Moving test accuracy research forward

Lee, J.A.

Publication date
2023

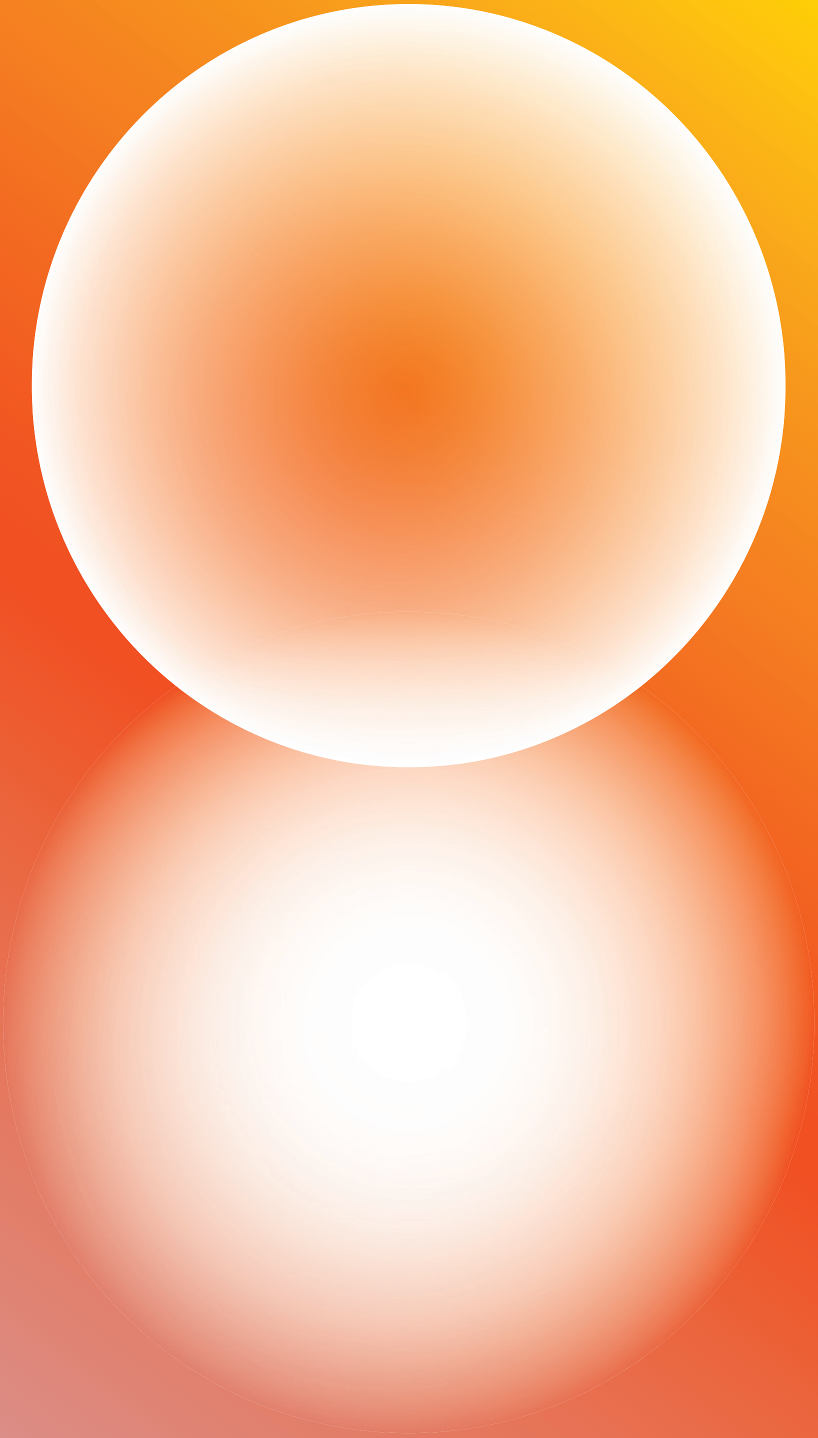
Document Version
Final published version

Link to publication

Citation for published version (APA):
Lee, J. A. (2023). *Moving test accuracy research forward*. [Thesis, fully internal, Universiteit van Amsterdam].

PHD THESIS 2023

JENNY LEE



MOVING TEST ACCURACY RESEARCH FORWARD

Moving Test Accuracy Research Forward

JENNY LEE

ISBN: 978-94-6469-454-3
Cover design: Lucille Fenouil | www.lucillefenouil.com
Printing: ProefschriftMaken | www.proefschriftmaken.nl

Financial support for this thesis was provided by the Liver Investigation: Testing Marker Utility in Steatohepatitis (LITMUS) project, which received funding from the Innovative Medicines Initiative 2 Joint Undertaking, under grant agreement No. 777377. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. Financial support for printing this thesis was provided by AMC medical research.

Moving Test Accuracy Research Forward

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 8 september 2023, te 13.00 uur

door Jin A Lee
geboren te Tokyo

Promotiecommissie

Promotor:	prof. dr. P.M.M. Bossuyt	AMC-UvA
Copromotor:	dr. M.H. Zafarmand	AMC-UvA
Overige leden:	prof. dr. M. Nieuwdorp	AMC-UvA
	prof. dr. A. Abu-Hanna	AMC-UvA
	prof. dr. R. de Jonge	Vrije Universiteit Amsterdam
	prof. dr. S.M.	Francque Universiteit Antwerpen
	dr. M.M.G. Leeflang	AMC-UvA
	prof. dr. Y. Takwoingi	University of Birmingham

Faculteit der Geneeskunde

Contents

01

General Introduction

General introduction

The COVID-19 pandemic brought the significance of diagnostic tests to the frontline of our society. Copious amounts of tests and testing strategies to detect the SARS-CoV-2 virus were introduced and the results of these tests dictated our not only healthcare providers in alleviating the burden of the outbreak, but also the mobility and personal decisions of millions around the globe. Unanimous efforts were invested into developing rapid and accurate diagnostics to detect, monitor, and prevent the spread of COVID-19. We relied on technologies like polymerase chain reaction (PCR), antigens, and antibodies for developing tests to detect this novel disease. But with the rapidly changing landscape of diagnostics came the question: how accurate are these tests?

When we study test accuracy, it is important to consider the various settings that medical tests are used in. The US Food and Drug Administration (FDA) outlined several different contexts of use for biomarkers, including diagnostic, monitoring, and predictive, among others. The context of use will often determine the most suitable approach for evaluating test performance, as well as emphasize different features of a test that are most important. Ideally, we have a perfect test that maximizes both the true negative and positive results, but this is seldom the case and there is a trade-off between the two that we need to balance. The performance of a test may also vary depending on patient characteristics. Some tests are more accurate for men compared to women, while other tests may perform better with younger patients compared to the elderly. Understanding the accuracy of a test therefore encompasses a more general evaluation of whether a test is fit for purpose and serves its intended use, as well as understanding its more operational characteristics, which can improve our understanding of optimal and suboptimal settings to use a test.

To investigate the performance of a biomarker that could be used to detect a disease or condition (here and now), we conduct diagnostic accuracy studies. This field of research has grown tremendously, and it is no surprise why. There is a lot of excitement around discovering and validating biomarkers that carry the potential to aid disease detection. Their use has proliferated to various decision points along disease management, and clinical trials for drug development. In such clinical trial settings, biomarkers can be implemented as tools to expedite recruitment, as a screening test to enrich the recruitment pool with those more likely to benefit from receiving treatment, reducing the burdens that come with more invasive and costly testing procedures.

A diagnostic accuracy study is designed in such a way that the index test is evaluated by comparing its results to the best available method to detect the disease or condition, known as the reference standard. Most often, both the test and the reference standard generate a dichotomous result, which classifies the patient as either positive or negative, and the target condition as present or absent. These data, constructed in the form of a 2x2 contingency table, allow us to calculate performance metrics such as sensitivity, specificity, and predictive values (Table 1).

Table 1. A 2x2 contingency table including sensitivity, specificity, and predictive values

		Reference Standard		
		Condition present (+)	Condition absent (-)	
Index Test	Test positive (+)	True positives (TP)	False positives (FP)	Positive predictive value (PPV) TP/(TP+FP)
	Test negative (-)	False negatives (FN)	True negative (TN)	Negative predictive value (PPV) TN/(FN+TN)
		Sensitivity TP/(TP+FN)	Sensitivity TN/(FP+TN)	

Sensitivity and specificity indicate the proportion of test positives among those with the target condition (true positives) and test negatives among those without the target condition (true negatives), respectively. Predictive values indicate the proportion of disease positives among all those who test positive (positive predictive value) or the disease negatives among test negatives (negative predictive value). These metrics, among others, express the correspondence between the test and reference standard.

Not all tests generate a dichotomous result. In such cases, a positivity threshold can dichotomize a continuous test result, common for many biomarkers and diagnostic models on a continuous or ordinal scale, allowing classification of those with and without the target condition. When evaluating a biomarker as a medical test, a receiver operating characteristic (ROC) curve can be constructed (Figure 1). ROC curves illustrate the sensitivity and specificity at any possible positivity threshold. The area under the ROC

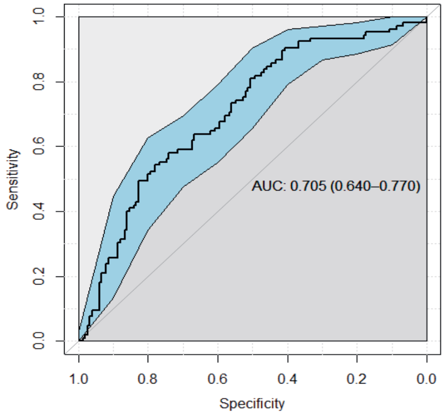


Figure 1. An example of a ROC curve including the area under the ROC curve (AUC)

curve (AUC) expresses the overall ability of a test to distinguish those with and without the target condition. ROC curves provide an intuitive visual of the trade-off between the sensitivity, meaning, proportion of true positives, and the specificity, true negatives. These characteristics are a direct result of the selected positivity threshold. The context in which a test may be most beneficial influences the selection of the positivity threshold, to the extent that the same biomarker with a different threshold may be considered a different test in itself.

For this thesis, we conducted studies and explored methods for both synthesizing the available evidence and generating new data on test performance. The earlier studies inspired the later chapters where we expanded the evaluation of bias, an essential element of the evidence synthesis process, and applied methods to alleviate a well understood challenge in diagnostic accuracy studies: accommodating for variability and its influence on the positivity threshold (1). We studied the performance of several non-invasive tests and further generated new diagnostic models using supervised machine learning techniques. These studies were conducted largely in the context of evaluating non-invasive biomarkers for the detection of key outcomes in patients with non-alcoholic fatty liver disease (NAFLD).

Non-alcoholic fatty liver disease (NAFLD)

NAFLD is a multifactorial condition characterized by the accumulation of fat in the hepatocytes. As a progressive disease, the histological spectrum spans from simple steatosis, to non-alcoholic steatohepatitis (NASH) with or without fibrosis, and a small but significant subset of patients may progress to more severe stages like cirrhosis and hepatocellular carcinoma (HCC). In parallel with metabolic conditions, such as obesity and diabetes, the fast-growing prevalence of NAFLD has set it to become a significant cause of liver cancer and transplantation (2). Despite its prevalence and growing clinical significance, there are no licensed therapies for NASH, and accurate diagnosis, and thereby timely disease management, remains a challenge.

The current clinical reference standard for detecting outcomes with NAFLD is liver histology. A patient undergoes a biopsy to evaluate the degree of hepatic steatosis, lobular inflammation, ballooning, and fibrosis. A biopsy, however, poses risk for the patients, is resource intensive, and has limitations including inter- and intra-observer variability (3). This, and the growing number of drugs under development for NASH, have been the driving force for regulatory approval of non-invasive diagnostic alternatives, predominantly for those with active NASH or NASH with a degree of fibrosis.

Outline of thesis

For this thesis, a series of systematic reviews and meta-analyses were performed to landscape the performance of selected biomarkers proposed for detecting conditions within NAFLD. In **Chapter 2** we synthesized the existing literature on circulating cytokeratin-18 (CK-18) as a candidate marker for detecting NASH. We discovered and managed the heterogeneous use of positivity thresholds by applying a linear mixed effects multiple thresholds model to accommodate for the different thresholds.

Chapter 3 focuses on an imaging modality, vibration controlled transient elastography (VCTE), for the staging of liver fibrosis. Here we performed an individual patient data (IPD) meta-analysis to evaluate the ability of VCTE to detect fibrosis stages. The performance of VCTE was compared to liver fibrosis tests commonly used in practice.

Chapter 4 presents a systematic review on the prognostic accuracy of three known and accessible non-invasive multi-marker scores for NAFLD-related events. Unlike diagnostic accuracy studies, which evaluate the ability of a test to detect a target condition, either present or absent at the time of testing, prognostic accuracy studies evaluate the performance of a test to predict the occurrence of a future event.

The challenges brought forth by the systematic review reported in Chapter 4 inspired the development of a modified risk-of-bias tool, which we delve into in **Chapter 5**. In this chapter, we introduce QUAPAS (Quality Assessment of Prognostic Accuracy Studies). QUAPAS is a risk-of-bias and applicability assessment tool for prognostic accuracy studies. No such tool was previously available, presenting a challenge when conducting reviews of prognostic accuracy studies. We systematically modified an existing instrument, QUADAS-2 for diagnostic accuracy studies, in the absence of an appropriate risk-of-bias tool for the systematic review in Chapter 4.

Following the evidence synthesis phase, we evaluated seventeen biomarkers, multi-marker scores and VCTE for detecting NAFLD conditions in a comparative diagnostic accuracy study, conducted in a large multicenter study group (**Chapter 6**). Here we additionally proposed new positivity thresholds for recruitment of patients in future drug trials, with the aim of reducing those subjected to biopsies. This work inspired **Chapter 7**, where new prediction models were developed, with clinical and biomarker data, to stage and grade NASH, at-risk NASH, and fibrosis stages using supervised machine learning techniques.

Finally, in **Chapter 8**, we evaluate methods proposed for incorporating covariate information in ROC curve analysis, in the setting of D-dimer as a diagnostic test for venous thromboembolism. We further explored the implications of covariates, and their subgroups, on the appropriate selection of the positivity threshold.

This thesis closes with a direction for future research for conducting systematic reviews and primary analysis of test accuracy studies (**Chapter 9**).

02

Accuracy of cytokeratin 18 (M30 and M65) in detecting non-alcoholic steatohepatitis and fibrosis: a systematic review and meta-analysis

Jenny Lee
Yasaman Vali
Jérôme Boursier
Kevin Duffin
Joanne Verheij
Julia Brosnan
Koos Zwinderman
Quentin M. Anstee
Patrick M. Bossuyt
Mohammad Hadi Zafarmand

PLoS One, 2020, 15.9: e0238717.

Abstract

Introduction: Association between elevated cytokeratin 18 (CK-18) levels and hepatocyte death has made circulating CK-18 a candidate biomarker to differentiate non-alcoholic fatty liver from non-alcoholic steatohepatitis (NASH). Yet studies produced variable diagnostic performance. We aimed to provide summary estimates with increased precision for the accuracy of CK-18 (M30, M65) in detecting NASH and fibrosis among non-alcoholic fatty liver disease (NAFLD) adults.

Methods: We searched five databases to retrieve studies evaluating CK-18 against a liver biopsy in NAFLD adults. Reference screening, data extraction and quality assessment (QUADAS-2) were independently conducted by two authors. Meta-analyses were performed for five groups based on the CK-18 antigens and target conditions, using one of two methods: linear mixed-effects multiple thresholds model or bivariate logit-normal random-effects model.

Results: We included 41 studies, with data on 5,815 participants. A wide range of disease prevalence was observed. No study reported a pre-defined cut-off. Thirty of 41 studies provided sufficient data for inclusion in any of the meta-analyses. Summary AUC [95% CI] were: 0.75 [0.69 - 0.82] (M30) and 0.82 [0.69-0.91] (M65) for NASH; 0.73 [0.57-0.85] (M30) for fibrotic NASH; 0.68 (M30) for significant (F2-4) fibrosis; and 0.75 (M30) for advanced (F3-4) fibrosis. Thirteen studies used CK-18 as a component of a multimarker model.

Conclusions: For M30 we found lower diagnostic accuracy to detect NASH compared to previous meta-analyses, indicating a limited ability to act as a stand-alone test, with better performance for M65. Additional external validation studies are needed to obtain credible estimates of the diagnostic accuracy of multimarker models.

Introduction

Non-alcoholic fatty liver disease (NAFLD), a condition with a complex and multifactorial etiology, has rapidly emerged as the most common cause of chronic liver disease in the United States and Europe (1, 2). The global prevalence is approximately 25%, representing a wide histological spectrum from simple steatosis (NAFL), non-alcoholic steatohepatitis (NASH) (3) to hepatic fibrosis. Fibrosis is the strongest predictor for long-term clinical outcomes in NAFLD patients, thereby, a key target event for patient stratification and clinical trial recruitment (4).

The clinical reference standard for detecting NASH activity and fibrosis stages is a liver biopsy, a practice with well-established limitations (5-7). As such, only patients at highest risk should be pre-selected for such an invasive and resource intensive procedure. The discovery of less invasive methods with performance comparable to liver biopsy has become essential.

Several blood-based biomarkers have been studied for their ability to identify NASH or fibrosis. Cytokeratin 18 (CK-18) is the main intermediate filament protein in hepatocytes and is released upon the initiation of cell death. The association between elevated CK-18 levels and cell death in the liver (8, 9) has made circulating CK-18 (both M30 and M65 antigens) a candidate marker for detecting NASH and fibrosis (10), as a stand-alone test and, more recently, as part of multimarker models.

Although the M30 and M65 antigens are of the same protein, there is a mechanistic distinction between the two. M30 measures the caspase-cleaved CK-18 revealed during apoptosis, while M65 measures the full-length protein, including both caspase-cleaved and intact CK-18, which is released from cells undergoing necrosis (11).

In recommendations by the EASL-EASD-EASO Clinical Practice Guidelines (12) the performance of CK-18 M30 to differentiate NASH from NAFL was judged modest, as per data from a meta-analysis of 11 studies (13). The Asia-Pacific Working Party on NAFLD (14) similarly concluded modest performance, referencing a meta-analysis of 10 studies (15). A single study mentioned in both guidelines criticized CK-18 for its limited performance for detecting NASH at a threshold of 165 U/L (10). However, it is not clear what thresholds would then maximize the test's sensitivity or specificity.

We found several limitations and methodological concerns in the above-mentioned meta-analyses. One performed a meta-analysis on only the M30 antigen in detecting NASH, with the rationale that M65 performed similarly (13). However, it has been shown that M65 outperforms M30 (9). Further, we found several methodological concerns in the systematic review by Chen et al. (2014) such as overlapping patient populations included in the meta-analysis (15).

