25 (3.8)
Ferritin, *mmol/L* (median, IQR)	360 (121–701)
Test results, transient elastography	
Transient elastography score, *kPa* (median, IQR)	4.2 (3.6–5.2)
Interquartile range (median, IQR)	0.7 (0.4–0.9)
Index test results	
NFS (median, IQR)	-2.87 (-3.52 to -2.11)
FIB-4 (median, IQR)	0.84 (0.62–1.10)
APRI (median, IQR)	0.21 (0.17–0.26)

ALT, alanine transaminase; AST, aspartate transaminase; BMI, body mass index; IFG, impaired fasting glucose; IQR, interquartile range; SD, standard deviation.

Table 2. Frequency Table for Persons With Positive or Negative Tests Results According to Reference (Transient Elastography) and Index (Noninvasive Fibrosis Scores) Test Results Using the Recommended Upper Thresholds

	Reference test positive		Reference test negative		
	Index test positive	Index test negative	Index test positive	Index test negative	Total
NFS	1	14	1	733	749
FIB-4	2	13	2	732	749
APRI	2	13	1	733	749

that when the noninvasive score test result is negative, the posttest probability of disease is not substantially altered (Table 3). The positive likelihood ratios were high for all tests; however, the CIs were very wide because of the small number of people with positive test results indicating a high level of uncertainty in these estimates. Using the lower threshold resulted in reduced positive likelihood ratios with no meaningful change in the negative likelihood ratios (Table 3).

Diagnostic Algorithm for Clinical Practice Using the Nonalcoholic Fatty Liver Disease Fibrosis Score

Figure 2 shows a suggested diagnostic algorithm for the use of NFS according to the expected population prevalence of advanced fibrosis. Although a negative result generally indicates the absence of advanced fibrosis with a high degree of certainty across all populations, positive predictive values are highly variable.

Discussion

The results of the current population-based study of noninvasive scores for the evaluation of fibrosis in NAFLD have important implications for clinical practice. We demonstrate that in low-risk cohorts, a negative result using noninvasive scoring systems is associated with a very high specificity approximating 100% (95% CI, 99–100) and negative predictive value of 98%, indicating that people without advanced fibrosis can be confidently excluded, and with minimal chance of a missed diagnosis. When a lower threshold for exclusion of advanced fibrosis with the NFS or FIB-4 test was applied, the specificity and negative predictive values did not change to a clinically meaningful extent. Tests with a high specificity and negative predictive value are useful in clinical settings, because diagnostic uncertainty is minimized and follow-up testing can be avoided. This is particularly helpful in NAFLD, where the confirmatory tests are either risky, such as liver biopsy,[8] or expensive with limited availability, such as transient elastography.

In contrast, the sensitivity and positive predictive values of noninvasive fibrosis scores were low and imprecise. Using the upper threshold for NFS and FIB-4, the positive predictive value was estimated at 50% with a wide CI, indicating that further testing is needed to distinguish true cases from false positives. The clinical implications of a high number of false positives depend on attributes of the confirmatory test. such as side effects, in addition to the degree of anxiety associated with a false-positive diagnosis.[35] The positive likelihood ratio was high but imprecise because of the small number of people with positive test results and the true estimate remains uncertain, whereas the negative likelihood ratio was estimated with a high degree of precision and approximated 1.0, irrespective of whether upper or lower thresholds were used. This reflects the very low sensitivity regardless of threshold, in the context of a preserved test specificity. Furthermore, we did not find evidence to suggest that 1 score was substantially more accurate than another in this population.

These data concur with previous studies of diagnostic accuracy of noninvasive scores that consistently demonstrate a high specificity and negative predictive value with upper thresholds irrespective of prevalence,[29] but sensitivity and positive predictive values vary widely, resulting in different recommendations for follow-up management. In a tertiary liver clinic population where

Table 3. Test Performance Characteristics of Noninvasive Fibrosis Scores in the General Population Where the Prevalence of Advanced Fibrosis Is Low (n = 749)