An updated and more methodologically robust meta-analysis would be able to generate, in principle, summary estimates with increased precision and more general validity. To address this need, we aimed to conduct a systematic review and meta-analysis of the accuracy of both CK-18 antigens (M30 and M65) in identifying NASH, fibrotic NASH, and fibrosis stages among NAFLD adults.

Materials and Methods

This systematic review was conducted as part of the evidence synthesis efforts of the LITMUS (Liver Investigation: Testing Marker Utility in Steatohepatitis) project, funded the European Union's IMI2, aiming to evaluate biomarkers for use in NAFLD. The protocol of the complete systematic review is available in PROSPERO (registration number: CRD42018106821). This study report was prepared using the PRISMA-DTA statement, see PRISMA checklist in S1 Table.

Search strategy

A comprehensive search strategy, containing words in the title/abstract or text words across the record and the medical subject heading (MeSH), was developed with a search specialist. MEDLINE (via OVID), EMBASE (via OVID), PubMed, Science Citation Index, and CENTRAL (The Cochrane Library) were searched to retrieve potentially eligible studies from inception to August 2018 (see S2 Table). We further conducted a manual screening of relevant systematic reviews and reference lists and contacted partners within the LITMUS consortium. The search was updated in May 2019, and again in June 2020.

Study selection

Search results of all databases were merged and deduplicated using Endnote. Titles were screened by one reviewer (YV); a second reviewer independently screened 10% (MHZ). Abstract and full text screening was conducted by two independent reviewers (JL and YV), following pre-established inclusion and exclusion criteria. Any discrepancies were resolved by discussion between the two reviewers. Title and abstract screening phases were conducted on Rayyan QCRI (https://rayyan.qcri.org).

Inclusion and exclusion criteria

We searched for studies including adults (≥18 years) with clinical suspicion or biopsy proven NAFLD, with paired data on liver histology and CK-18 (M30 or M65). Diagnostic accuracy studies reported in full articles in peer-reviewed journals, or as conference abstracts, in any language were eligible. Studies with insufficient information for making decisions on inclusion, for evaluating methodological quality, or for calculating diagnostic accuracy were excluded. Study groups with a mix of conditions (e.g. viral hepatitis) were only included if outcomes were separately reported for NAFLD patients.

The target conditions for this systematic review were NASH, fibrotic NASH, and liver fibrosis. The NAFLD Activity Score (NAS) (16) is the most commonly used pathologic criterion for evaluating NASH. We considered a threshold value of NAS ≥4 with at least one point for each criteria of steatohepatitis for the characterization of NASH. See S3 Table for different histological scoring systems developed to characterize NAFLD progression. Fibrotic NASH was defined using the above-mentioned criteria for NASH and at least F1 or more.

A five-point scoring system (F0-F4), developed by the NASH clinical research network (NASH CRN) (17), is the most commonly used for fibrosis staging. Studies assessing significant (≥F2) and advanced (≥F3) fibrosis were included. See S4 Table for different scoring systems for liver fibrosis, and S5 Table for a conversion grid of the different scoring systems.

Data extraction and quality assessment

The following information was extracted: study characteristics, clinical characteristics, index test features, liver biopsy features, and data that allowed construction of a 2x2 contingency table (true positives, true negatives, false positive and false negatives) to assess the performance of the index test. For studies that reported accuracy data for multiple thresholds, all data were extracted.

When pertinent data were not reported, the corresponding study author was contacted. Data were extracted independently and cross-checked by two reviewers (JL and YV).

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool (18) was used to assess methodological quality of all available full text studies. Two reviewers (JL and YV) independently evaluated the risk of bias and concerns about applicability of the included primary studies using the four domains of QUADAS-2, assigning each study with a judgement of ‘low’, ‘high’, or ‘unclear’ risk.

Statistical analysis

Included studies were classified into five groups for meta-analysis, based on the availability of data on the CK-18 antigens and target conditions: (1) CK-18 M30 for detecting NASH, (2) CK-18 M65 for detecting NASH, (3) CK-18 M30 for detecting fibrotic NASH, (4) CK-18 M30 for detecting significant fibrosis, and (5) CK-18 M30 for detecting advanced fibrosis.

Sensitivity and specificity estimates from each study, with respective 95% confidence intervals (95% CI), were graphically illustrated as forest plots, for each reported threshold, using RevMan.

Two different meta-analytical methods were applied for the combinations of CK-18 antigens and target conditions based on the number of reported threshold values. For groups 1-3, we applied a linear mixed effects multiple thresholds model (diagmeta package in R) as a majority of the primary studies reported multiple threshold values. The multiple thresholds model utilizes the number of true and false positives and true and false negatives at every threshold to produce summary receiver operating characteristic (SROC) curves. With the model, we could calculate estimates of sensitivity, specificity at any given threshold. We calculated the threshold value that would maximize Youden's J statistic (also called Youden's index): the sum of sensitivity and specificity minus 1.

We computed estimates of positive and negative predictive values in settings with different disease prevalence. We further assessed thresholds of the index test required to achieve pre-specified high values of sensitivity and specificity. The minimally acceptable performance levels of AUC and sensitivity and specificity for the index test was 0.80, for it to exceed that of other NAFLD-related screening and diagnostic biomarkers.

As a majority of the primary studies in groups 4 and 5 reported only a single threshold value, we applied a bivariate logit-normal random-effects model (mada package in R) to compute summary estimates of sensitivity and specificity. SROC curves were constructed to represent the overall diagnostic accuracy of the index test.

Publication bias was not formally evaluated as no accepted statistical tests can reliably discriminate publication bias from other sources of bias in diagnostic meta-analyses (19). Heterogeneity between and within studies was incorporated by calculating 95% prediction intervals (20). The confidence interval around the summery point reflect the statistical imprecision around the mean. The prediction region around the summary point indicates the region where we would expect results from a new study in the future to lie. It reflects both the uncertainty around the mean and the between study heterogeneity and is therefore wider than the confidence region.

We investigated the influence of studies with compromised methodological quality by excluding those at high risk of bias or with applicability concerns in a sensitivity analysis. We further evaluated the effect of pooling data from various ELISA assays by excluding studies that either did not disclose the assay used or used one from a manufacturer that was not PEVIVA. Sensitivity analysis was also conducted among solely biopsy-proven NAFLD patients, excluding those with clinically suspected NAFLD.

All analyses were conducted using R for Windows (Version 3.6.0; R Foundation for Statistical Computing, Vienna, Austria).

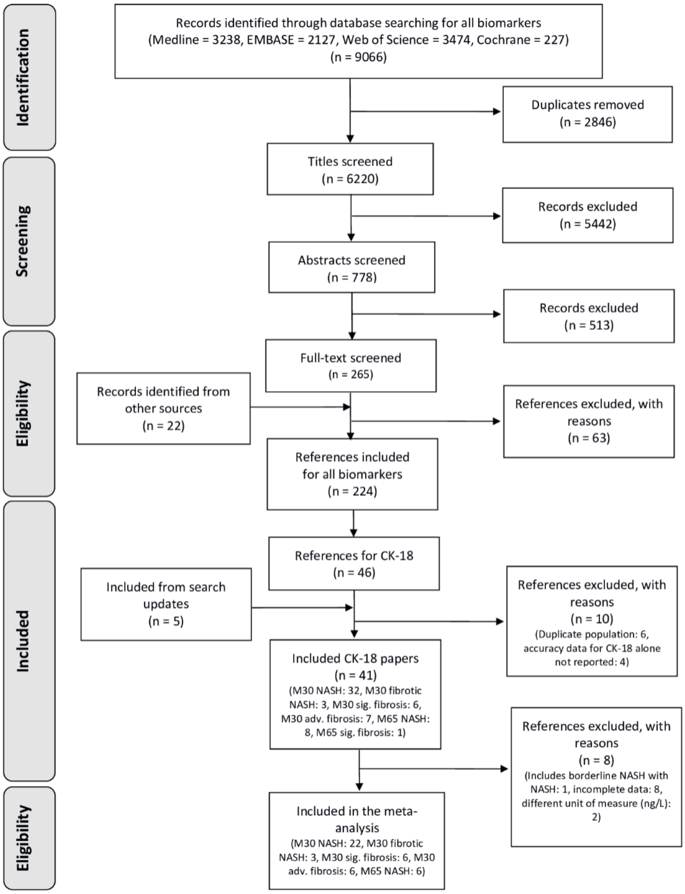

Identification
Records identified through database searching for all biomarkers (Medline = 3238, EMBASE = 2127, Web of Science = 3474, Cochrane = 227) (n = 9066)
Duplicates removed (n = 2846)
Screening
Titles screened (n = 6220)
Records excluded (n = 5442)
Abstracts screened (n = 778)
Records excluded (n = 513)
Eligibility
Full-text screened (n = 265)
Records identified from other sources (n = 22)
References excluded, with reasons (n = 63)
References included for all biomarkers (n = 224)
Included
References for CK-18 (n = 46)
Included from search updates (n = 5)
References excluded, with reasons (n = 10) (Duplicate population: 6, accuracy data for CK-18 alone not reported: 4)
Included CK-18 papers (n = 41) (M30 NASH: 32, M30 fibrotic NASH: 3, M30 sig. fibrosis: 6, M30 adv. fibrosis: 7, M65 NASH: 8, M65 sig. fibrosis: 1)
References excluded, with reasons (n = 8) (Includes borderline NASH with NASH: 1, incomplete data: 8, different unit of measure (ng/L): 2)
Eligibility
Included in the meta-analysis (M30 NASH: 22, M30 fibrotic NASH: 3, M30 sig. fibrosis: 6, M30 adv. fibrosis: 6, M65 NASH: 6)

Figure 1. PRISMA flow diagram of included primary studies

Results

Search results

Our initial search of all biomarkers identified 6,220 studies post deduplication. Following the pre-defined inclusion and exclusion criteria, 778 studies were eligible for abstract screening, of which 265 underwent full-text review. A total of 46 study reports were included for CK-18. Following the exclusion of 10 and inclusion of five studies from the two search updates, a total of 41 studies (5,815 participants) could be included in the present systematic review (Figure 1). Thirty studies were included in one or more of the meta-analyses.

Study characteristics

Characteristics of the included studies can be found in Table 1. A majority of the studies (32/41) had included NAFLD patients with mean BMI <35. A relatively wide range of disease prevalence was observed; 21% to 85% for NASH, 21% to 62% for fibrotic NASH, 18% to 59% for significant fibrosis and 19% to 36% for advanced fibrosis. The publication year spanned from 2006 to 2020; 27 studies were published after 2012.

Thirty-two studies investigated the accuracy of M30 in detecting NASH, and three for fibrotic NASH. The accuracy of M30 in detecting significant and advanced fibrosis was studied in six and seven studies, respectively. We further identified eight diagnostic accuracy studies of M65 for NASH and one study of M65 for significant fibrosis.

Quality assessment

The methodological quality of the 41 studies, assessed with QUADAS-2, is summarized in S1 and S2 Figure. Ten studies were scored as high risk of bias in the patient selection domain (9, 10, 29, 31, 36, 37, 40, 47, 50, 58). No study had low risk of bias in the index test domain, with 22 judged as high risk, due to the lack of a pre-established threshold value for CK-18.

Table 1. Characteristics of all included studies

	Study ID	Country	*N* (female)	Target condition	Prev (%)	Population	BMI, mean (SD)	ALT, median (IQR)	AST, median (IQR)	Comorbidities (%)
1.	Aida 2014 (21)	Japan	116 (75)	NASH	44	Biopsy proven NAFLD	27.2 (18.8-45.9)†	52 (31-266)	42 (13-256)	NR
2.	Cao 2013 (22)	China	95 (73)	NASH	46	Biopsy proven or clinical suspicion of NAFLD	28.5 (2.8)	57.0 (48.0-71.0)	48.0 (44.0-54.0)	DM: 24 HTN: 48 DL: 86
3.	Chan 2014 (23)	Malaysia	93 (48)	NASH	42	US diagnosed NAFLD	29.4 (3.8)	70 (44-109)	41 (28-64)	DM: 59 HTN: 88 DL: 97
4.	Chuah 2019 (24)	Malaysia	196 (99)	Fibrotic NASH	21	US diagnosed NAFLD	29.8 (4.5)	67 (44-105)	39 (29-61)	T2DM: 46 HTN: 58 DL: 80 Obesity: 86
5.	Cusi 2013 (10)	USA	318 (113)	NASH	63	Obese patients with biopsy proven NAFLD	33.3 (0.9)	40 (1)‡	55 (2)‡	NR
6.	Darweesh 2019 (25)	Egypt	25 (55.6)	Steatosis	NR	Biopsy proven NAFLD	33.52 (4.56)	50.57 (31.06)	48.29 (46.51)	NR
7.	Dvorak 2014 (26)	Czech Republic	56 (NR)	NASH	68	Biopsy proven NAFLD	29.6 (4.3)	120 (90)‡	66 (60)‡	NR
8.	Ergelen 2015 (27)	Turkey	87 (44)	Sig. fibrosis Adv. fibrosis	39 22	Biopsy proven NAFLD	30.6 (5.4)	77.8 (56.1)‡	49.6 (30.5)‡	NR
9.	Feldstein 2009 (8)	USA	139 (88)	NASH	50	Biopsy proven NAFLD	34.2 (30.3-37.8)†	43.0 (31.0-62.0)	66.0 (46.0-109.0)	DM: 19 HTN: 43 HL: 60
10.	Grigorescu 2012 (28)	NR	79 (23)	NASH	75	Biopsy proven NAFLD	30 (3.8)	76.8 (39.3)‡	35.9	T2DM: 16 HTN: 19
11.	Hasegawa 2015 (29)	Japan	41 (7)	NASH	49	US and CT diagnosed NAFLD	NR	75.3 (68.4)‡	53.6 (46.8)‡	NR

12. Huang 2017 (30)	Taiwan	76 (22)	Sig. fibrosis Adv. fibrosis	18 9	Biopsy proven NAFLD	28.7 (4.4)	117 (87.9)‡	63.1 (33.3)‡	DM: 54 HTN: 65
13. Joka 2011 (9)	Germany	22 (7)	NASH	55	Biopsy proven NAFLD	27 (1)	75.5 (9.5)‡	NR	NR
14. Kamada 2013 (31)	Japan	126 (56)	NASH	85	Biopsy proven NAFLD	27.5 (5.1)	95.8 (72.0)	62.9 (39.3)	NR
15. Kawanka 2015 (32)	Japan	146 (78)	NASH	71	Biopsy proven NAFLD	26.8	61 (12-264)	38 (14-204)	NR
16. Kazankov 2016 (33)	Australia, Italy	331 (112)	NASH	40	Biopsy proven NAFLD	29.2 (4.5)	67.6	40.7	DM: 20
17. Kim 2013 (34)	Korea	108 (35)	NASH	62	Biopsy proven NAFLD	28.71 (3.77)	108.68 (82.07)‡	63.54 (41.62)‡†	MetS: 48
18. Kobayashi 2017 (35)	Japan	229 (107)	NASH Fibrotic NASH	61 45	Biopsy proven NAFLD	26.6	79.2	50.7	DM: 45 HTN: 42 DL: 56
19. Liu 2016 (36)	China	48 (13)	NASH	65	Biopsy proven NAFLD	26.9 (0.5)	68.7 (7.4)‡	NR	NR
20. Liu 2019 (37)	China	82 (23.5)	NASH	47	Biopsy proven NAFLD	26.8 (3.3)	80.5 (76.4)	47.9 (31.8)	DM: 32 HTN: 35
21. Malik 2009 (38)	USA	95 (37)	NASH	63	Biopsy proven NAFLD	31.3 (4.2)	74.5 (9.7)‡	NR	T2DM: 27 HTN: 49
22. Mohammed 2019 (39)	Egypt	62 (62)	NASH	66	US proven NAFLD	30.8 (4.02)	75.53 (22.3)	69 (29.5)	MetS: 59
23. Musso 2010 (40)	NR	40 (12)	NASH	58	Biopsy proven NAFLD	25.1 (1.6)	120.7 (8)‡	48 (3)‡	MetS: 43
24. Papatheodoridis 2010 (41)	Greece	58 (26)	NASH	52	Biopsy proven NAFLD	28.6 (4.5)	75.4	39.5	DM: 16
25. Pimentel 2016 (42)	USA	183 (73)	NASH Adv. fibrosis	49 19	Biopsy proven NAFLD	34 (7)	50.6 (32)‡	75.8 (50)‡	T2DM: 36 HTN: 52
26. Rosso 2016 (43)	Italy	105 (29)	Sig. fibrosis Adv. fibrosis	59 36	Biopsy proven NAFLD	28.1 (3.9)	65 (57-79)	36 (33-41)	NR

27. Shen 2012 (44)	China	147 (65)	NASH	47	Biopsy proven NAFLD	27.4 (3.9)	73 (45) ‡	NR	T2DM: 48 HTN: 43
28. Tada 2018 (45)	Japan	170 (91)	NASH	76	Biopsy proven NAFLD	27.6 (24.9-30.7) †	79 (49-126)	52 (35-82)	DM: 51 HTN: 28 DL: 44
29. Tamimi 2011 (46)	USA	95 (47)	NASH	43	Clinically suspected NASH	31.4 (5.1)	53.5 (32-87)	54 (38-75)	DM: 27 HTN: 45 MetS: 50 HL: 53
30. Valva 2018 (47)	Argentina	34 (15)	Sig. fibrosis	18	Biopsy proven NAFLD	NR	81.5 (31-279)	52.5 (22-208)	Obesity: 25
31. Wieckowska 2006 (48)	USA	39 (21)	NASH	31	Biopsy proven NAFLD	31.5 (4.0)	73.0 (54.0-104.0)	58.0 (46.0-76.0)	DM: 31 HTN: 46 HL: 46
32. Yang 2015 (49)	China	179 (93)	NASH	38	Biopsy proven NAFLD	NR	116 (30.2) ‡	60 (22.1) ‡	NR
33. Yilmaz 2007 (50)	Turkey	83 (38)	Sig. fibrosis	20	Suspected NAFLD	30.3 (4.8)	60 (10-184)	42 (16-102)	DM: 15 HTN: 34 MetS: 35
34. Younes 2018 (51)	Italy	292 (91)	NASH Adv. fibrosis	77 25	Biopsy proven NAFLD	28.9 (4.1)	66 (61-71)	36 (35-38)	MetS: 32 DM: 20
35. Younossi 2008 (52)	USA	69 (46)	NASH	32	Biopsy proven NAFLD	NR	27.1 (18.4) ‡	36.6 (27.3) ‡	NR
36. Zheng 2020 (53)	China	38 (36.2)	NASH	53	Biopsy proven NAFLD (ALT ≤ 35 (men), ≤ 23 (women))	26.05 (3.33)	27.70 (7.77)	25.77 (6.75)	DM: 36 MetS: 55 HTN: 35
37. Anty 2010 (54)	France	310 (267)	NASH	NR	Morbidly obese, bariatric surgery patients	44.7 (5.5)	35.3 (35.7) ‡	NR	NAS<5 DM: 19.6 MetS: 47.6 NAS≥5 DM: 43.6 MetS: 82.1

38. Boursier 2018 (55)	France, Belgium	846 (525)	NASH Fibrotic NASH Sig. fibrosis Adv. fibrosis	54 23 51 17	Biopsy proven NAFLD, obese patients, morbidly obese patients referred for bariatric surgery	38.5 (7.6)	49.7 (31.7) ‡	35.5 (19.6) ‡	MetS: 68 DM: 27
39. Diab 2008 (56)	USA	55 (68)	NASH	40	Bariatric surgery patients	48 (43-54) †	23.0 (18.0-29.0)	21.5 (16.0-33.0)	DM: 41 HTN: 67 DL: 57
40. Pirvulescu 2012 (57)	Romania	59 (42)	NASH (incl. borderline NASH)	22	Overweight, obese and morbidly obese patients referred for bariatric surgery	47.3 (8.1)	37.8 (13.6) ‡	29.3 (10.1) ‡	NR
41. Younossi 2011 (58)	USA	79 (61)	NASH	51	Biopsy proven NAFLD	47.56 (8.07)	36.44 (28.05) ‡	27.22 (19.39) ‡	DM: 24

† Median and interquartile range

‡ Mean and standard deviation

NR: not reported, DM: diabetes mellitus, T2DM: type 2 diabetes mellitus, HTN: hypertension, DL: dyslipidemia, MetS: metabolic syndrome, US: ultrasound, CT: computerized tomography scan

Seven studies were scored as unclear risk of bias in the reference standard domain, for failing to report whether biopsy reviewers were blinded to clinical data (8, 9, 29, 32, 40, 49, 57). Only three studies were classified as at high risk of bias for flow and timing (10, 25, 40). We further graded four studies with high concern regarding applicability in the patient selection domain (30, 46, 54, 57).

NASH

Accuracy of CK-18 M30 in detecting NASH

A total of 22 studies (3,503 participants, 2,010 with NASH) were included in the meta-analysis of the diagnostic accuracy of M30 in detecting NASH (S3 Figure). Ten studies reported multiple threshold values, resulting in 47 thresholds (41 unique values) included in our model. The thresholds spanned from 111 to 670 U/L. The multiple thresholds model produced a summary area under the receiver operating characteristic curve (AUC) of 0.75 (95% CI: 0.69 - 0.82) with a mean sensitivity of 0.61 (95% CI: 0.51 - 0.71) and mean specificity of 0.81 (95% CI: 0.71 - 0.88). The Youden-threshold value was 304 U/L (Figure 2A).

Using the multiple thresholds model, we calculated the positive predictive value (PPV) and the negative predictive value (NPV) under different clinical settings (5% to 70% NASH prevalence) for desired levels of sensitivity and specificity (Table 2). Optimizing sensitivity (0.80 to 0.90), we found corresponding specificity values, ranging from 0.51 to 0.23 at threshold values 127 to 191 U/L (Table 2). High NPV (0.91 - 0.96) values were observed at lower prevalence setting of 10% and 20%. The corresponding PPV ranged from 0.12 to 0.29.

When fixing specificity values (0.80 to 0.90), the corresponding sensitivity ranged from 0.48 to 0.61 (Table 2) with threshold values between 304 and 399 U/L. High NPV (0.87 to 0.95) were again seen for low prevalence settings (10 to 20%). A graphical representation of the predictive values in different prevalence settings can be seen in Figure 3A-B.

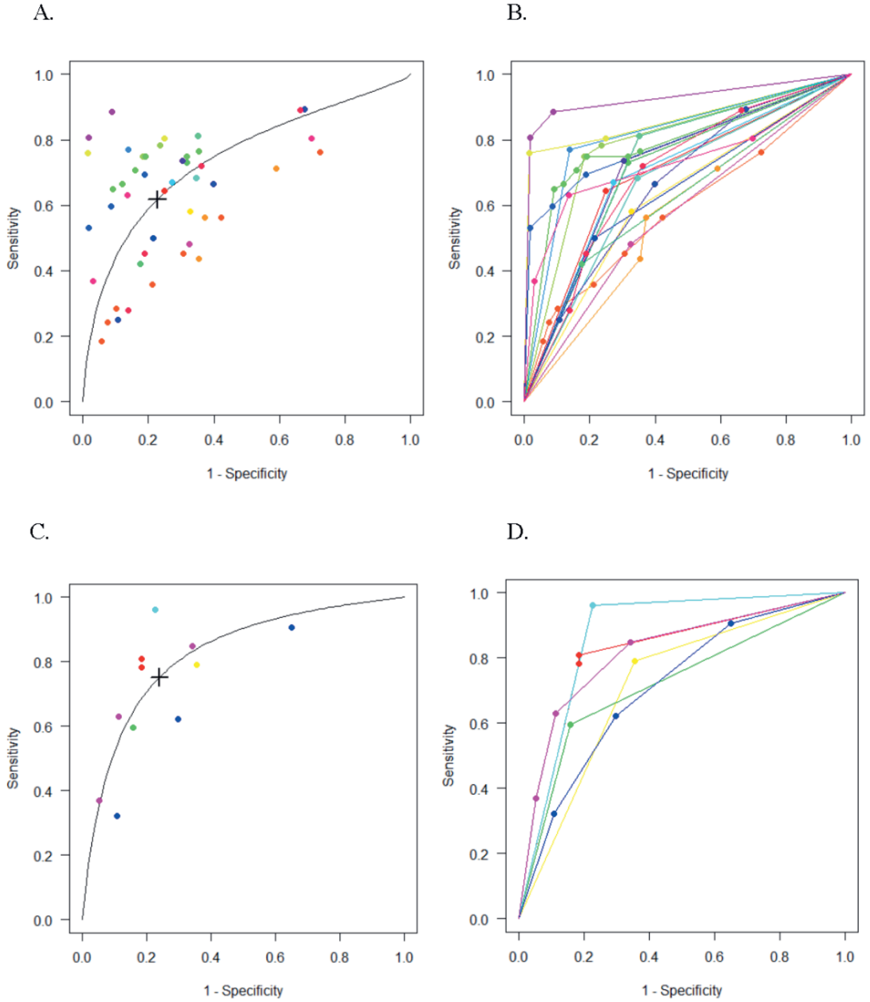


Figure 2. Multiple threshold SROC and ROC curves for detecting NASH. Multiple threshold SROC and ROC curves for CK-18 M30 (A-B) and M65 (C-D) in detecting NASH. Each point represents a reported threshold value, points of the same color represent thresholds reported within the same study. The x-axis indicates 1 – specificity, and the y-axis, sensitivity. The cross in the SROC curve indicates the Youden-based threshold value: A. Youden-threshold: 304 U/L, sensitivity: 0.61 (95% CI: 0.51 - 0.71), specificity: 0.81 (95% CI: 0.71 - 0.88), AUC: 0.75 (95% CI: 0.69 - 0.82) for CK-18 M30. C. Youden-threshold: 478 U/L, sensitivity: 0.75 (95% CI: 0.51 - 0.90), specificity: 0.76 (95% CI: 0.49 -0.91), AUC: 0.82 (95% CI: 0.69 - 0.91) for CK-18 M65.

Accuracy of CK-18 M65 in detecting NASH

In the meta-analysis of M65 in detecting NASH, we analyzed six studies with a total of 414 participants (220 with NASH) (S4 Figure). Eleven unique threshold values were included in the model, ranging from 340 to 1183 U/L. The combined AUC was 0.82 (95% CI: 0.69 - 0.91) with a mean sensitivity of 0.75 (95% CI: 0.51 - 0.90) and mean specificity of 0.76 (95% CI: 0.49 -0.91) at Youden-threshold of 478 U/L (Figure 2C).

We again investigated the PPV and NPV in various clinical settings (Table 3, Figure 3C-D). Fixing sensitivity from 0.80 to 0.90, the specificity ranged from 0.70 to 0.51 at threshold values of 337 to 437 U/L (Table 3). NPV in lower prevalence settings (10-20%) ranged from 0.93 to 0.98 with corresponding PPV from 0.17 to 0.40. Similar patterns were observed for optimizing specificity over sensitivity (Table 3). Within a NASH prevalence of 10% or 20% we found PPV and NPV ranged from 0.28 to 0.56 and from 0.88 to 0.96, respectively.

Accuracy of CK-18 M30 in detecting fibrotic NASH

Three studies provided sufficient data for analysis of M30 in detecting fibrotic NASH, with a combined total of 1,271 participants (343 with fibrotic NASH) (S5 Figure). Two studies investigated M30 as part of a multimarker models; authors of both studies (24, 55) provided accuracy data for M30 at seven threshold values we selected based on the data from the present meta-analysis (133, 200, 248, 292, 356, 395, and 464 U/L). This allowed us to apply the multiple thresholds model (15 thresholds), to calculate an AUC of 0.73 (95% CI: 0.57 - 0.85), mean sensitivity of 0.63 (95% CI: 0.39 - 0.82) and mean specificity of 0.73 (95% CI: 0.51 - 0.88) at a Youden-threshold value of 371 U/L.

Fibrosis

Accuracy of CK-18 M30 in detecting significant and advanced fibrosis

We identified several studies that investigated CK-18 for fibrosis staging. For significant fibrosis, we included a single threshold value (ranging from 122 to 285 U/L) from five

Table 2. Performance of M30 in detecting NASH: positive and negative predictive values for different NASH prevalence

A. Fixed sensitivity values (0.80, 0.85, 0.90)

Prev	Fixed 0.80 sensitivity					Fixed 0.85 sensitivity					Fixed 0.90 sensitivity				
	Cut-off	Sp	PPV	NPV	Mis%	Cut-off	Sp	PPV	NPV	Mis%	Cut-off	Sp	PPV	NPV	Mis%
0.05	191	0.51	0.08	0.98	48	161	0.38	0.07	0.98	59	127	0.23	0.06	0.98	72
0.10			0.15	0.96	46			0.13	0.96	56			0.12	0.96	69
0.20			0.29	0.91	43			0.26	0.91	52			0.23	0.91	62
0.30			0.41	0.85	40			0.37	0.86	47			0.34	0.86	56
0.40			0.52	0.79	37			0.48	0.79	43			0.45	0.79	49
0.50			0.62	0.72	35			0.58	0.72	38			0.55	0.72	43
0.70			0.79	0.52	29			0.76	0.52	29			0.74	0.52	30

B. Fixed specificity values (0.80, 0.85, 0.90)

Prev	Fixed 0.80 specificity					Fixed 0.85 specificity					Fixed 0.90 specificity				
	Cut-off	Se	PPV	NPV	Mis%	Cut-off	Se	PPV	NPV	Mis%	Cut-off	Se	PPV	NPV	Mis%
0.05	304	0.61	0.14	0.98	21	340	0.56	0.17	0.97	17	399	0.48	0.21	0.97	12
0.10			0.25	0.95	22			0.28	0.94	18			0.33	0.94	15
0.20			0.42	0.89	24			0.47	0.88	21			0.52	0.87	19
0.30			0.56	0.82	26			0.60	0.81	25			0.65	0.79	24
0.40			0.66	0.75	28			0.70	0.73	28			0.75	0.71	28
0.50			0.75	0.66	31			0.78	0.64	31			0.82	0.62	33
0.70			0.87	0.46	35			0.89	0.43	37			0.91	0.41	42

Prev: prevalence, Sp: specificity, Se: sensitivity, PPV: positive predictive value, NPV: negative predictive value, Mis%: percent misclassified

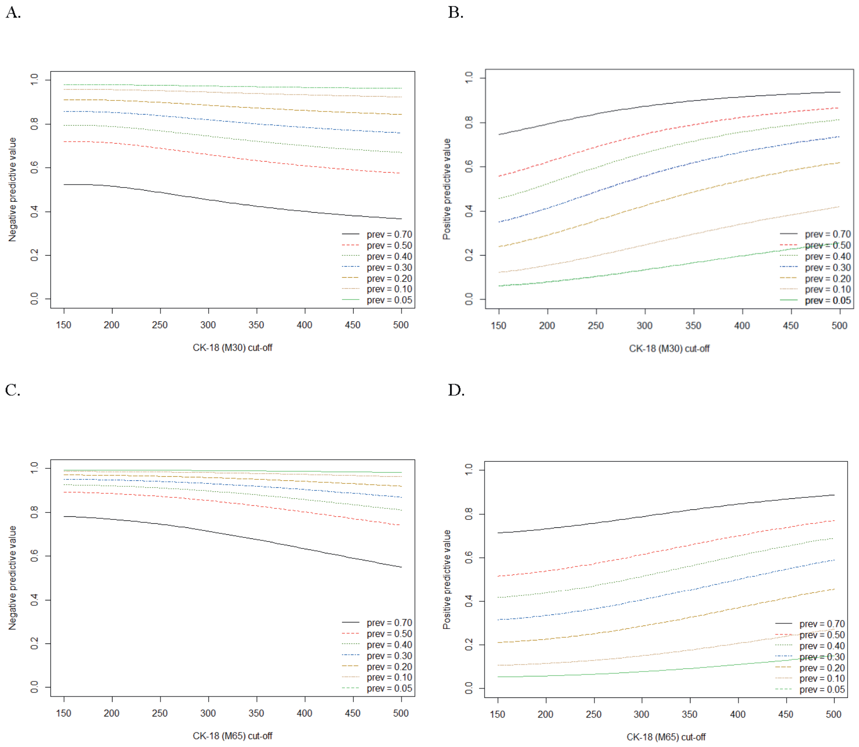


Figure 3. Positive and negative predictive values and thresholds for detecting NASH. Plots illustrating the negative and positive predictive values of M30 (A-B) and M65 (C-D) in detecting NASH at corresponding threshold values, projected by the multiple thresholds model. Each colored line represents a different prevalence setting, ranging from 5% to 70%. The y-axis indicates the predictive value and the x-axis indicated the threshold values for CK-18.

studies (27, 43, 47, 55, 59) with a total of 1,155 participants (554 had significant fibrosis) (S6 Figure). The resulting AUC was 0.68. See S7A Figure for SROC curve and corresponding 95% CI and prediction region. One study (50) assessed the ability of M65 to detect significant fibrosis; at a threshold of 244 U/L, sensitivity was 0.71 for a specificity of 0.71 (AUC: 0.74).

For advanced fibrosis, five studies (27, 42, 43, 51, 55) were included in the meta-analysis (1,513 participants, 313 with advanced fibrosis) (S8 Figure). We calculated an AUC of 0.75, with included threshold values ranging from 216 to 396 U/L (see S7B Figure). One study had to be excluded from the meta-analysis of both significant and advanced fibrosis due to discrepancies in the 2x2 contingency table (30).

Multimarker models including CK-18

Thirteen studies additionally used CK-18 as an ingredient of a multimarker model (Table 4). There was greatest interest in detecting NASH (8/13 studies), with AUCs among the eight models ranging from 0.79 to 0.96. The highest performance was observed in NASH-score (BMI, alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), HOMA-IR, M65 and adiponectin), which produced an AUC of 0.96 (57).

One model was developed with the aim of detecting fibrotic NASH (55). Composed of three ingredients (HOMA, AST and CK-18) the AUC from the validation group (n = 846) was 0.85. MACK-3 had an AUC of 0.80 when evaluated in a separate study (24).

Two studies (27, 43) investigated the combined use of M30 with transient elastography (TE) (FibroScan) to detect fibrosis. One study found combining TE and M30 to detect significant (AUC: 0.89) and advanced fibrosis (AUC: 0.93) did not significantly improve the diagnostic ability from either TE or CK-18 as a stand-alone test (27). Another study, however, found some improvement in AUC by combining M30 to TE compared to TE alone; in adding M30 they found an improvement in AUC by 0.03 for significant fibrosis, and 0.05 for advanced fibrosis (43).

Sensitivity analysis

A sensitivity analysis was conducted excluding four studies with two or more domains of high risk of bias or applicability concerns (9, 10, 31, 40) for M30 and NASH. The AUC was 0.75 (95% CI: 0.68 – 0.81), with a mean sensitivity of 0.62 (95% CI: 0.51 - 0.72), and mean specificity of 0.78 (95% CI: 0.66 - 0.86).