Test	Cutoff	Sensitivity % 95% CI	Specificity % 95% CI	PPV % 95% CI	NPV % 95% CI	+LR Estimate 95% CI	-LR Estimate 95% CI
NFS	0.676	7 (0–32)	100 (99–100)	(50–99)	98 (97–99)	49 (3–746)	0.93 (0.82–1.07)
	—1.455	33 (9–7)	91 (88–92)	7 (1–12)	99 (97–99)	3.5 (0.9–6.2)	0.73 (0.47–1.00)
FIB-4	3.25	13 (2–40)	100 (99–100)	50 (7–93)	98 (97–99)	49 (7–325)	0.87 (0.71–1.06)
	1.3	33 (9–57)	87 (85–90)	5 (0–10)	98 (97–99)	2.6 (0.7–4.6)	0.76 (0.49–1.03)
APRI	1.0	13 (2–40)	100 (99–100)	67 (9–99)	98 (97–99)	98 (9–1021)	0.87 (0.71–1.06)

LR, likelihood ratio; NPV, negative predictive value; PPV, positive predictive value.

the prevalence of advanced fibrosis exceeded 30%, positive predictive values were estimated at 90% with a positive likelihood ratio of 11,[9] suggesting a high degree of confidence in the diagnosis. In a separate liver clinic cohort where the prevalence of advanced fibrosis was 20%, the positive predictive values were lower at 30%–37%.[29] In population-based type 2 diabetics where the prevalence of advanced fibrosis was 5%, the positive predictive values of FIB-4 and APRI were 13%.[36] Follow-up testing for a positive test result may include transient elastography, magnetic resonance spectroscopy, or liver biopsy, all associated with substantial cost, inconvenience, or risk, but whether these approaches improve patient-important outcomes has not been studied. In general, diagnostic tests should not just be considered within a narrow framework of test accuracy, but before they are recommended for widespread use, should be supported by evidence that they ultimately improve patient outcomes and are cost effective.[37] These conclusions are best established by testing within the framework of a randomized clinical trial.[38]

Our findings also inform the use of noninvasive scores for the case detection of advanced fibrosis in large epidemiologic studies. In a recent examination of the National Health and Nutrition Survey in the US cohort, 3.2% of the population or 1.4 million Americans were classified as having advanced fibrosis according to the standard threshold using the NFS. This group was found to have a threefold increase in cardiovascular mortality compared with those with a negative score.[18] Given the low positive predictive value found in the current study, more than half of those with a positive test result are likely to be a false-positive result, making accurate conclusions about mortality difficult to calculate. Similarly,