We identified four studies that used an ELISA assay that was not from PEVIVA (40, 42, 52, 53). Among the 18 studies that used the M30 Apoptosense ELISA by PEVIVA, the AUC was 0.74 (95% CI: 0.67; 0.80), with paired sensitivity and specificity of 0.60 (95% CI: 0.49; 0.70)

Table 3. Performance of M65 in detecting NASH: positive and negative predictive values for different NASH prevalence

A. Fixed sensitivity values (0.80, 0.85, 0.90)

Prev	Fixed 0.80 sensitivity					Fixed 0.85 sensitivity					Fixed 0.90 sensitivity				
	Cut-off	Sp	PPV	NPV	Mis%	Cut-off	Sp	PPV	NPV	Mis%	Cut-off	Sp	PPV	NPV	Mis%
0.05	437	0.70	0.13	0.99	30	391	0.63	0.11	0.99	36	337	0.51	0.09	0.99	47
0.10			0.23	0.97	29			0.20	0.97	35			0.17	0.98	45
0.20			0.40	0.93	28			0.36	0.94	33			0.32	0.95	41
0.30			0.54	0.89	27			0.49	0.91	30			0.44	0.95	37
0.40			0.64	0.84	26			0.60	0.86	28			0.55	0.88	33
0.50			0.73	0.78	25			0.69	0.81	26			0.65	0.84	30
0.70			0.86	0.60	23			0.84	0.64	22			0.81	0.67	22

B. Fixed specificity values (0.80, 0.85, 0.90)

Prev	Fixed 0.80 specificity					Fixed 0.85 specificity					Fixed 0.90 specificity				
	Cut-off	Se	PPV	NPV	Mis%	Cut-off	Se	PPV	NPV	Mis%	Cut-off	Se	PPV	NPV	Mis%
0.05	515	0.71	0.16	0.98	20	575	0.63	0.18	0.98	16	665	0.52	0.22	0.97	12
0.10			0.28	0.96	21			0.32	0.95	17			0.37	0.94	14
0.20			0.47	0.92	22			0.51	0.90	19			0.56	0.88	18
0.30			0.60	0.86	23			0.64	0.84	22			0.69	0.81	21
0.40			0.70	0.80	24			0.74	0.76	24			0.78	0.74	25
0.50			0.79	0.73	25			0.81	0.70	26			0.84	0.65	29
0.70			0.89	0.53	26			0.91	0.50	30			092	0.45	37

Prev: prevalence, Sp: specificity, Se: sensitivity, PPV: positive predictive value, NPV: negative predictive value, Mis%: percent misclassified

Table 4. Summary of studies that additionally included CK-18 in multimarker model

Author	Target condition and population	Scoring system	Ingredients	AUC
Anty 2010	NAFLD grading among morbidly obese	The Nice Model	Metabolic syndrome, ALT, CK-18	Training: 0.88 Validation: 0.83
Boursier 2018	Fibrotic NASH among NAFLD	MACK-3	HOMA, AST, CK-18	Validation: 0.85
Cao 2013	NASH among NAFLD		ALT, platelets, M30, and TG	0.92
Chuah 2019	Fibrotic NASH among NAFLD	MACK-3	HOMA, AST, CK-18	0.80
Ergelen 2015	Fibrosis among NAFLD		TE, M30	F ≥2: 0.89 F ≥3: 0.93
Grigorescu 2012	NASH among NAFLD		M65, IL-6 and adiponectin	0.90
Pirvulescu 2012	NASH (including borderline NASH) among morbidly obese patients	NASH-score	BMI, ALT, AST, ALP, HOMA-IR, M65, and adiponectin	0.96
Rosso 2016	Fibrosis among NAFLD		TE, M30	F ≥2: 0.84 F ≥3: 0.87
Tada 2018	NASH among NAFLD	FIC-22	FIB-4 and CK-18	0.82
Tamimi 2011	NASH among NAFLD		Soluble fas and CK-18	Training: 0.93 Validation 0.79
Yang 2015	NASH among NAFLD		M30†, FGF-21, IL-1Ra, PEDF, and OPG	Training NPV: 0.76 and PPV: 0.85 Validation NPV: 0.80 and PPV: 0.76
Younossi 2008	NASH among NAFLD		M30 and M65, adiponectin, resistin	Training: 0.91 Validation: 0.73
Younossi 2011	NASH among NAFLD	NAFLD diagnostic panel	Diabetes, gender, BMI, triglycerides, M30, and M65	NASH: 0.81

NAFLD: non-alcoholic fatty liver disease, NASH: non-alcoholic steatohepatitis, ALT: alanine aminotransferase, CK-18: cytokeratin 18, AST: aspartate aminotransferase, TG: trigylceride, HOMA-IR: homeostatic model assessment for insulin resistance, TE: transient elastography, IL-6: interleukin 6, BMI: body max index, FGF-21: fibroblast growth factor 21, IL-1Ra: interleukin-1 receptor antagonist, NPV: negative predictive value, PPV: positive predictive value, PEDF: pigment epithelium-derived factor, OPG: osteoprotegerin, † Unit of measure for M30 is ng/L.

and 0.80 (95% CI: 0.70; 0.87), respectively. We additionally conducted sensitivity analysis solely among studies that included biopsy-proven NAFLD patients (19/22 studies for M30 and NASH), and found an AUC of 0.74 (95% CI: 0.67; 0.80). No marked differences were observed when excluding studies with high risk of bias or applicability concerns, different ELISA assays or cohorts with clinical suspicion of NAFLD.

Discussion

Main findings

Among NAFLD adults, the diagnostic accuracy of M30 to distinguish NASH from NAFL was under the minimally acceptable performance level, fixed a priori at AUC of 0.80. More promising results were observed for M65 and NASH, although it is of note that only six studies could be included in this meta-analysis, compared to 22 for M30. The superior performance of M65 should further be interpreted with caution, as its ability to detect fibrotic NASH, the most clinically relevant target condition, is limited.

At lower prevalence, mirroring primary care settings, high NPVs above 0.85 were achieved for both M30 and M65 antigens at fixed sensitivity and specificity values above 0.80 (Table 2 and Table 3).

Our meta-analysis on the accuracy of M30 in detecting fibrotic NASH also showed modest performance. MACK-3 showed more promise for detecting fibrotic NASH, but the evidence is still limited to two studies, and the model presents with limitations such as adequate performance among subgroups with metabolic syndrome and a large gap of patients who lie between the high and low threshold values (24, 55).

Results for both significant and advanced fibrosis were below the minimally acceptable performance level, demonstrating sub-optimal ability of M30 to function as a stand-alone test for fibrosis staging, even more so when considering the available accurate elastography methods and multimarker models for detecting liver fibrosis.

As expected, we observed a wide range of reported threshold values for both CK-18 antigens. This can be explained by the variability of methods employed for choosing a threshold and general lack of established recommendations. With our meta-analysis we suggest high and low thresholds for M30 and M65, which can be selected in accordance to the intention of use (ruling-in or ruling-out NASH). It is of note that the threshold suggestions for the M30 and M65 antigens are strictly for results produced by the PEVIVA assays, as it is understood that different CK-18 assays show poor inter-test reliability and majority of our studies used CK-18 assays from PEVIVIA (42).

Strengths and limitations

By employing novel meta-analytical methods, we were able to incorporate all data available in the primary studies, eliminating arbitrary selection of a single threshold for our meta-analyses. This allowed greater freedom to investigate which clinical setting would optimize the use of CK-18. A more comprehensive evaluation of the clinical performance, including projections of accuracy data (sensitivity, specificity, PPV, NPV) in various prevalence settings was possible. The multiple thresholds model further allowed us to assess the diagnostic accuracy of CK-18 at threshold values not investigated in the original studies. We were however limited in the sense that the data projected by our models are based on the cumulative distribution of CK-18 in the diseased and non-diseased populations of the primary studies, which had higher prevalence than one would expect in a primary care setting.

The approach for selecting either a single ‘optimal’ threshold value or a set of thresholds were very heterogeneous in our included studies. While some used the Youden or equivalent methods, others chose to optimize either the sensitivity or specificity, and a concerning few did not report how a threshold value was calculated. This was however anticipated as there is no recommended threshold for CK-18. We further observed sparse reporting of the histological procedure, including quality of biopsies and expertise of histological evaluation (S6 Table), which raises concerns regarding the reliability of the reference standard test.

In context of published literature

For M30 and NASH (22 studies), we found lower diagnostic accuracy compared to previous meta-analyses. He (2017) (14 studies) reported an AUC of 0.82 (60); Kwok (2014) (seven studies) reported a summary sensitivity of 0.66, at a specificity of 0.82 (13); Chen (2013) (nine studies) found an AUC of 0.84 (15); and Musso (2010) (nine studies) found an AUC of 0.82 (61). Parameters such as mean age, BMI and disease prevalence were not sources of major heterogeneity between the present and previously published meta-analyses (61). Our meta-analysis did however include a greater number of studies, incorporating more recent publications with lower performance. Among the six studies published after 2017, the AUC ranged between 0.59 and 0.77 for M30 in detecting NASH, a noticeable drop compared to pioneering work from 2008-10 (AUC: 0.71 to 0.88). The lowest AUC (0.59) was found in the largest study (N = 846) conducted in 2018. Interestingly, this study also found M30 to be most accurate in detecting patients with fibrotic NASH, achieving an AUC of 0.72 (55). In parallel with the incrementally less impressive results, the excitement for CK-18 as a NAFLD biomarker has tempered with each subsequent study, serving as an exemplar of the entire biomarker space.

The only other meta-analysis performed on the diagnostic ability of both M30 and M65 concluded that both antigens had similar ability to distinguish NASH from NAFL (M30 had AUC of 0.82, M65 had AUC of 0.80) (60). Among the three studies that investigated both M30 and M65 within the same cohort, all found better performance for M65 compared to M30 (9, 26, 57). Although M30 has been more popularly studied as a diagnostic biomarker for NASH, our meta-analysis demonstrates the need for more evidence to establish the performance of M65. Further studies conducting head-to-head comparisons of M30 and M65 within the same cohort would be valuable for assessing superior performance of either antigen.

Fibrotic NASH has become an emerging target condition of interest in NAFLD research (17). Despite the established role of hepatocyte apoptosis in the progression of liver damage (11), there have been contradictory opinions regarding the usefulness of CK-18 for fibrosis staging. Our results showed limited ability of CK-18 to function as a stand-alone test for detecting fibrotic patients compared to existing biomarkers.

Even still, the involvement of CK-18 in the disease pathway of NAFLD indicates potential for CK-18 to be used in combination with other biomarkers. Several promising models that included CK-18 (M30 and/or M65) were identified in our systematic review, most of which exceeded the minimally acceptable performance level of an AUC ≥0.80. Unfortunately, most models are limited to a single validation within the original studies with the exception of M30 with TE, and MACK-3, which raises the concern of how well the models would perform in practice. Additional validation studies for the proposed multimarker models should be conducted to ensure reliability of their performance. We do acknowledge that other studies including CK-18 in a composite scoring system may exist, despite not being eligible for inclusion in the present systematic review (62, 63). For example, a recent study developed a model for distinguishing NASH from NAFL, finding an AUC of 0.73 (0.66-0.81), with even better accuracy for detecting advanced fibrosis (63).

Implications for current practice and future research

Both the EASL-EASD-EASO and Asia-Pacific Working Party guidelines suggest that CK-18 has limited ability to function as a stand-alone test for distinguishing NASH from NAFL given its modest performance (12, 14). However, in a setting with 20% prevalence, a sensitivity of 0.90 and a NPV of 0.91 were achieved at a threshold value of 127 U/L (M30), demonstrating high negative values for ruling-out those without NASH. In such a scenario CK-18 could be of value as a first-line test at a primary care level for further evaluation by a specialist, even more so when considering the low cost and accessibility. This however comes at the cost of lower specificity, resulting in a high number of false positive results, as well as the compromise of 62% misclassified patients in the same setting with 20% prevalence. Alternatively, should CK-18 be used to rule-in NASH, a higher threshold of 399 U/L would be more appropriate. The trade-off between sensitivity and specificity as well as predictive values should be considered before selecting a threshold to be use in clinical practice, as a substantial number of patients without NASH could be referred for further, more invasive and risky evaluation.

CK-18 can potentially improve risk stratification in combination with other synergistic markers, such as TE or NFS, by testing for elevated M30 levels among patients under the low threshold or in patients with intermediate TE/NFS values (between the high and low threshold) (64). In the study by Liebig et al., risk stratification was considerably improved with this approach, showing more than 70% of patients with low TE/NFS but elevated M30 revealing presence of NASH (mostly with fibrosis). As with CK-18, other highly validated tests also run the risk of misclassified patients, for example, those with low or intermediate risk by TE who would not be considered for a biopsy despite presence of NASH. In such a step-wise diagnostic regime, a high cut-off for M30 should be selected to optimize specificity and rule-in those with NASH.

Acknowledgments

The authors thank Dr. Kee-Huat Chuah (Gastroenterology and Hepatology Unit, Department of Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia) for kindly sharing additional data, which enabled additional data included in the meta-analysis. We further thank the greater LITMUS systematic review team for their contribution in developing the protocol.

References

1. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology. 2016;64(1):73-84.
2. Younossi ZM, Stepanova M, Afendy M, Fang Y, Younossi Y, Mir H, et al. Changes in the prevalence of the most common causes of chronic liver diseases in the United States from 1988 to 2008. Clinical Gastroenterology and Hepatology. 2011;9(6):524-30. e1.
3. Satapathy SK, Sanyal AJ. Epidemiology and Natural History of Nonalcoholic Fatty Liver Disease. Seminars in liver disease. 2015;35(3):221-35.
4. Angulo P, Kleiner DE, Dam-Larsen S, Adams LA, Bjornsson ES, Charatcharoenwitthaya P, 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(2):389-97.e10.
5. 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.
6. Myers RP, Fong A, Shaheen AA. Utilization rates, complications and costs of percutaneous liver biopsy: a population-based study including 4275 biopsies. Liver international : official journal of the International Association for the Study of the Liver. 2008;28(5):705-12.
7. 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.
8. Feldstein AE, Wieckowska A, Lopez AR, Liu YC, Zein NN, McCullough AJ. Cytokeratin-18 fragment levels as noninvasive biomarkers for nonalcoholic steatohepatitis: a multicenter validation study. Hepatology. 2009;50(4):1072-8.
9. Joka D, Wahl K, Moeller S, Schlue J, Vaske B, Bahr MJ, et al. Prospective biopsy-controlled evaluation of cell death biomarkers for prediction of liver fibrosis and nonalcoholic steatohepatitis. Hepatology. 2012;55(2):455-64.
10. Cusi K, Chang Z, Harrison S, Lomonaco R, Bril F, Orsak B, et al. Limited value of plasma cytokeratin-18 as a biomarker for NASH and fibrosis in patients with non-alcoholic fatty liver disease. Journal of Hepatology. 2014;60(1):167-74.
11. Eguchi A, Wree A, Feldstein AE. Biomarkers of liver cell death. J Hepatol. 2014;60(5):1063-74.
12. EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease. J Hepatol. 2016;64(6):1388-402.
13. Kwok R, Tse YK, Wong GL, Ha Y, Lee AU, Ngu MC, et al. Systematic review with meta-analysis: non-invasive assessment of non-alcoholic fatty liver disease--the role of transient elastography and plasma cytokeratin-18 fragments. Aliment Pharmacol Ther. 2014;39(3):254-69.
14. Wong VW, Chan WK, Chitturi S, Chawla Y, Dan YY, Duseja A, et al. Asia-Pacific Working Party on Non-alcoholic Fatty Liver Disease guidelines 2017-Part 1: Definition, risk factors and assessment. Journal of gastroenterology and hepatology. 2018;33(1):70-85.
15. Chen J, Zhu Y, Zheng Q, Jiang J. Serum cytokeratin-18 in the diagnosis of non-alcoholic steatohepatitis: A meta-analysis. Hepatology research : the official journal of the Japan Society of Hepatology. 2014;44(8):854-62.
16. 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.
17. 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. J Clin Gastroenterol. 2011;45(1):55-8.
18. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36.
19. Burkner PC, Doebler P. Testing for publication bias in diagnostic meta-analysis: a simulation study. Statistics in medicine. 2014;33(18):3061-77.

20. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005;58(10):982-90.
21. Aida Y, Abe H, Tomita Y, Nagano T, Seki N, Sugita T, et al. Serum cytokeratin 18 fragment level as a noninvasive biomarker for non-alcoholic fatty liver disease. International journal of clinical and experimental medicine. 2014;7(11):4191-8.
22. Cao W, Zhao C, Shen C, Wang Y. Cytokeratin 18, alanine aminotransferase, platelets and triglycerides predict the presence of nonalcoholic steatohepatitis. PLoS ONE [Electronic Resource]. 2013;8(12):e82092.
23. Chan WK, Sthaneshwar P, Mustapha NRN, Mahadeva S. Limited utility of plasma M30 in discriminating non-alcoholic steatohepatitis from steatosis-a comparison with routine biochemical markers. Journal of Gastroenterology and Hepatology (Australia). 2014;3:182-3.
24. Chuah KH, Wan Yusoff WNI, Sthaneshwar P, Nik Mustapha NR, Mahadeva S, Chan WK. MACK-3 (combination of hoMa, Ast and CK18): A promising novel biomarker for fibrotic non-alcoholic steatohepatitis. Liver International. 2019;01:01.
25. Darweesh SK, AbdElAziz RA, Abd-ElFatah DS, AbdElazim NA, Fathi SA, Attia D, et al. Serum cytokeratin-18 and its relation to liver fibrosis and steatosis diagnosed by FibroScan and controlled attenuation parameter in nonalcoholic fatty liver disease and hepatitis C virus patients. European Journal of Gastroenterology & Hepatology. 2019;31(5):633-41.
26. Dvorak K, Stritesky J, Petrtyl J, Vitek L, Sroubkova R, Lenicek M, et al. Use of non-invasive parameters of non-alcoholic steatohepatitis and liver fibrosis in daily practice--an exploratory case-control study. PLoS ONE [Electronic Resource]. 2014;9(10):e111551.
27. Ergelen R, Akyuz U, Aydin Y, Eren F, Yilmaz Y. Measurements of serum procollagen-III peptide and M30 do not improve the diagnostic accuracy of transient elastography for the detection of hepatic fibrosis in patients with nonalcoholic fatty liver disease. European Journal of Gastroenterology & Hepatology. 2015;27(6):667-71.
28. Grigorescu M, Crisan D, Radu C, Grigorescu MD, Sparchez Z, Serban A. A novel pathophysiological-based panel of biomarkers for the diagnosis of nonalcoholic steatohepatitis. Journal of physiology and pharmacology : an official journal of the Polish Physiological Society. 2012;63(4):347-53.
29. Hasegawa Y, Kim SR, Hatae T, Ohta M, Fujinami A, Sugimoto K, et al. Usefulness of Cytokeratin-18M65 in Diagnosing Non-Alcoholic Steatohepatitis in Japanese Population. Digestive Diseases. 2015;33(6):715-20.
30. Huang JF, Yeh ML, Huang CF, Huang CI, Tsai PC, Tai CM, et al. Cytokeratin-18 and uric acid predicts disease severity in Taiwanese nonalcoholic steatohepatitis patients. PLoS ONE [Electronic Resource]. 2017;12(5):e0174394.
31. Kamada Y, Akita M, Takeda Y, Yamada S, Fujii H, Sawai Y, et al. Serum Fucosylated Haptoglobin as a Novel Diagnostic Biomarker for Predicting Hepatocyte Ballooning and Nonalcoholic Steatohepatitis. PLoS ONE [Electronic Resource]. 2013;8(6):e66328.
32. Kawanaka M, Nishino K, Nakamura J, Urata N, Oka T, Goto D, et al. Correlation between serum cytokeratin-18 and the progression or regression of non-alcoholic fatty liver disease. Annals of Hepatology. 2015;14(6):837-44.
33. Kazankov K, Barrera F, Moller HJ, Rosso C, Bugianesi E, David E, et al. The macrophage activation marker sCD163 is associated with morphological disease stages in patients with non-alcoholic fatty liver disease.[Erratum appears in Liver Int. 2017 Nov;37(11):1745; PMID: 29065254]. Liver International. 2016;36(10):1549-57.
34. Kim YS, Jung ES, Hur W, Bae SH, Choi JY, Song MJ, et al. Noninvasive predictors of nonalcoholic steatohepatitis in Korean patients with histologically proven nonalcoholic fatty liver disease. Clinical and Molecular Hepatology. 2013;19(2):120-30.
35. Kobayashi N, Kumada T, Toyoda H, Tada T, Ito T, Kage M, et al. Ability of Cytokeratin-18 Fragments and FIB-4 Index to Diagnose Overall and Mild Fibrosis Nonalcoholic Steatohepatitis in Japanese Nonalcoholic Fatty Liver Disease Patients. Digestive Diseases. 2017;35(6):521-30.
36. Liu XL, Pan Q, Zhang RN, Shen F, Yan SY, Sun C, et al. Disease-specific miR-34a as diagnostic marker of non-alcoholic steatohepatitis in a Chinese population. World Journal of Gastroenterology. 2016;22(44):9844-52.