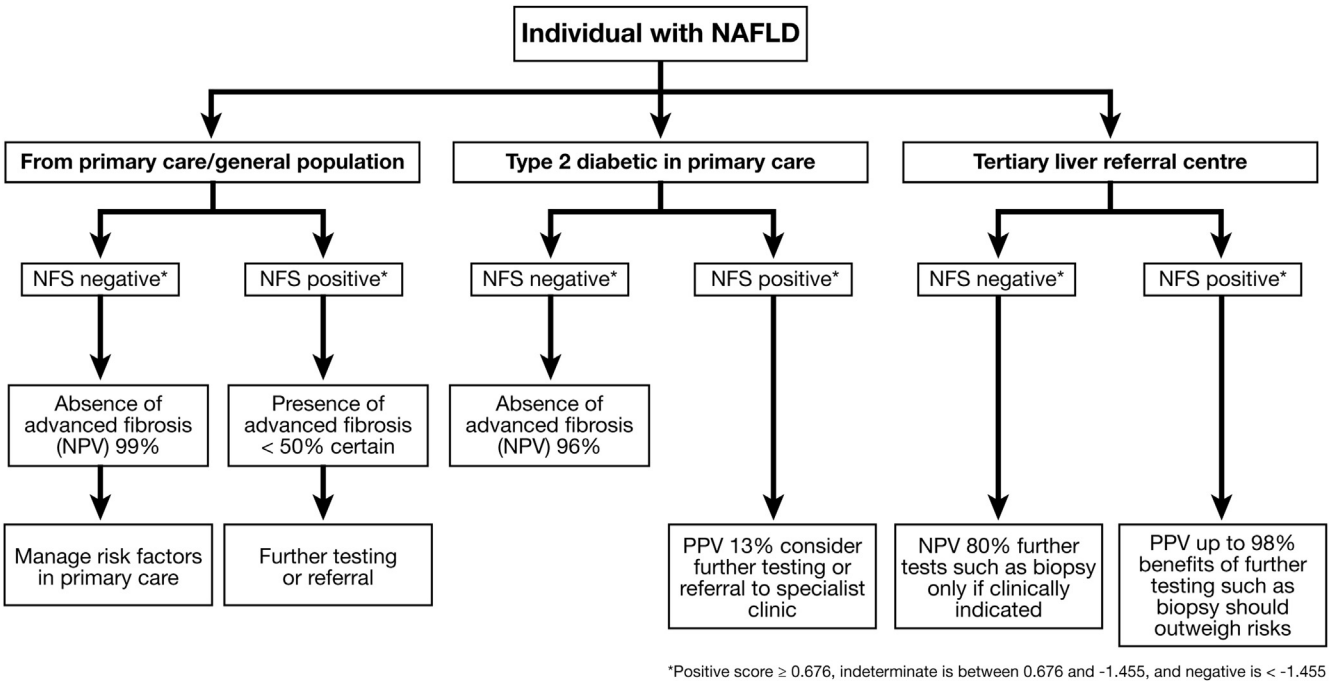


Figure 2. Diagnostic algorithm describing predictive values of a positive or negative NFS in populations with a different prevalence of advanced fibrosis. Recommendations for follow-up testing are shown. Positive NFS is >0.676, negative is < -1.455. Data for primary care population taken from current study, data for type 2 diabetic population taken from Armstrong et al,[15] and data for high-prevalence population taken from Angulo et al.[9] NPV, negative predictive value; PPV, positive predictive value.

although current international clinical practice guidelines suggest that noninvasive scores should be used for case finding,[7] our and other data suggest that on the contrary, these tests have highest diagnostic yield when used to exclude advanced fibrosis in low-risk people rather than for case detection.

Although it is accepted that predictive values vary with the population tested, our data have also shown a change in sensitivity and specificity that should be explained. The sensitivity of the NFS was estimated at 43% in the original cohort,[9] whereas it was 7% in our study. Sensitivity and specificity vary when the disease severity differs, a phenomenon known as spectrum bias.[39] This arises because diagnostic tests are usually developed in cohorts with a severe form of the disorder and tested against a normal group, making the disease easy to detect, which optimizes diagnostic accuracy. However, when the same test is applied to a population where the disease spectrum includes less obvious cases, the sensitivity falls.[39] Changes in sensitivity and specificity and predictive values should be expected when applying these scoring systems to a low-risk population.

The major limitation to this study was the small number of people with positive test scores. Although this truthfully reflected the low prevalence within the general population, sensitivity and positive predictive values could not be calculated with precision, and the range within which we can be 95% confident that the true value lies, is extremely wide. A larger sample size can overcome this but it is challenging to perform a large number of transient elastography scans in the general population using the rigorous methodology of the present study. Indeed, assuming a test sensitivity of 50%, the sample size required for a 95% CI width of 10% in a population where the prevalence is 2% approximates 20,000 persons.[40] In addition, we undertook transient elastography as the reference standard for fibrosis, because liver biopsy is not a feasible test in large studies, and measurement error may be associated with this approach. To accurately determine the degree of measurement error, data on the results of all 3 tests in the same population are needed; these data would be very difficult to obtain. Despite this, measurement error caused by a suboptimal reference standard is likely to result in misclassification of patients with bias of results towards the null, resulting in underestimation of test accuracy. Our population consisted of those with exclusively Asian ancestry, and whether these results can be generalized to other ethnic populations, such as white persons, is unclear. Finally, selection bias may have affected our results because there are no data available on nonrespondents to know if they are comparable; however, the participants seemed to reflect the population at large.[34] There were several strengths to our approach, including a large, general population based sample; an absence of time delay between index and reference tests meaning that disease status was unchanged; and reporting was performed according to standard recommendations,[20] such that the methodology and results are explicit.

Although these data provide important information for the interpretation of noninvasive scores in low-risk populations, they do not support the widespread adoption of noninvasive scoring systems for fibrosis screening in such situations as primary care. Further research regarding whether the use of noninvasive scores can ultimately improve patient-important outcomes is needed. In addition, whether scores can be used in a serial manner, and how much change represents a true change in disease status, is unknown. Validation of our results in other ethnic cohorts would also be helpful.