37. Liu WY, Zheng KI, Pan XY, Ma HL, Zhu PW, Wu XX, et al. Effect of PNPLA3 polymorphism on diagnostic performance of various noninvasive markers for diagnosing and staging nonalcoholic fatty liver disease. Journal of Gastroenterology and Hepatology. 2019.

38. Malik R, Chang M, Bhaskar K, Nasser I, Curry M, Schuppan D, et al. The clinical utility of biomarkers and the nonalcoholic steatohepatitis CRN liver biopsy scoring system in patients with nonalcoholic fatty liver disease. Journal of Gastroenterology & Hepatology. 2008;24(4):564-8.

39. Mohammed MA, Omar NM, Mohammed SA, Amin AM, Gad DF. FICK-3 Score Combining Fibrosis-4, Insulin Resistance and Cytokeratin-18 in Predicting Non-alcoholic Steatohepatitis in NAFLD Egyptian Patients. Pak. 2019;22(10):457-66.

40. Musso G, Gambino R, Durazzo M, Cassader M. Noninvasive assessment of liver disease severity with liver fat score and CK-18 in NAFLD: Prognostic value of liver fat equation goes beyond hepatic fat estimation. Hepatology. 2010;51(2):715-7.

41. Papatheodoridis GV, Hadziyannis E, Tsochatzis E, Georgiou A, Kafiri G, Tiniakos DG, et al. Serum apoptotic caspase activity in chronic hepatitis C and nonalcoholic Fatty liver disease. J Clin Gastroenterol. 2010;44(4):e87-95.

42. Pimentel CF, Jiang ZG, Otsubo T, Feldbrugge L, Challies TL, Nasser I, et al. Poor Inter-test Reliability Between CK18 Kits as a Biomarker of NASH. Digestive Diseases & Sciences. 2016;61(3):905-12.

43. Rosso C, Caviglia GP, Abate ML, Vanni E, Mezzabotta L, Touscoz GA, et al. Cytokeratin 18-Aspartate396 apoptotic fragment for fibrosis detection in patients with non-alcoholic fatty liver disease and chronic viral hepatitis. Digestive & Liver Disease. 2016;48(1):55-61.

44. Shen J, Chan HL, Wong GL, Chan AW, Choi PC, Chan HY, et al. Assessment of non-alcoholic fatty liver disease using serum total cell death and apoptosis markers. Alimentary Pharmacology & Therapeutics. 2012;36(11):1057-66.

45. Tada T, Kumada T, Toyoda H, Saibara T, Ono M, Kage M. New scoring system combining the FIB-4 index and cytokeratin-18 fragments for predicting steatohepatitis and liver fibrosis in patients with nonalcoholic fatty liver disease. Biomarkers. 2018;23(4):328-34.

46. Tamimi TI, Elgouhari HM, Alkhouri N, Yerian LM, Berk MP, Lopez R, et al. An apoptosis panel for nonalcoholic steatohepatitis diagnosis. Journal of Hepatology. 2010;54(6):1224-9.

47. Valva P, Rios D, Casciato P, Gadano A, Galdame O, Mullen E, et al. Nonalcoholic fatty liver disease: biomarkers as diagnostic tools for liver damage assessment in adult patients from Argentina. European Journal of Gastroenterology & Hepatology. 2018;30(6):637-44.

48. Wieckowska A, Zein NN, Yerian LM, Lopez AR, McCullough AJ, Feldstein AE. In vivo assessment of liver cell apoptosis as a novel biomarker of disease severity in nonalcoholic fatty liver disease. Hepatology. 2006;44(1):27-33.

49. Yang M, Xu D, Liu Y, Guo X, Li W, Guo C, et al. Combined Serum Biomarkers in Non-Invasive Diagnosis of Non-Alcoholic Steatohepatitis. PLoS ONE [Electronic Resource]. 2015;10(6):e0131664.

50. Yilmaz Y, Dolar E, Ulukaya E, Akgoz S, Keskin M, Kiyici M, et al. Soluble forms of extracellular cytokeratin 18 may differentiate simple steatosis from nonalcoholic steatohepatitis. World Journal of Gastroenterology. 2007;13(6):837-44.

51. Younes R, Rosso C, Petta S, Cucco M, Marietti M, Caviglia GP, et al. Usefulness of the index of NASH - ION for the diagnosis of steatohepatitis in patients with non-alcoholic fatty liver: An external validation study. Liver International. 2018;38(4):715-23.

52. Younossi ZM, Jarrar M, Nugent C, R, hawa M, Afendy M, et al. A novel diagnostic biomarker panel for obesity-related nonalcoholic steatohepatitis (NASH). Obesity Surgery. 2008;18(11):1430-7.

53. Zheng KI, Liu WY, Pan XY, Ma HL, Zhu PW, Wu XX, et al. Combined and sequential non-invasive approach to diagnosing non-alcoholic steatohepatitis in patients with non-alcoholic fatty liver disease and persistently normal alanine aminotransferase levels. BMJ open diabetes res. 2020;8(1):03.

54. Anty R, Iannelli A, Patouraux S, Bonnafous S, Lavallard VJ, Senni-Buratti M, et al. A new composite model including metabolic syndrome, alanine aminotransferase and cytokeratin-18 for the diagnosis of non-alcoholic steatohepatitis in morbidly obese patients. Alimentary Pharmacology & Therapeutics. 2010;32(11):1315-22.

55. Boursier J, Anty R, Vonghia L, Moal V, Vanwolleghem T, Canivet CM, et al. Screening for therapeutic trials and treatment indication in clinical practice: MACK-3, a new blood test for the diagnosis of fibrotic NASH. Alimentary Pharmacology & Therapeutics. 2018;47(10):1387-96.

56. Diab DL, Yerian L, Schauer P, Kashyap SR, Lopez R, Hazen SL, et al. Cytokeratin 18 fragment levels as a noninvasive biomarker for nonalcoholic steatohepatitis in bariatric surgery patients. Clinical Gastroenterology & Hepatology. 2008;6(11):1249-54.

57. Pirvulescu I, Gheorghe L, Csiki I, Becheanu G, Dumbrava M, Fica S, et al. Noninvasive clinical model for the diagnosis of nonalcoholic steatohepatitis in overweight and morbidly obese patients undergoing bariatric surgery. Chirurgia (Bucuresti). 2012;107(6):772-9.

58. Younossi ZM, Page S, Rafiq N, Birerdinc A, Stepanova M, Hossain N, et al. A biomarker panel for non-alcoholic steatohepatitis (NASH) and NASH-related fibrosis. Obesity Surgery. 2011;21(4):431-9.

59. Yilmaz Y, Ulukaya E, Dolar E. Serum M30 levels: a potential biomarker of severe liver disease in nonalcoholic fatty liver disease and normal aminotransferase levels. Hepatology. 2009;49(2):697-.

60. He L, Deng L, Zhang Q, Guo J, Zhou J, Song W, et al. Diagnostic Value of CK-18, FGF-21, and Related Biomarker Panel in Nonalcoholic Fatty Liver Disease: A Systematic Review and Meta-Analysis. BioMed research international. 2017;2017:9729107.

61. Musso G, Gambino R, Cassader M, Pagano G. Meta-analysis: natural history of non-alcoholic fatty liver disease (NAFLD) and diagnostic accuracy of non-invasive tests for liver disease severity. Annals of medicine. 2011;43(8):617-49.

62. Shen J, Chan HL-Y, Wong GL-H, Choi PC-L, Chan AW-H, Chan H-Y, et al. Non-invasive diagnosis of non-alcoholic steatohepatitis by combined serum biomarkers. Journal of Hepatology. 2012;56(6):1363-70.

63. Canbay A, Kälsch J, Neumann U, Rau M, Hohenester S, Baba HA, et al. Non-invasive assessment of NAFLD as systemic disease-A machine learning perspective. PLoS One. 2019;14(3):e0214436-e.

64. Liebig S, Stoeckmann N, Geier A, Rau M, Schattenberg JM, Bahr MJ, et al. Multicenter Validation Study of a Diagnostic Algorithm to Detect NASH and Fibrosis in NAFLD Patients With Low NAFLD Fibrosis Score or Liver Stiffness. Clin Transl Gastroenterol. 2019;10(8):e00066-e.

Supplementary Material

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0238717#sec029

03

Diagnostic accuracy of non-invasive tests for advanced fibrosis in patients with NAFLD – An individual patient data meta-analysis

Ferenc E. Mózes   Jenny Lee   Emmanuel A. Selvaraj
Arjun N. A. Jayaswal   Michael Trauner   Jérôme Boursier
Céline Fournier   Katharina Staufer   Rudolf E. Stauber
Elisabetta Bugianesi   Ramy Younes   Silvia Gaia
Monica Lupșor-Platon   Salvatore Petta   Toshihide Shima
Takeshi Okanoue   Sanjiv Mahadeva   Wah-Kheong Chan
Peter J. Eddowes   Philip N. Newsome   Vincent Wai-Sun Wong
Victor de Lédinghen   Jian-Gao Fan   Feng Shen
Jeremy F. Cobbold   Yoshio Sumida   Akira Okajima
Jörn M. Schattenberg   Christian Labenz   Won Kim
Myoung Seok Lee   Johannes Wiegand   Thomas Karlas
Yusuf Yilmaz   Guruprasad Padur Aithal
Naaventhan Palaniyappan   Christophe Cassinotto
Sandeep Aggarwal Harshit Garg   Geraldine Ooi
Atsushi Nakajima   Masato Yoneda   Marianne Ziol
Nathalie Barget   Andreas Geier   Theresa Tuthill
Julia M. Brosnan   Quentin M. Anstee   Stefan Neubauer
Stephen A. Harrison   Patrick M. Bossuyt   Michael Pavlides

Gut, 2022, 71.5: 1006-1019.

Abstract

Objective: Liver biopsy is still needed for fibrosis staging in many patients with non-alcoholic fatty liver disease. The aims of this study were to evaluate the individual diagnostic performance of liver stiffness measurement by vibration controlled transient elastography (LSM- VCTE), Fibrosis-4 index (FIB-4) and NAFLD Fibrosis Score (NFS) and to derive diagnostic strategies that could reduce the need for liver biopsies.

Design: Individual patient data meta-analysis of studies evaluating LSM-VCTE against liver histology was conducted. FIB-4 and NFS were computed where possible. Sensitivity, specificity and area under the receiver operating curve (AUROC) were calculated. Biomarkers were assessed individually and in sequential combinations.

Results: Data were included from 37 primary studies (n=5735; 45% female; median age: 54 years; median BMI: 30 kg/m^2; 33% had type 2 diabetes; 30% had advanced fibrosis). AUROCs of individual LSM-VCTE, FIB-4 and NFS for advanced fibrosis were 0.85, 0.76 and 0.73. Sequential combination of FIB-4 cut-offs (<1.3; ≥2.67) followed by LSM-VCTE cut-offs (<8.0; ≥10.0kPa) to rule-in or rule-out advanced fibrosis had sensitivity and specificity (95% CI) of 66% (63-68) and 86% (84-87) with 33% needing a biopsy to establish a final diagnosis. FIB-4 cut-offs (<1.3; ≥3.48) followed by LSM cut-offs (<8.0; ≥20.0kPa) to rule out advanced fibrosis or rule in cirrhosis had a sensitivity of 38% (37-39) and specificity of 90% (89-91) with 19% needing biopsy.

Conclusion: Sequential combinations of markers with a lower cut-off to rule-out advanced fibrosis and a higher cut-off to rule-in cirrhosis can reduce the need for liver biopsies.

Introduction

Non-alcoholic fatty liver disease (NAFLD) is the hepatic manifestation of the metabolic syndrome with high prevalence worldwide [1]. Most patients remain asymptomatic for long periods of time (years/decades) with slowly progressive disease, but a minority [2] progress to cirrhosis, liver failure, and hepatocellular carcinoma (HCC).

NAFLD comprises several histological features ranging from simple steatosis to steatosis with lobular inflammation and ballooned hepatocytes (steatohepatitis), both of which can be accompanied by varying degrees of fibrosis. The currently accepted reference standard for diagnosing NAFLD is liver biopsy as its diagnostic features are based on histology [3]. Liver biopsy, however, is invasive and carries a risk of complications [4], is limited by sampling variability [5] and high observer dependent variability in pathological reporting [6,7].

NAFLD is often diagnosed after incidental findings of elevated liver transaminases on blood tests, or liver steatosis or cirrhosis on imaging. One challenge clinicians face is to identify which of these patients are at high risk of progression or clinical outcomes, as they would benefit from specialist follow-up. There is now substantial evidence showing that those with at least advanced fibrosis (F3-4) are at higher risk of liver-related events in later life [8–10].

A large body of evidence also exists on how non-invasive tests (NITs) could be used to risk-stratify patients for the presence of advanced fibrosis. These approaches usually involve sequential application of two NITs, with the first tier of a simple, inexpensive, serum-based test performed in the community (e.g. Fibrosis-4 index (FIB-4) or NAFLD fibrosis score (NFS)), followed by a second tier of liver stiffness measurement (LSM) (e.g., vibration controlled transient elastography; VCTE), or a proprietary serum-based test (e.g. enhanced liver fibrosis test; ELF). A lower and an upper threshold are usually used in each tier of testing to rule out (those with a NIT result less than the lower threshold) or rule in (those with a NIT result more than the upper threshold) patients at high risk of advanced fibrosis. Patients with indeterminate results in both tiers of testing would need a liver biopsy for risk stratification. The main value of these approaches lies in their high negative predictive value to rule out patients with low risk of advanced fibrosis who can be safely managed in primary care.

Despite the increasing evidence to support these approaches, some aspects of their application require further clarifications. First, there is no consensus on which NIT thresholds to use for this purpose. For example, FIB-4 upper cut-offs of 3.25 [11] and 2.67 [12] have been described, while other investigators omit the FIB-4 upper cut-off altogether [13]. There is also some uncertainty about the performance of NITs in specific patient subgroups, such as those with diabetes or obesity. Furthermore, for patients who are ruled in as being at high risk of advanced fibrosis (F3-4), liver biopsy is often needed to identify those with cirrhosis who would need surveillance for HCC [14]. Developing approaches that can minimise the need for liver biopsy in secondary care is therefore an area of unmet need.

To address these problems, we conducted an individual patient data meta-analysis (IPDMA) with three main aims: 1) to evaluate the performance of LSM-VCTE and compare it to the performance of FIB-4 and NFS as screening tests to rule out advanced fibrosis; 2) to evaluate NIT combination strategies to minimise the number of cases that would need a liver biopsy in secondary care; 3) to explore factors that influence diagnostic accuracy.

Methods

This IPDMA was reported in accordance with the recommendations of the PRISMA-IPD Statement [15] and was registered as PROSPERO CRD42019157661.

Criteria for considering studies for the IPD meta-analysis

Patients

Studies reporting data on adults (≥18 years) with NAFLD and paired liver histology and liver stiffness measurements by (LSM-VCTE) were eligible. When studies reported study groups of participants with unselected aetiologies, only IPD of those with NAFLD were sought.

Index tests

The index test of main interest was LSM-VCTE performed with FibroScan® (Echosens, France). Results for serum-based biomarkers NSF [16], FIB-4 [17], aspartate aminotransferase (AST) to alanine aminotransferase (ALT) ratio [18], and AST-to-platelet ratio index (APRI) [19]) were also computed where data was available. Supporting Table 1 summarises the definition of NITs considered in this IPDMA.

Universally accepted cut-offs for diagnosing different groups of fibrosis stages do not exist (several suggested cut-offs are presented in Supporting Table 2). For LSM-VCTE, <7.9 kPa and ⍰⍰9.6 kPa are the most used for respectively ruling out and in, advanced fibrosis [20].

Reference standard

Only studies reporting histological classification of liver fibrosis based on the NASH CRN staging system were considered [21].

Target conditions

Advanced fibrosis (F3-4) and cirrhosis (F4) were the target conditions of interest. To fulfil the aims of the study, cut-offs were selected to rule out or rule in advanced fibrosis, and to rule out advanced fibrosis or rule in cirrhosis.

Study design

All study designs were considered if they were reporting on patients with NAFLD undergoing both liver biopsy and LSM-VCTE within 6 months. No language restrictions were applied.

Establishing collaborations

Authors of eligible studies were contacted by email and reminders were sent if a response was not received within 2 weeks. Only data from studies that received ethical approval were used. Additional ethical approval was not sought for the meta-analysis as only anonymised data were provided.

Data verification

Range checks of measurement values provided for individual patients were carried out and authors were asked to provide clarifications where necessary. Missing data were queried until received or confirmed as unavailable. Missing data were handled in the analysis by pairwise deletion.

LSM-VCTE with median stiffness ≥7.1 kPa and IQR-to-median LSM ratio >30% were considered unreliable [22]. These were included in the main analysis and were later compared in a subgroup analysis to reliable measurements, to assess whether they can be reliably used to diagnose advanced fibrosis.

Authors were provided with a template table of required data (Supporting Table 3) and were asked to de-duplicate data were possible. We also checked for duplicate entries and were identified these were removed.

Data analysis

Quality and bias assessment

The quality of studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS-2) [23].

IPD meta-analysis

The original data sets were merged, a study identification variable was added, and descriptive statistical analysis of the data sets was conducted. Dichotomous variables are displayed as percentages. Continuous variables are reported as means with standard deviations, or medians with interquartile ranges according to the distribution of the data.

Analyses were done per-protocol, as we did not have information on failed LSM-VCTE. To express the diagnostic performance of NITs, non-parametric, empirical receiver operating characteristic (ROC) curves were constructed for the target conditions of interest. Diagnostic performance was expressed as the area under the ROC curve (AUROC) with 95% confidence intervals (95% CI), based on De Long’s method. AUROCs were compared using De Long’s test statistic.

Thresholds to maximise the Youden index (i.e. sensitivity+specificity-1), for 90% sensitivity, and for 90% specificity were reported. The diagnostic performance of previously published cut-offs was also evaluated. Sequential combinations of serum biomarkers and LSM-VCTE were evaluated, by computing sensitivity, specificity, and proportions of misclassified and indeterminate patients.

Positive and negative predictive values (PPV and NPV) were estimated for prevalence within the range of those reported in the original studies. The number of false positive and false negative results for 100 theoretical cases were also reported.

The main analysis was conducted to maximise data for each NIT. For a valid comparison of the performance of NITs, a separate analysis was conducted in the subgroup of patients where all three of VCTE, FIB-4, and NFS were available in each participant.