In conclusion, noninvasive fibrosis scoring systems can reliably exclude the presence of advanced fibrosis in low-risk populations without the need for further testing, and with minimal risk of a missed diagnosis. However, scoring systems are insufficiently sensitive for case detection. Positive test results are associated with a high rate of false positives and should prompt further testing.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Clinical Gastroenterology and Hepatology* at www.cghjournal.org, and at http://dx.doi.org/10.1016/j.cgh.2017.02.031.

References

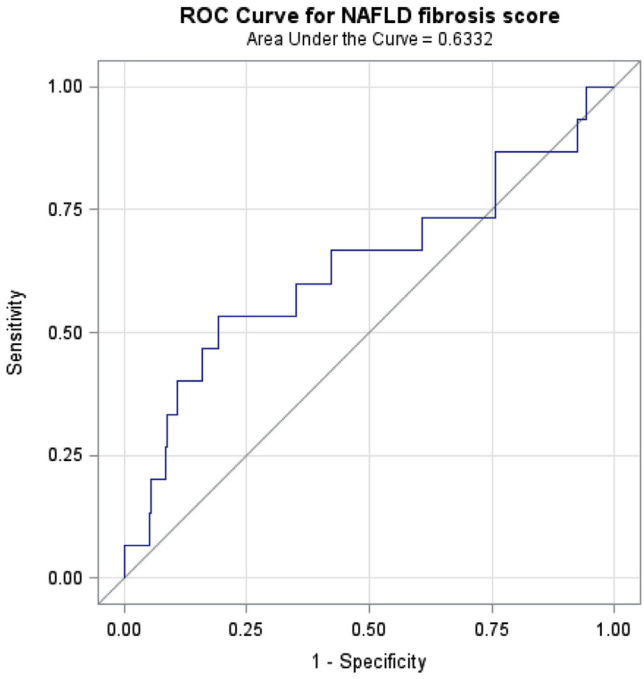
1. Angulo P, Kleiner DE, Dam-Larsen S, et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology 2015;149:389–397.
2. Younossi ZM, Stepanova M, Rafiq N, et al. Pathologic criteria for nonalcoholic steatohepatitis: interprotocol agreement and ability to predict liver-related mortality. Hepatology 2011; 53:1874–1882.
3. Ekstedt M, Hagstrom H, Nasr P, et al. Fibrosis stage is the strongest predictor for disease-specific mortality in NAFLD after up to 33 years of follow-up. Hepatology 2015;61:1547–1554.
4. Matteoni CA, Younossi ZM, Gramlich T, et al. Nonalcoholic fatty liver disease: a spectrum of clinical and pathological severity. Gastroenterology 1999;116:1413–1419.
5. Haflidadottir S, Jonasson JG, Norland H, et al. Long-term follow-up and liver-related death rate in patients with non-alcoholic and alcoholic related fatty liver disease. BMC Gastroenterol 2014;14:166.
6. Browning JD, Szczepaniak LS, Dobbins R, et al. Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. Hepatology 2004;40:1387–1395.
7. Chalasani N, Younossi Z, Lavine JE, et al. The diagnosis and management of non-alcoholic fatty liver disease: practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association. Hepatology 2012; 55:2005–2023.
8. Grant A, Neuberger J. Guidelines on the use of liver biopsy in clinical practice. British Society of Gastroenterology. Gut 1999; 45(Suppl 4):IV1–IV11.
9. Angulo P, Hui JM, Marchesini G, et al. The NAFLD Fibrosis Score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology 2007;45:846–854.

10. Lin ZH, Xin YN, Dong QJ, et al. Performance of the aspartate aminotransferase-to-platelet ratio index for the staging of hepatitis C-related fibrosis: an updated meta-analysis. Hepatology 2011;53:726–736.
11. Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006;43:1317–1325.
12. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009;62:5–12.
13. Treeprasertsuk S, Bjornsson E, Enders F, et al. NAFLD Fibrosis Score: a prognostic predictor for mortality and liver complications among NAFLD patients. World J Gastroenterol 2013; 19:1219–1229.
14. Poynard T, Lebray P, Ingiliz P, et al. Prevalence of liver fibrosis and risk factors in a general population using non-invasive biomarkers (FibroTest). BMC Gastroenterol 2010;10:40.
15. Armstrong MJ, Hazlehurst JM, Parker R, et al. Severe asymptomatic non-alcoholic fatty liver disease in routine diabetes care; a multi-disciplinary team approach to diagnosis and management. Q J Med 2014;107:33–41.
16. Dowman JK, Tomlinson JW, Newsome PN. Systematic review: the diagnosis and staging of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis. Aliment Pharmacol Ther 2011;33:525–540.
17. Sanyal AJ, Friedman SL, McCullough AJ, et al. Challenges and opportunities in drug and biomarker development for nonalcoholic steatohepatitis: findings and recommendations from an American Association for the Study of Liver Diseases–U.S. Food and Drug Administration Joint Workshop. Hepatology 2015; 61:1392–1405.
18. Kim D, Kim WR, Kim HJ, et al. Association between noninvasive fibrosis markers and mortality among adults with nonalcoholic fatty liver disease in the United States. Hepatology 2013; 57:1357–1365.
19. European Association for the Study of the Liver. EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease. J Hepatol 2016;64:1388–1402.
20. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:5527.
21. Boursier J, de Ledinghen V, Poynard T, et al. An extension of STARD statements for reporting diagnostic accuracy studies on liver fibrosis tests: the Liver-FibroSTARD standards. J Hepatol 2015;62:807–815.
22. Wong VW, Chu WC, Wong GL, et al. Prevalence of non-alcoholic fatty liver disease and advanced fibrosis in Hong Kong Chinese: a population study using proton-magnetic resonance spectroscopy and transient elastography. Gut 2012;61:409–415.
23. Roulot D, Costes JL, Buyck JF, et al. Transient elastography as a screening tool for liver fibrosis and cirrhosis in a community-based population aged over 45 years. Gut 2011; 60:977–984.
24. Wong VW, Vergniol J, Wong GL, et al. Diagnosis of fibrosis and cirrhosis using liver stiffness measurement in nonalcoholic fatty liver disease. Hepatology 2010;51:454–462.
25. Wong VW, Wong GL, Chim AM, et al. Validation of the NAFLD fibrosis score in a Chinese population with low prevalence of advanced fibrosis. Am J Gastroenterol 2008;103:1682–1688.
26. Qureshi K, Clements RH, Abrams GA. The utility of the "NAFLD fibrosis score" in morbidly obese subjects with NAFLD. Obes Surg 2008;18:264–270.
27. Shah AG, Lydecker A, Murray K, et al. Comparison of noninvasive markers of fibrosis in patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2009;7:1104–1112.
28. Wai CT, Greenson JK, Fontana RJ, et al. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. Hepatology 2003;38:518–526.
29. McPherson S, Stewart SF, Henderson E, et al. Simple non-invasive fibrosis scoring systems can reliably exclude advanced fibrosis in patients with non-alcoholic fatty liver disease. Gut 2010;59:1265–1269.
30. Kruger FC, Daniels CR, Kidd M, et al. APRI: a simple bedside marker for advanced fibrosis that can avoid liver biopsy in patients with NAFLD/NASH. South African Medical Journal Suid-Afrikaanse Tydskrif Vir Geneeskunde 2011;101:477–480.
31. http://nafldscore.com/. [Accessed October 1, 2015].
32. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1994;271:389–391.
33. Foucher J, Castera L, Bernard PH, et al. Prevalence and factors associated with failure of liver stiffness measurement using FibroScan in a prospective study of 2114 examinations. Eur J Gastroenterol Hepatol 2006;18:411–412.
34. http://www.census2011.gov.hk/en/census-result.html. [Accessed June 1, 2016].
35. Brewer NT, Salz T, Lillie SE. Systematic review: the long-term effects of false-positive mammograms. Ann Intern Med 2007; 146:502–510.
36. Morling JR, Fallowfield JA, Guha IN, et al. Using non-invasive biomarkers to identify hepatic fibrosis in people with type 2 diabetes mellitus: the Edinburgh type 2 diabetes study. J Hepatol 2014;60:384–391.
37. Bossuyt PM, Reitsma JB, Linnet K, et al. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem 2012; 58:1636–1643.
38. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144:850–855.
39. Montori VM, Wyer P, Newman TB, et al. Tips for learners of evidence-based medicine: 5. The effect of spectrum of disease on the performance of diagnostic tests. CMAJ 2005; 173:385–390.
40. PASS 13 Power Analysis and Sample Size Software (2014). NCSS, LLC. Kaysville, Utah. Available at www.ncss.com/software/pass. Accessed May 1, 2016.