To fulfil the aim of developing testing strategies that reduce the number of patients in need of a liver biopsy, lower cut-offs for ruling out advanced fibrosis and upper cut-offs for ruling in cirrhosis were used. The rationale for this approach is illustrated in Supporting Figure 1. The upper cut-offs for identifying cirrhosis were chosen at 95% and 98% specificity in a derivation set and tested in validation set. Derivation and validation sets were obtained by random sampling from the IPD study group in a 3:2 ratio. These upper cut-offs were combined with lower cut-offs from the literature for ruling out advanced fibrosis and the algorithm was tested in the whole IPD study group. For ease of reference, we also examined the cut-offs of 8 kPa and 10 kPa (corresponding to the most common VCTE cut-offs in the literature of 7.9 kPa and 9.6 kPa rounded to the nearest integer) and also rounded our cirrhosis cut-offs to the nearest integer to facilitate application in clinical practice.

Only test-positive and test-negative patients were included in the calculation of diagnostic performance indices, and patients in the indeterminate group were excluded from calculations.

Subgroup analysis was performed according to biopsy length (<20 mm, ≥20 mm), number of portal tracts in biopsy samples (<11, ≥11), biopsy quality (intermediate: 10 mm ≤ length <20 mm; high: length ≥20 mm and ≥11 tracts), age (four quartiles), sex, body-mass index (BMI; BMI<25 kg/m^2, 25 kg/m^2 ≤BMI<30 kg/m^2, BMI≥30 kg/m^2), presence of type 2 diabetes mellitus (T2DM), continent of provenance (Europe, Asia), probes used (M, XL), reliability criteria for LSM-VCTE (reliable (median LSM<7.1 kPa or median LSM≥7.1 kPa and IQR/median LSM<0.30) versus unreliable (median LSM≥7.1 kPa and IQR/median LSM≥0.30) [22]; reliable (IQR/median LSM<0.30) versus unreliable (IQR/median LSM≥0.30)), and aminotransferase levels (ALT or AST<40, 40≤ALT or AST<100, ALT or AST≥100; ALT<40 and AST<40, ALT≥40 or AST≥40.

All statistical analyses were performed using R (version 1.2.1335, R Foundation for Statistical Computing, Vienna, Austria) with the pROC package [24,25]; 95% confidence intervals were calculated using 500 stratified bootstrap replicates using the boot package [26,27].

VCTE probe types

The analysis to account for probe type is described in the Supporting Materials.

Patient and public involvement

Patients and the public were not involved in the conduct of this study as there was no direct patient participation in the study.

Results

Search process and data collection

10392 articles were identified in a search performed for a larger systematic review evaluating the diagnostic performance of LSM-VCTE and other index tests for the staging of fibrosis and diagnosis of non-alcoholic steatohepatitis (NASH) in adult patients with NAFLD. After removing

duplicates, and screening titles, abstracts, and full texts, 59 studies examining VCTE were identified. The authors of 37 studies shared useable data (Figure 1). Authors of more than one study supplied data in a single dataset and, overall, we received 30 data sets including data from 6571 patients. After removing duplicates (n=628) and patients with missing biopsy (n=14) or LSM-VCTE (n=194) data, the final dataset consisted of 5735 unique patients.

Study and population characteristics

The characteristics of the 30 data sets are summarised in Table 1. Studies were conducted in Europe (67%), Asia (40%) and Australia (3%). Data availability is shown in Supporting Table 3. FIB-4 and NFS were determined in 5393 (94%) and 3248 (57%) cases, respectively. Median age was 54 years, 2570 (45%) patients were female, 33% had diabetes and 43%

had BMI≥30 kg/m^2. Overall, 30% had advanced fibrosis and 11% had cirrhosis. Details of the IPD study group are included in Table 2, and Supporting Tables 4 and 5.

Study quality

The methodological quality of the studies assessed with the QUADAS-2 tool is summarised in Supporting Figures 2 and 3. Only one study had low risk of bias or low applicability concerns in all QUADAS-2 domains [28]. The flow and timing domain were judged to have high risk or unclear risk of bias in 65% of studies, as these either excluded technical failures from their final diagnostic performance analysis or did not report them.

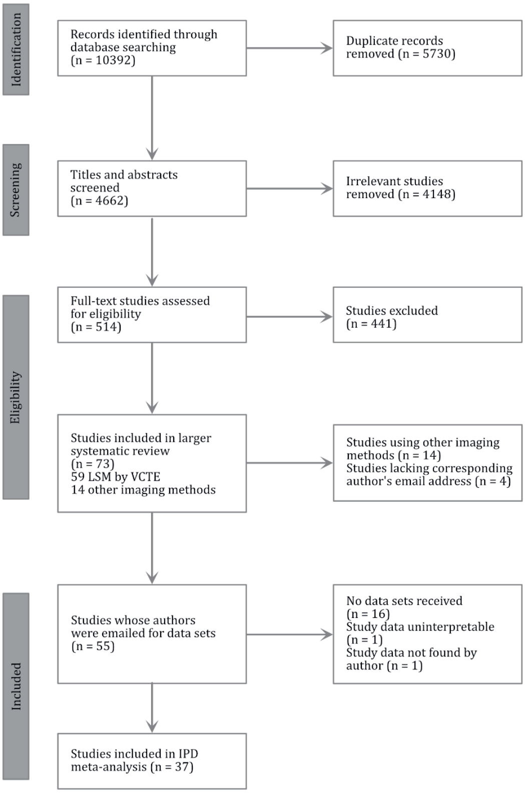

Identification
Records identified through database searching (n = 10392)
Duplicate records removed (n = 5730)
Screening
Titles and abstracts screened (n = 4662)
Irrelevant studies removed (n = 4148)
Eligibility
Full-text studies assessed for eligibility (n = 514)
Studies excluded (n = 441)
Studies included in larger systematic review (n = 73) 59 LSM by VCTE 14 other imaging methods
Studies using other imaging methods (n = 14) Studies lacking corresponding author's email address (n = 4)
Included
Studies whose authors were emailed for data sets (n = 55)
No data sets received (n = 16) Study data uninterpretable (n = 1) Study data not found by author (n = 1)
Studies included in IPD meta-analysis (n = 37)

Figure 1. PRISMA flow chart illustrating the identification and selection process for studies finally included in this individual patient data meta-analysis.

Table 1. Details of individual patient data included in this meta-analysis.

Data set ID	Country	Study Design	Number of participants (n)	Age (yr)	BMI (kg/m²)	WC (cm)	M/F	Recruitment interval	Hardware used	Probe used
Agrawal 2017 [48]	UK	MC, P, CS	25	47.8 (19-70)	27.7 (15.8-35.7)	95.4 (39-120)	18/7	2009-2012	-	M
Aykut 2014 [49]	Turkey	SC, P, CS	88	46.0 (24-62)	30.3 (18.3-41.8)	101.5 (70-143)	50/38	-	FibroScan 502 Touch	M
Boursier [50–52]	France	MC, P, CS	1063	56.1 (18-83)	31.6 (16.7-55.5)	108.3 (58-174)	613/450	-	-	M or XL
Cassinotto 2013 [53]	France	SC, P, CS	61	55.9 (22-81)	30.1 (16.7-46.6)	103.6 (72-125)	40/21	2010-2012	-	M and XL
Cassinotto 2016 [54]	France	MC, P, CS	286	56.6 (18-80)	32.2 (20.3-57.4)	109.8 (68-168)	171/115	2011-2015	-	M and XL
Chan 2015 [55]	Malaysia	SC, P, CS	146	50.4 (18-73)	29.4 (6.9-41.2)	98.3 (79-127)	80/66	2012-2013	FibroScan 502 Touch	M
Chan 2017 [56]	Malaysia, Hong Kong	MC, P, CC	153	54.0 (24-76)	29.9 (20.1-44.8)	98.4 (69-141)	68/85	-	FibroScan 502 Touch	M and XL
Eddowes 2016 [57,58]	UK	MC, P, CS	358	53.3 (19-77)	34.2 (19.5-53.2)	117.2 (65-158)	206/152	-	-	M or XL
Eddowes 2019 [30]	UK	MC, P, CS	50	50.2 (18-73)	33.6 (23.6-47.8)	109.4 (89-132)	28/22	2014-2015	-	M or XL
Gaia 2011 [59]	Italy	SC, P, CS	68	46.8 (28-65)	28.0 (21.2-40.2)	-	48/20	2007-2009	-	M
Garg 2018 [60]	India	SC, P, CS	76	38.2 (20-65)	45.2 (32.3-73.8)	-	16/60	2014-2016	FibroScan 502 Touch	XL

Karlas 2015 [61]	Germany	SC, P, CS	41	45.7 (28-64)	47.7 (33.7-60.1)	-	13/28	-	FibroScan 502	XL
Labenz 2018 [62]	Germany	SC, P, CS	126	47.4 (20-73)	31.6 (23.2-50.4)	-	72/54	-	FibroScan 402	M or XL
Lee 2017 [63]	Korea	SC, P, CS	94	55.5 (19-82)	27.2 (19.1-36.3)	-	41/53	2014-2015	-	M or XL
Lupsor 2010 [64]	Romania	SC, P, CS	72	42.4 (20-69)	29.7 (21.0-41.5)	102.4 (60-124)	51/21	2007-2009	-	M
Mahadeva 2013 [65]	Malaysia	SC, P, CS	131	49.9 (23-73)	28.7 (18.6-43.1)	93.5 (43-128)	66/65	2009-2010	-	M
Okajima 2017 [66]	Japan	SC, P, CS	173	56.3 (18-81)	27.2 (16.5-40.3)	-	84/89	2013-2015	-	M
Ooi 2018 [67]	Australia	MC, P, CS	82	44.5 (18-67)	46.2 (29.1-74.0)	136.5 (101-192)	23/59	2015-2016	-	M or XL
Pavlides 2017 [68]	UK	SC, P, CS	70	53.5 (25-77)	34.5 (23.0-57.3)	112.5 (80-149)	42/28	2011-2015	-	M or XL
Petta 2015 [69,70]	Italy	MC, P&R, CS	234	45.5 (15-78)	28.2 (15.7-40.7)	99.4 (69-126)	169/65	2008-2013	-	M
Petta 2016 [28]	France, Hong Kong, Italy	MC, P, CS	260	54.6 (15-87)	29.4 (16.5-46.6)	100.9 (74-148)	122/138	-	-	M
Petta 2017 [71]	Italy	MC, P, CS	474	45.5 (19-77)	29.2 (15.2-49.5)	99.6 (47-164)	275/199	-	-	M
Seki 2017 [72]	Japan	SC, P, CS	181	57.7 (16-82)	27.1 (16.9-38.1)	95.1 (71-117)	91/90	2013-2015	-	M
Shen 2015 [73]	China	MC, P, CS	101	59.0 (16-67)	27.0 (20.1-37.3)	92.9 (75-120)	74/27	2012-2014	FibroScan 502	M

Staufer 2019 [74]	Austria	MC, P, CS	186	49.6 (19-83)	32.5 (19.0-56.9)	-	106/80	2011-2016	FibroScan 502 Touch	M or XL
Wong 2019 [75–78]	Hong Kong, France	MC, P, CS	464	53.8 (20-83)	30.5 (17.3-48.0)	102.0 (71-148)	201/263	2009-2017	-	M and XL
Wong 2010 [20]	Hong Kong, France	MC, P, CS	273	51.6 (21-77)	28.8 (16.5-54.0)	96.2 (65-144)	147/126	2003-2009	-	M
Yoneda 2008 [79]	Japan	MC, P, CS	97	52.1 (19-76)	26.5 (17.9-38.5)	-	41/56	< 2008	-	M
Younes 2017 [80]	Italy	MC, P, CS	289	44.8 (15-78)	28.8 (17.5-41.7)	98.9 (47-128)	199/90	-	-	M
Ziol 2009 [81]	France	SC, P, CS	13	49.3 (39-60)	29.4 (23.8-34.6)	-	10/3	2003-2005	-	-
Abbreviations: BMI - body mass index; WC – waist circumference; M – males; F – females; MC – multi-centre, SC – single-centre, P – prospective, R – retrospective, CS – cross-sectional, CC – case-control; - Data not available										

Table 2. Demographic details of the entire cohort, and patients without (F0-2) and with (F3-4) advanced fibrosis.

	Entire cohort (N = 5735)	**F0-2 (N = 4013)**	**F3-4 (N = 1722)**
Females (%)	45	43	48
BMI ≥ 30 kg/m^2 (%)	43	45	53
Waist circumference (cm)	103 (15)	102 (15)	106 (14)
Diabetes (%)	33	30	58
Age (years)*	54 (19)	50 (19)	59 (14)
BMI (kg/m^2)*	30 (7)	29 (8)	30 (7)
Biopsy data			
Steatosis S0/S1/S2/S3 (%)	3/35/36/26	3/36/36/25	2/32/38/28
Ballooning B0/B1/B2 (%)	24/47/29	30/49/21	10/45/46
Inflammation I0/I1/I2/I3 (%)	13/60/24/3	17/62/20/1	5/55/34/6
NAS score+	4 (2)	4 (2)	5 (1)
NASH (%)	50	43	67
Liver function tests			
ALT (IU/L)*	55 (48)	53 (48)	60 (48)
AST (IU/L)*	40 (30)	36 (25)	50 (34)
Platelets (×10^9/l) +	230 (72)	241 (67)	205 (75)
Albumin (g/l) +	43 (9)	43 (7)	43 (13)
GGT (IU/L)*	69 (87)	62 (78)	87 (102)
NITs			
LSM (kPa)*	10.7 (6.1)	6.7 (3.5)	13.3 (12.0)
FIB-4*	1.7 (1.2)	1.1 (0.9)	1.9 (1.7)
NFS+	-1.5 (1.7)	-1.9 (1.6)	-0.6 (1.8)
APRI*	0.6 (0.4)	0.4 (0.3)	0.6 (0.6)
AST/ALT*	0.8 (0.4)	0.7 (0.4)	0.8 (0.5)

*Data are reported as median (IQR).

+Data are reported as mean (SD).

Validating the diagnostic performance of LSM by VCTE and serum-based tests for detecting advanced fibrosis

LSM-VCTE, FIB-4, NFS, APRI, and AST/ALT had corresponding AUROCs of 0.85, 0.76, 0.73, 0.70, 0.64 for identifying advanced fibrosis (Table 3), and 0.90, 0.80, 0.78, 0.72, 0.69; for the identification of cirrhosis (Supporting Table 6). LSM-VCTE performed significantly better ($p<10^{-15}$) in detecting both advanced fibrosis and cirrhosis than all serum-based tests. This relationship was preserved when performing a head-to-head comparison of LSM-VCTE, FIB-4 and NFS in the same group of patients (Supporting Tables 7 and 8).

When considering cut-offs from the literature, we evaluated lower and higher cut-offs separately. For any given test, as would be expected, low thresholds yielded higher sensitivity and high thresholds were associated with higher specificity (Supporting Table 9). Indicative PPV and NPV are also provided for the range of prevalences (5% - 50%) reported in the primary studies (Supporting Tables 10-14).

APRI and AST/ALT ratio had only modest diagnostic performance for advanced fibrosis (AUROC≤0.70, Table 3), and were therefore not considered further.

None of the thresholds regarded in isolation resulted in both a high sensitivity (≥80%) and high specificity (≥80%) (Figure 2, Table 2, Supporting Tables 9 and 15, and Supporting Figure 4). Therefore, we explored the use of a lower and an upper cut-off. LSM-VCTE literature cut-offs performed well in only two cases (<7.1 kPa and ≥14.1 kPa: 83% sensitivity, 90% specificity; and <7.9 kPa and ≥9.6 kPa: 84% sensitivity, 78% specificity), while for other LSM-VCTE, NFS and FIB-4 thresholds a high specificity was observed (FIB-4: 91% for <1.3 & ≥2.67, 95% for <1.3, ≥3.25) but sensitivity was <60% (Table 4). In addition, the proportion of indeterminate cases was >30% for serum-based NITs. Threshold pairs derived from the IPD study group did not reduce the proportion of misclassified and indeterminate patients seen with literature-based threshold pairs (Table 4).

We further evaluated the performance of LSM-VCTE, FIB-4 and NFS to diagnose advanced fibrosis in sequential combinations of serum-based NITs and LSM-VCTE. When selecting threshold combinations for FIB-4 and NFS available in the literature (<1.3 & ≥2.67, <1.3 & ≥3.25 for FIB-4; <-1.455 & ≥0.676 for NFS) and pairing them with the best threshold pair for LSM-VCTE (<7.9 kPa & ≥9.6 kPa, identified as the one with highest sensitivity and lowest indeterminate proportion), the proportion of patients in the indeterminate group was 5%. While both the FIB-4+LSM-VCTE and NFS+LSM-VCTE sequential combinations had specificity >80%, their sensitivity was ≤80% (Table 5). A better sensitivity was reached by using thresholds derived from the IPD study group (<0.88 & ≥2.31 for FIB-4; <-2.55 & ≥0.28 for NFS), but the proportion of indeterminate cases was near 20% in those cases and the proportions of patients needing LSM-VCTE was also larger than when using literature cut-offs (Table 5).

Table 3. Diagnostic performance of non-invasive tests for advanced fibrosis (F3-F4).

	LSM by VCTE (n = 5489)			FIB-4 (n = 5393)			NFS (n = 3248)			APRI (n = 5477)			AST/ALT (n = 5434)		
Advanced fibrosis, %	30			30			29			30			30		
AUC	0.85 (0.84-0.86)			0.76 (0.74-0.77)			0.73 (0.71-0.75)			0.70 (0.69-0.72)			0.64 (0.62-0.65)		
	YI	90% Se	90% Sp	YI	90% Se	90% Sp	YI	90% Se	90% Sp	YI	90% Se	90% Sp	YI	90% Se	90% Sp
Threshold	9.1	7.4	12.1	1.44	0.88	2.31	-1.39	-2.55	0.28	0.49	0.29	0.91	0.64	0.51	1.34
Sensitivity, %	77 (75-79)	90 (89-91)	55 (52-57)	69 (67-72)	90 (88-91)	38 (36-41)	75 (72-78)	90 (88-92)	29 (26-32)	67 (64-69)	90 (89-92)	32 (30-34)	775 (73-77)	90 (87-91)	16 (14-18)
Specificity, %	78 (76-79)	60 (59-61)	90 (89-91)	70 (69-72)	39 (37-40)	90 (89-91)	63 (61-65)	36 (33-37)	90 (89-91)	63 (62-65)	29 (28-30)	90 (89-91)	47 (45-48)	25 (23-26)	90 (89-91)
Misclassified, %	22 (22-23)	31 (31-32)	21 (20-21)	30 (30-31)	46 (46-47)	26 (25-26)	34 (34-36)	48 (49-50)	28 (28-29)	36 (36-37)	53 (53-54)	27 (27-28)	45 (45-46)	56 (56-57)	32 (32-33)

For each non-invasive test thresholds were selected according to Youden's index (YI), and fixed at 90% sensitivity (90% Se) and 90% specificity (90% Sp). 95% confidence intervals were estimated with 500 bootstrap replicates.

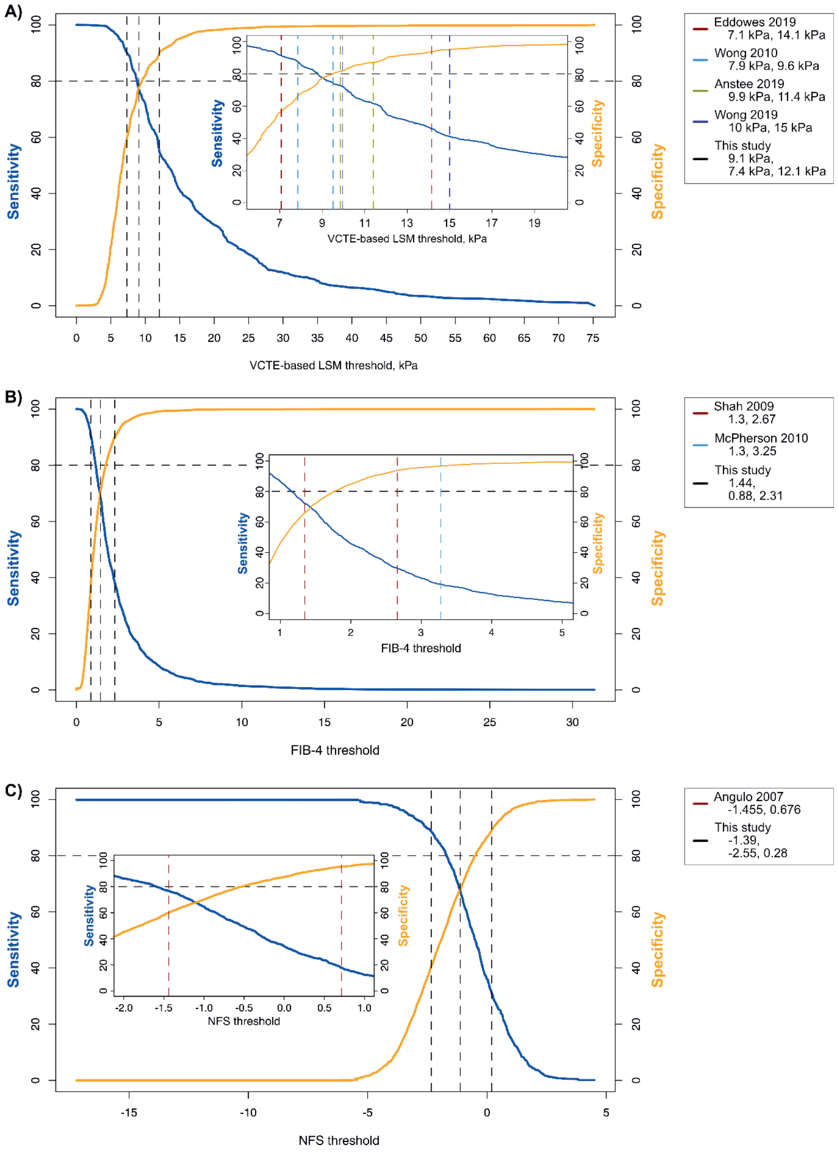


Figure 2. Distribution of sensitivities and specificities over the possible threshold ranges for LSM by VCTE (A), FIB-4 (B) and NFS (C) when considering the diagnosis of advanced fibrosis. Insets show the distribution of cut-offs identified from the literature. Horizontal dashed lines are representing the minimum acceptable criteria for considering a test as having high sensitivity (≥80%) and high specificity (≥80%).

Algorithms to minimise the need for liver biopsy

In the derivation set, the cut-offs for 95% and 98% specificity for the diagnosis of cirrhosis were respectively 20.4 kPa and 27.6 kPa for LSM-VCTE, 3.48 and 4.63 for FIB-4 and 1.01 and 1.57 for NFS. These cut-offs performed similarly in the validation set (Supporting Tables 16 and 17).

Algorithms combining FIB-4 (lower cut-off of 1.3 as described in the literature and upper cut-offs of 3.48 and 4.63 as described above) and LSM by VCTE (lower cut-off rounded to 8.0 kPa and upper cut-offs rounded to 20.0 kPa and 28.0 kPa, as described above) were then compared to the traditional way of applying these tests, also with rounded cut-offs for LSM by VCTE (8 kPa and 10 kPa) (Figure 3). This approach increased the number of patients requiring a liver stiffness measurement (from 34% to 40% and 44%) but decreased the number of patients needing liver biopsy (from 33% to 19% and 24% when using the 95% and 98% specificity cut-offs, respectively) (Supporting Table 18 and Figure 3).

Subgroup and sensitivity analyses

In subgroup analysis for the diagnosis of advanced fibrosis (Supporting Table 19), NITs performed better in patients with lower BMI (AUROCs LSM-VCTE: 0.91, $p<0.005$; FIB-4: 0.81, $p<0.001$; NFS: 0.76, $p<0.025$), without T2DM (LSM-VCTE: 0.87, $p<10^{-6}$; FIB-4: 0.77, $p<0.01$), and with biopsies shorter than 20mm (LSM-VCTE: 0.87, $p<0.005$; FIB-4: 0.80, $p<0.001$; NFS: 0.79, $p<0.05$), or with fewer than 11 portal tracts (LSM-VCTE: 0.86, $p=0.01$; FIB-4: 0.79, $p=0.04$; NFS: 0.78, $p<0.005$). Diagnostic performance was also lower in patients in the youngest age quartile (<43 years, AUROC: 0.58, $p<0.001$) and in females (AUROC: 0.71, $p=0.03$) for NFS, while continent of provenance did not have a significant effect for any NITs. In patients with normal levels of ALT (ALT<40) FIB-4 performed worse (AUROC: 0.73) than in patients with ALT≥40 and ALT<100 (AUROC: 0.77, $p<0.01$). NFS performed better in patients with AST<40 (AUROC: 0.76), then in patients with AST≥100 (AUROC: 0.65, $p<0.01$). FIB-4 performed better in patients with at least one abnormal aminotransferase measurement (AUROC: 0.72, $p=0.014$). For cirrhosis, the trends were similar, except that for the diagnosis of cirrhosis, LSM by VCTE performed better in the youngest age group (AUROC: 0.97, $p<10^{-4}$) and NIT diagnostic performance was independent of aminotransferase levels (Supporting Table 20).

Table 4. Diagnostic accuracy of pairs of cut-offs from the literature for LSM by VCTE, FIB-4 and NFS for diagnosing advanced fibrosis.

	LSM by VCTE (n = 5489)				**FIB-4 (n = 5393)**			**NFS (n = 3248)**		
Advanced fibrosis, %	30				30			29		
AUROC	0.85 (0.84-0.86)				0.76 (0.74-0.77)			0.73 (0.71-0.75)		
Source	Anstee 2019 [29]	Eddowes 2019 [30]	Wong 2019 [75]	Wong 2010 [20]	This study	Shah 2009 [82]	McPherson 2010 [83]	This study	Angulo 2007 [16]	This study
Thresholds	<9.9, ≥11.4	<7.1, ≥14.1	<10, ≥15	<7.9, ≥9.6	<7.4, ≥12.1*	<1.3, ≥2.67	<1.3, ≥3.25	<0.88, ≥2.31*	<-1.455, ≥0.676	<-2.55, ≥0.28*
Sensitivity, %	69 (67-71)	83 (80-86)	59 (57-61)	84 (82-87)	84 (81-87)	54 (52-56)	44 (42-46)	80 (76-83)	47 (44-50)	74 (70-79)
Specificity, %	86 (85-88)	90 (88-92)	94 (93-96)	78 (76-80)	87 (85-88)	91 (89-92)	95 (93-96)	79 (77-81)	91 (89-93)	78 (76-81)
Misclassified, %	17 (16-19)	7 (6-8)	12 (11-13)	17 (16-19)	10 (9-11)	12 (11-13)	10 (9-11)	10 (9-11)	11 (10-13)	10 (8-11)
Indeterminate, %	7 (6-8)	39 (37-40)	18 (17-19)	13 (12-14)	31 30-33)	34 (33-35)	39 (37-40)	52 (50-53)	39 (37-41)	56 (54-59)

*Cut-offs determined from the IPD study group. Lower cut-offs correspond to a lower limit of 90% sensitivity, upper cut-offs correspond to a lower limit of 90% specificity. 95% confidence intervals were determined with 500 bootstrap replicates.

Table 5. Diagnostic performance of combinations of NFS and LSM by VCTE, and FIB-4 and LSM by VCTE tests to diagnose patients with advanced fibrosis.

	FIB-4 & LSM by VCTE (n = 5159)	NFS & LSM by VCTE (n = 3094)	FIB-4 & LSM by VCTE (n = 5159)	NFS & LSM by VCTE (n = 3094)	FIB-4 & LSM by VCTE (n = 5159)	NFS & LSM by VCTE (n = 3094)
Advanced fibrosis, %	30	28	30	28	30	28
Thresholds for blood-based NIT	<0.88, ≥2.31*	<-2.55, ≥0.28*	<1.3, ≥2.67^{+}	<-1.455, ≥0.676^{+}	<1.3, ≥2.67^{+}	<-1.455, ≥0.676^{+}
Thresholds for LSM by VCTE, kPa	<7.4, ≥12.1*	<7.4, ≥12.1*	<7.9, ≥9.6^{+}	<7.9, ≥9.6^{+}	<8.0, ≥10.0^{+}	<8.0, ≥10.0^{+}
Sensitivity, %	80 (77-83)	77 (74-81)	67 (64-69)	65 (62-68)	66 (63-68)	64 (62-67)
Specificity, %	81 (79-83)	83 (81-85)	85 (84-87)	86 (84-88)	86 (84-87)	86 (84-88)
PPV, %	62 (60-65)	61 (58-64)	66 (64-68)	63 (61-67)	66 (64-68)	64 (61-67)
NPV, %	91 (90-92)	91 (89-93)	86 (85-87)	87 (85-88)	86 (85-87)	86 (85-88)
Indeterminate, %	18 (17-19)	20 (18-21)	5 (4-5)	5 (5-6)	5 (4-6)	5 (5-6)
Misclassification, %	16 (14-17)	15 (13-17)	19 (18-21)	19 (17-21)	19 (18-20)	19 (17-21)
Patients undergoing LSM by VCTE, %	51 (50-53)	56 (54-59)	34 (32-35)	38 (36-40)	34 (33-35)	38 (37-40)

95% confidence intervals were estimated with 500 bootstrap replicates.

*Thresholds were determined from the IPD study group as corresponding to 90% sensitivity (lower value) and 90% specificity (upper value)

$^{+}$Threshold were determined from the literature. For LSM by VCTE, a threshold pair yielding the highest sensitivity and specificity while having the smallest proportion of indeterminate cases in diagnosing advanced fibrosis was chosen.

The diagnostic performance of LSM-VCTE was significantly lower in patients with unreliable liver stiffness measurements ($p<10^{-8}$; both for advanced fibrosis and cirrhosis) when applying the Boursier-criteria [22], but not when only considering IQR/median LSM<0.30. The proportion of unreliable results was 12% both in the advanced fibrosis and cirrhosis groups (Supporting Table 21).

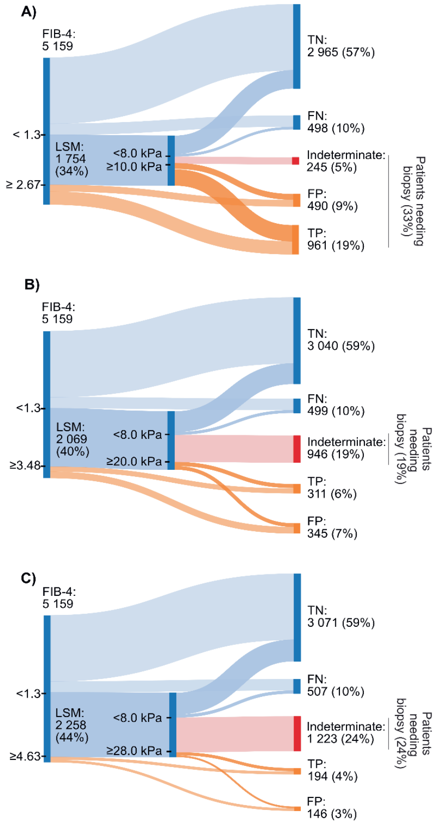
There was no difference in the diagnostic performance of LSM-VCTE between the M and XL probes in the subgroup of patients who had undergone LSM by both probes (Supporting Table 22).

In a sensitivity analysis of patients with LSM matched to BMI (only M probe measurements if BMI<30 kg/m^2 and only XL probe measurements if BMI≥30 kg/m^2), there was no significant difference between the diagnostic performance of LSM-VCTE when comparing to the entire IPD study group (Supporting Table 23).

Discussion

Through an extensive collaboration network with authors of primary studies we were able to collect the largest dataset of its kind ever to be reported on. This includes a diverse set of study groups from Europe, Asia, and Australia, 30% of whom had advanced fibrosis. We believe that our findings are therefore relevant for patients typical of secondary care in these territories and may be applied in the development of new strategies or in the consolidation of existing practices in evaluating patients for referral to secondary care.

A few studies evaluated the diagnostic performance of LSM-VCTE and other NITs, but most report on fewer than 500 patients. One similarly large study reported on patients screened for inclusion in clinical trials, where the prevalence of advanced fibrosis was 71% [29], making it difficult to make generalisations about its applicability in routine practice or compare its results to ours. A smaller study with 1073 NAFLD patients of whom 29% had advanced fibrosis (ref) examined the diagnostic performance of LSM by VCTE. The authors of that study reported AUC and specificity values similar to our findings, however they reported increased sensitivity. Other smaller studies reported similar prevalence of advanced fibrosis and similar AUROCs for LSM-VCTE [30–33].



		FIB-4	LSM
< Lower cutoff	TN	2311 (44%)	654 (13%)
	FN	396 (8%)	102 (2%)
≥ Upper cutoff	TP	459 (9%)	502 (10%)
	FP	239 (4%)	251 (5%)

		FIB-4	LSM
< Lower cutoff	TN	2311 (44%)	729 (14%)
	FN	396 (8%)	103 (2%)
≥ Upper cutoff	TP	169 (3%)	142 (3%)
	FP	214 (4%)	131 (3%)

		FIB-4	LSM
< Lower cutoff	TN	2311 (44%)	760 (15%)
	FN	396 (8%)	111 (2%)
≥ Upper cutoff	TP	102 (2%)	92 (2%)
	FP	92 (2%)	54 (1%)

Figure 3. Sankey diagrams showing the distribution of patients in true positive, true negative, false positive, false negative and indeterminate groups for a sequential combination of FIB-4 and LSM by VCTE when using different thresholds for each testing tier. A lower threshold was used to rule out patients without advanced fibrosis and an upper threshold ruled in patients with advanced fibrosis when applying both tests (A). In an alternative model a lower threshold was used to rule out patients without advanced fibrosis, but the upper threshold ruled in only patients with cirrhosis (B, C). Two different pairs of thresholds were chosen for this hybrid strategy: the lower cut-off for both FIB-4 and LSM by VCTE were determined from the literature; upper cut-offs were both determined as corresponding to 95% specificity in detecting cirrhosis (B) or both corresponding to 98% specificity in detecting cirrhosis (C). In the application of the algorithm described in (A) 33% of patients would need to have a liver biopsy for the diagnosis of cirrhosis (those in the indeterminate group to rule out advanced fibrosis and those in the rule in group to identify cirrhosis). With the application of an upper cut-off to rule in cirrhosis without the need of biopsy, only patients in the indeterminate group need to have a biopsy. The latter strategy results in fewer patients undergoing biopsy (18% and 24% depending on the threshold used).

Overall, the diagnostic performance of LSM-VCTE for advanced fibrosis was good (AUROC=0.85), while that of FIB-4 and NFS in the same group was moderate (AUROCs=0.76 for FIB-4, AUROC=0.73 for NFS). None of the studied NITs had both sufficiently high sensitivity and specificity (≥ 80%) when used with single cut-offs. Diagnostic performance was higher for detecting cirrhosis, as reported in previous studies [30,34,35]. LSM-VCTE had the highest sensitivity and specificity, both in the case of a single cut-off (9.1 kPa obtained by maximising the Youden index; 77% and 78%) and for two cut-offs (<7.4 kPa & ≥12.1 kPa; 84% and 87%). Of the LSM-VCTE cut-off pairs tested, <7.1 kPa and ≥14.1 kPa, first published by Eddowes et al. in 2019 [30], performed well for advanced fibrosis, with sensitivity of 83% and specificity of 90%, but with a proportion of 39% of patients ending up with an indeterminate result, similar to 41% indeterminate patients reported in the original paper [30].

LSM-VCTE thresholds identified in our study group (<9.1 kPa; <7.4 kPa & ≥12.1 kPa) were similar to thresholds reported in the literature (<9.9 kPa; <7.1 kPa & ≥14.1 kPa, <7.9 kPa & ≥9.6 kPa). However, thresholds for FIB-4 (<1.44; <0.88 & ≥2.31) and NFS (<-1.39; <-2.55 & ≥0.28) defined in our IPD study group spanned a wider range than those reported in the literature (<1.3 & ≥2.67 or <1.3 & ≥3.25 for FIB-4; <-1.455 & ≥0.676 for NFS).

Our findings are in line with the existing literature suggesting that sequential combinations of NITs increase sensitivity and specificity [29]. Additionally, we have found NFS+LSM-VCTE and FIB-4+LSM-VCTE combinations to have similar sensitivity and specificity as recently reported by Boursier et al. [36]. Such combined testing strategies can reduce the number of indeterminate cases and reduce the costs associated with liver biopsies.

Furthermore, we propose an approach that could minimise the need for liver biopsies further, by using upper cut-offs with 95% and 98% specificity for the identification of cirrhosis. The rationale for this approach is explained in the Supporting Discussion. When using the 95% specificity cut-off, the proportion of patients needing liver biopsy decreases from 33% to 19% (Figure 3). However, in this approach, 345 of 656 patients “ruled-in” as having cirrhosis do not have histologically diagnosed cirrhosis. While this may seem like a high proportion of patients with false positive results, this must be interpreted in the light of two factors. First, the limitations of liver biopsy could mean that these patients are falsely classified as not having cirrhosis histologically. Furthermore, patients without cirrhosis on histology and with high NIT values could have equivalent risks as patients with cirrhosis on histology. For example, it is known from the hepatitis C literature [37] that patients without cirrhosis on liver biopsy but with a high FIB-4 (>3.25) still had a significant risk of developing HCC after hepatitis C treatment, demonstrating that NITs can have added benefit beyond the histological diagnosis of cirrhosis alone. The rate of false positive results for cirrhosis can be decreased by choosing cut-offs with higher specificity, but this will come at the expense of doing more biopsies. Despite this encouraging result, this is an area where more information is needed, particularly longitudinal data comparing the prognostic value of LSM-VCTE and other NITs against histology, and ultimately, the cost effectiveness of the various cut-offs would need to be evaluated.