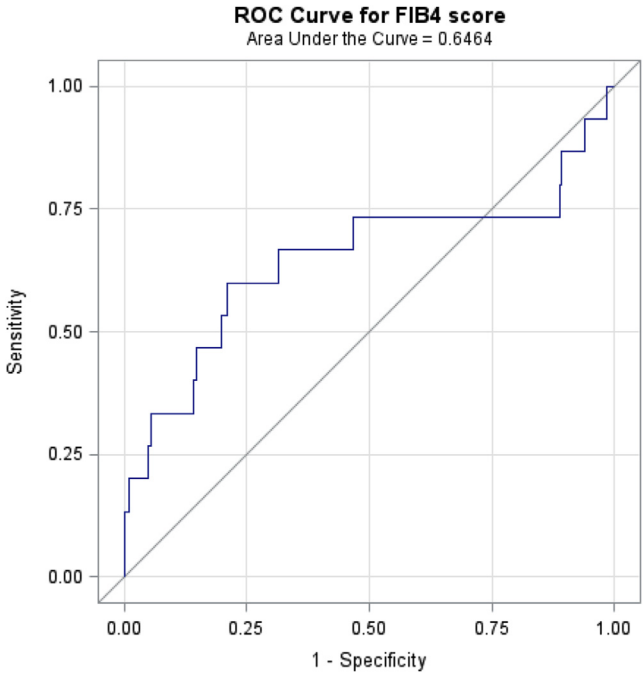
Reprint requests
Address requests for reprints to: Jacob George, PhD, Storr Liver Centre, Westmead Millennium Institute for Medical Research, University of Sydney and Westmead Hospital, Hawkesbury Road, Westmead 2145, Australia. e-mail: jacob.george@sydney.edu.au; fax: +612 9635 7582.

Conflicts of interest
The authors disclose no conflicts.

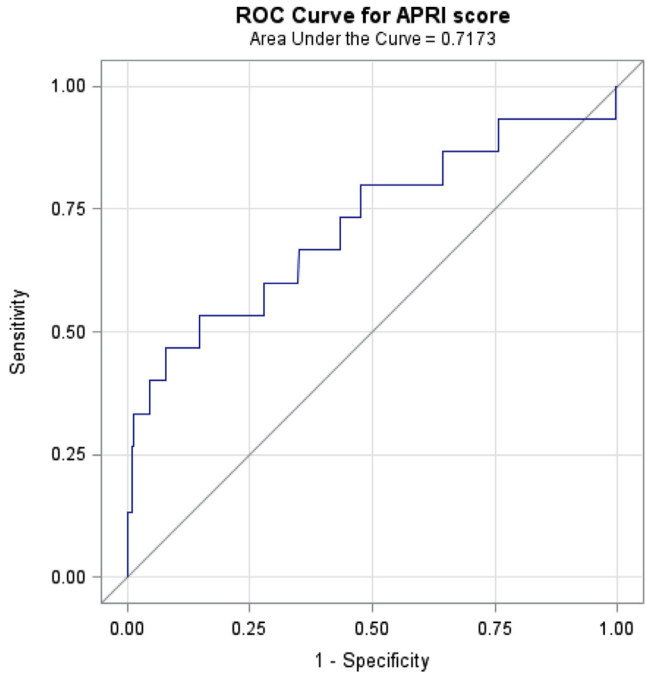
Funding
Suzanne E. Mahady is supported by a postgraduate research scholarship (1055948) funded by the National Health & Medical Research Council.



Supplementary Figure 1. ROC curve for diagnostic accuracy of NFS in a population with a low prevalence of advanced fibrosis. ROC, receiver operating characteristic curve.



Supplementary Figure 2. ROC curve for diagnostic accuracy of FIB-4 score in a population with a low prevalence of advanced fibrosis. ROC, receiver operating characteristic curve.



Supplementary Figure 3. ROC curve for diagnostic accuracy of APRI score in a population with a low prevalence of advanced fibrosis. ROC, receiver operating characteristic curve.