Surprisingly, subgroup analyses showed that the diagnostic accuracy of NITs was better in cases with poor biopsy quality. This finding is difficult to explain but a similar observation was reported previously in a large group of patients screened for clinical trials [29]. The use of local biopsy reports as reference standard and the well-known observer-dependent variability of biopsy interpretation, even among expert pathologists [7], are factors that may have contributed to our finding. Spectrum bias was excluded as a source of this finding due to a near-identical proportion of patients in both the advanced fibrosis and cirrhosis group having short biopsies (Supporting Table 5).

Subgroup analysis showed better diagnostic performance of NITs in patients with lower BMI [38,39], and patients without diabetes, in keeping with other studies [40,41]. This effect is likely to be primarily driven by BMI as there is thought to be a causal association between BMI and T2DM. NIT performance was impacted by age, with all NITs performing worse in the younger quartile of our study group for advanced fibrosis, but the trend was reversed for cirrhosis where NITs performed better in those younger than 43 years of age. The age dependence of FIB-4 and NFS is expected, as age is one of the parameters included in the algorithms, and has indeed been previously described [13,42]. It is, however, difficult to explain why performance of NITs is better in the younger age group for the diagnosis of cirrhosis.

Our study has several strengths, including the large size of the IPD study group and composition with prevalence of advanced fibrosis of 30%, which makes it relevant to routine practice. Furthermore, the proportion of unreliable VCTE measurements in our study was 12%, in keeping with the literature [22]. However, we acknowledge some limitations. We did not have any data from the USA and very few studies from Australia, so the results could not be globally applicable, due to differences in BMI across study populations. In addition, due to the nature of our study, we had to use the locally provided histology results possibly introducing bias. Furthermore, we covered a large chronological period, during which LSM-VCTE application underwent significant changes, initially with the introduction of the XL probe, followed by the advice to measure SCD and the introduction of the Automatic Probe Selection tool. There was therefore some heterogeneity in the performance of LSM-VCTE, with early studies using only the M probe to assess all patients, while only a subset of studies assessed SCD to guide probe selection. Furthermore, one third of the included studies was carried out in France, as the technology used for LSM by VCTE originates from there.

Lastly, our data confirm that LSM-VCTE had superior accuracy to serum-based tests, and this is independent of probe type, sex, ALT, AST, and participants' continent of origin. There was, however, some dependence on the presence of T2DM, BMI and for the detection of cirrhosis, and we did not check for subgroup-specific cut-offs, but these should be explored in future studies.

Our study examined some of the most widely available NITs. While it cannot be considered exhaustive, it can be regarded as the benchmark against which newer NITs can be tested. This is particularly important as new tests are continuously being developed (FibroTest-FibroSURE, ActiTest [43], ELF [44]). Furthermore, newer tests are also needed for patients with "at risk" NASH (NASH+F2-3) who would be candidates for clinical trials or treatments, once approved therapies become available (FAST score [45], NIS4 [46], cTAG [47]).

In conclusion, our study provides further validation of the use of sequential combination of FIB-4 and LSM-VCTE to rule out patients with NAFLD and advanced fibrosis who can be managed in primary care. We have shown how the use of upper cut-offs to rule in cirrhosis in combination with lower cut-offs to rule out advanced fibrosis can lead to a reduction in the number of patients who would need to undergo liver biopsy.

References

1 Younossi Z, Anstee QM, Marietti M, *et al.* Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. *Nat Rev Gastroenterol Hepatol* 2017;15:11–20. doi:10.1038/nrgastro.2017.109

2 Loomba R, Wong R, Fraysse J, *et al.* Nonalcoholic fatty liver disease progression rates to cirrhosis and progression of cirrhosis to decompensation and mortality: a real world analysis of Medicare data. *Aliment Pharmacol Ther* 2020;51:1149–59. doi:https://doi.org/10.1111/apt.15679

3 European Association for the Study of the Liver (EASL) EA for the S of D (EASD), EA for the S of O (EASO), European Association for the Study of Diabetes (EASD), European Association for the Study of Obesity (EASO). EASL-EASD-EASO Clinical Practice Guidelines for the Management of Non-Alcoholic Fatty Liver Disease. *Obes Facts* 2016;9:65–90. doi:10.1159/000443344

4 Thampanitchawong P, Piratvisuth T. Liver biopsy: complications and risk factors. *World J Gastroenterol* 1999;5:301–4. doi:10.3748/WJG.V5.I4.301

5 Ratziu V, Charlotte F, Heurtier A, *et al.* Sampling variability of liver biopsy in nonalcoholic fatty liver disease. *Gastroenterology* 2005;128:1898–906.http://www.ncbi.nlm.nih.gov/pubmed/15940625 (accessed 15 Feb 2018).

6 Standish RA, Cholongitas E, Dhillon A, *et al.* An appraisal of the histopathological assessment of liver fibrosis. *Gut* 2006;55:569–78. doi:10.1136/gut.2005.084475

7 Davison BA, Harrison SA, Cotter G, *et al.* Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. *J Hepatol* 2020;73:1322–32. doi:10.1016/j.jhep.2020.06.025

8 Ekstedt M, Franzén LE, Mathiesen UL, *et al.* Long-term follow-up of patients with NAFLD and elevated liver enzymes. *Hepatology* Published Online First: 2006. doi:10.1002/hep.21327

9 Ekstedt M, Hagström 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;61:1547–54. doi:10.1002/hep.27368

10 Taylor RS, Taylor RJ, Bayliss S, *et al.* Association Between Fibrosis Stage and Outcomes of Patients With Nonalcoholic Fatty Liver Disease: A Systematic Review and Meta-Analysis. *Gastroenterology* 2020;158:1611-1625.e12. doi:10.1053/j.gastro.2020.01.043

11 Srivastava A, Gailer R, Tanwar S, *et al.* Prospective evaluation of a primary care referral pathway for patients with non-alcoholic fatty liver disease. *J Hepatol* 2019;71:371–8. doi:10.1016/j.jhep.2019.03.033

12 Moolla A, Motohashi K, Marjot T, *et al.* A multidisciplinary approach to the management of NAFLD is associated with improvement in markers of liver and cardio-metabolic health. *Frontline Gastroenterol* 2019;10:337 LP – 346. doi:10.1136/flgastro-2018-101155

13 Davyduke T, Tandon P, Al-Karaghouli M, *et al.* Impact of Implementing a “FIB-4 First” Strategy on a Pathway for Patients With NAFLD Referred From Primary Care. *Hepatol Commun* 2019;3:1322–33. doi:10.1002/hep4.1411

14 Marchesini G, Day CP, Dufour JF, *et al.* EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease. *J Hepatol* 2016;64:1388–402. doi:10.1016/j.jhep.2015.11.004

15 Stewart LA, Clarke M, Rovers M, *et al.* Preferred Reporting Items for a Systematic Review and Meta-analysis of Individual Participant Data. *JAMA* 2015;313:1657. doi:10.1001/jama.2015.3656

16 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;45:846–54. doi:10.1002/hep.21496

17 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;43:1317–25. doi:10.1002/hep.21178

18 Sheth SG, Flamm SL, Gordon FD, *et al.* AST/ALT Ratio Predicts Cirrhosis in Patients With Chronic Hepatitis C Virus Infection. *Off J Am Coll Gastroenterol | ACG* 1998;93.https://journals.lww.com/ajg/Fulltext/1998/01000/AST_ALT_Ratio_Predicts_Cirrhosis_in_Patients_With.12.aspx

19 Lin Z-H, Xin Y-N, Dong Q-J, *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;53:726–36. doi:10.1002/hep.24105

20 Wong VWS, Vergniol J, Wong GLH, *et al.* Diagnosis of fibrosis and cirrhosis using liver stiffness measurement in nonalcoholic fatty liver disease. *Hepatology* 2010;51:454–62. doi:10.1002/hep.23312

21 Kleiner DE, Brunt EM, Van Natta M, *et al.* Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* 2005;41:1313–21. doi:10.1002/hep.20701

22 Boursier J, Zarski JP, de Ledinghen V, *et al.* Determination of reliability criteria for liver stiffness evaluation by transient elastography. *Hepatology* 2013;57:1182–91. doi:10.1002/hep.25993

23 Whiting PF, Rutjes AWS, Westwood ME, *et al.* QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med* 2011;155:529. doi:10.7326/0003-4819-155-8-201110180-00009

24 Team RC. R: A language and environment for statistical computing. 2020.

25 Robin X, Turck N, Hainard A, *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;12:77. doi:10.1186/1471-2105-12-77

26 Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. R package version 1.3-24. 2019.

27 Davison AC, Hinkley D V. *Bootstrap Methods and their Application*. Cambridge University Press 1997. doi:10.1017/cbo9780511802843

28 Petta S, Wong VW-S, Cammà C, *et al.* Improved noninvasive prediction of liver fibrosis by liver stiffness measurement in patients with nonalcoholic fatty liver disease accounting for controlled attenuation parameter values. *Hepatology* 2017;65:1145–55. doi:10.1002/hep.28843

29 Anstee QM, Lawitz EJ, Alkhouri N, *et al.* Noninvasive Tests Accurately Identify Advanced Fibrosis due to NASH: Baseline Data From the STELLAR Trials. *Hepatology* 2019;70:1521–30. doi:10.1002/hep.30842

30 Eddowes PJ, Sasso M, Allison M, *et al.* Accuracy of FibroScan Controlled Attenuation Parameter and Liver Stiffness Measurement in Assessing Steatosis and Fibrosis in Patients With Nonalcoholic Fatty Liver Disease. *Gastroenterology* 2019;156:1717–30. doi:10.1053/j.gastro.2019.01.042

31 Inadomi C, Takahashi H, Ogawa Y, *et al.* Accuracy of the Enhanced Liver Fibrosis test, and combination of the Enhanced Liver Fibrosis and non-invasive tests for the diagnosis of advanced liver fibrosis in patients with non-alcoholic fatty liver disease. *Hepatol Res* 2020;50:682–92. doi:10.1111/hepr.13495

32 Siddiqui MS, Vuppalanchi R, Van Natta ML, *et al.* Vibration-controlled Transient Elastography to Assess Fibrosis and Steatosis in Patients With Nonalcoholic Fatty Liver Disease. doi:10.1016/j.cgh.2018.04.043

33 Hsu C, Caussy C, Imajo K, *et al.* Magnetic Resonance vs Transient Elastography Analysis of Patients With Nonalcoholic Fatty Liver Disease: A Systematic Review and Pooled Analysis of Individual Participants. *Clin Gastroenterol Hepatol* 2019;17:630-637.e8. doi:10.1016/j.cgh.2018.05.059

34 Chen J, Yin M, Talwalkar JA, *et al.* Diagnostic Performance of MR Elastography and Vibration-controlled Transient Elastography in the Detection of Hepatic Fibrosis in Patients with Severe to Morbid Obesity. *Radiology* 2017;283:418–28. doi:10.1148/radiol.2016160685

35 Myers RP, Pomier-Layrargues G, Kirsch R, *et al.* Feasibility and diagnostic performance of the FibroScan XL probe for liver stiffness measurement in overweight and obese patients. *Hepatology* 2012;55:199–208. doi:10.1002/hep.24624

36 Boursier J, Guillaume M, Leroy V, *et al.* New sequential combinations of non-invasive fibrosis tests provide an accurate diagnosis of advanced fibrosis in NAFLD. *J Hepatol* 2019;71:389–96. doi:10.1016/j.jhep.2019.04.020

37 Ioannou GN, Feld JJ. What Are the Benefits of a Sustained Virologic Response to Direct-Acting Antiviral Therapy for Hepatitis C Virus Infection? *Gastroenterology* 2019;156:446-460.e2. doi:https://doi.org/10.1053/j.gastro.2018.10.033

38 Petta S, Wai-Sun Wong V, Bugianesi E, *et al.* Impact of Obesity and Alanine Aminotransferase Levels on the Diagnostic Accuracy for Advanced Liver Fibrosis of Noninvasive Tools in Patients With Nonalcoholic Fatty Liver Disease. *Am J Gastroenterol* 2019;114:916–28. doi:10.14309/ajg.0000000000000153

39 Joo SK, Kim W, Kim D, *et al.* Steatosis severity affects the diagnostic performances of noninvasive fibrosis tests in nonalcoholic fatty liver disease. *Liver Int* 2018;38:331–41. doi:10.1111/liv.13549

40 Alkayyali T, Qutranji L, Kaya E, *et al.* Clinical utility of noninvasive scores in assessing advanced hepatic fibrosis in patients with type 2 diabetes mellitus: a study in biopsy-proven non-alcoholic fatty liver disease. *Acta Diabetol* 2020;57:613–8. doi:10.1007/s00592-019-01467-7

41 Eren F, Kaya E, Yilmaz Y. Accuracy of Fibrosis-4 index and non-alcoholic fatty liver disease fibrosis scores in metabolic (dysfunction) associated fatty liver disease according to body mass index: failure in the prediction of advanced fibrosis in lean and morbidly obese individual. *Eur J Gastroenterol Hepatol* 9000;Publish Ah.https://journals.lww.com/eurojgh/Fulltext/9000/Accuracy_of_Fibrosis_4_index_and_non_alcoholic.97415.aspx

42 McPherson S, Hardy T, Dufour J-F, *et al.* Age as a Confounding Factor for the Accurate Non-Invasive Diagnosis of Advanced NAFLD Fibrosis. *Am J Gastroenterol* 2017;112:740–51. doi:10.1038/ajg.2016.453

43 Ratziu V, Massard J, Charlotte F, *et al.* Diagnostic value of biochemical markers (FibroTest-FibroSURE) for the prediction of liver fibrosis in patients with non-alcoholic fatty liver disease. *BMC Gastroenterol* 2006;6:6. doi:10.1186/1471-230X-6-6

44 Lichtinghagen R, Pietsch D, Bantel H, *et al.* The Enhanced Liver Fibrosis (ELF) score: normal values, influence factors and proposed cut-off values. *J Hepatol* 2013;59:236–42. doi:10.1016/j.jhep.2013.03.016

45 Newsome PN, Sasso M, Deeks JJ, *et al.* FibroScan-AST (FAST) score for the non-invasive identification of patients with non-alcoholic steatohepatitis with significant activity and fibrosis: a prospective derivation and global validation study. *lancet Gastroenterol Hepatol* 2020;5:362–73. doi:10.1016/S2468-1253(19)30383-8

46 Harrison SA, Ratziu V, Boursier J, *et al.* A blood-based biomarker panel (NIS4) for non-invasive diagnosis of non-alcoholic steatohepatitis and liver fibrosis: a prospective derivation and global validation study. *Lancet Gastroenterol Hepatol* 2020;5:970–85. doi:https://doi.org/10.1016/S2468-1253(20)30252-1

47 Dennis A, Mouchti S, Kelly M, *et al.* A composite biomarker using multiparametric magnetic resonance imaging and blood analytes accurately identifies patients with non-alcoholic steatohepatitis and significant fibrosis. *Sci Rep* 2020;10:1–11. doi:10.1038/s41598-020-71995-8

48 Agrawal S, Hoad CL, Francis ST, *et al.* Visual morphometry and three non-invasive markers in the evaluation of liver fibrosis in chronic liver disease. *Scand J Gastroenterol* 2017;52:107–15. doi:10.1080/00365521.2016.1233578

49 Aykut UE, Akyuz U, Yesil A, *et al.* A comparison of FibroMeter™ NAFLD Score, NAFLD fibrosis score, and transient elastography as noninvasive diagnostic tools for hepatic fibrosis in patients with biopsy-proven non-alcoholic fatty liver disease. *Scand J Gastroenterol* 2014;49:1343–8. doi:10.3109/00365521.2014.958099

50 Boursier J, Vergniol J, Guillet A, *et al.* Diagnostic accuracy and prognostic significance of blood fibrosis tests and liver stiffness measurement by FibroScan in non-alcoholic fatty liver disease. *J Hepatol* 2016;65:570–8. doi:10.1016/J.JHEP.2016.04.023

51 Boursier J, Lannes A, Oberti F, *et al.* The combination of Fibroscan with blood markers in the fibrometerVCTE significantly reduces the use of liver biopsy for the assessment of advanced fibrosis in non-alcoholic fatty liver disease. *J Hepatol* 2017;66.

52 Boursier J, Lannes A, Shili S, *et al.* The new fibrometervcte outperforms recommended liver fibrosis tests in NAFLD. *Hepatology* 2018;68.

53 Cassinotto C, Lapuyade B, Aït-Ali A, *et al.* Liver Fibrosis: Noninvasive Assessment with Acoustic Radiation Force Impulse Elastography—Comparison with FibroScan M and XL Probes and FibroTest in Patients with Chronic Liver Disease. *Radiology* 2013;269:283–92. doi:10.1148/radiol.13122208

54 Cassinotto C, Boursier J, de Lédinghen V, *et al.* Liver stiffness in nonalcoholic fatty liver disease: A comparison of supersonic shear imaging, FibroScan, and ARFI with liver biopsy. *Hepatology* 2016;63:1817–27. doi:10.1002/hep.28394

55 Chan W-K, Nik Mustapha NR, Mahadeva S. A novel 2-step approach combining the NAFLD fibrosis

score and liver stiffness measurement for predicting advanced fibrosis. *Hepatol Int* 2015;9:594–602. doi:10.1007/s12072-014-9596-7

56 Chan W-K, Nik Mustapha NR, Wong GL-H, *et al.* Controlled attenuation parameter using the FibroScan® XL probe for quantification of hepatic steatosis for non-alcoholic fatty liver disease in an Asian population. *United Eur Gastroenterol J* 2017;5:76–85. doi:10.1177/2050640616646528

57 Eddowes PJ, Newsome PN, Anstee Q, *et al.* Staging fibrosis and excluding advanced fibrosis in patients with NAFLD: Comparison of non-invasive markers in an interim analysis from a prospective multicentre study. *Hepatology* 2016;64.

58 Clet M, Miette V, Eddowes P, *et al.* Expanding the Use of the Vcte XL probe in morbid obese patients: Validation of a new automated adaptive measurement depths algorithm in a large UK multicenter cohort. *Hepatology* 2018;68.

59 Gaia S, Carenzi S, Barilli AL, *et al.* Reliability of transient elastography for the detection of fibrosis in Non-Alcoholic Fatty Liver Disease and chronic viral hepatitis. *J Hepatol* 2011;54:64–71. doi:10.1016/j.jhep.2010.06.022

60 Garg H, Aggarwal S, Shalimar, *et al.* Utility of transient elastography (fibroscan) and impact of bariatric surgery on nonalcoholic fatty liver disease (NAFLD) in morbidly obese patients. *Surg Obes Relat Dis* 2018;14:81–91. doi:10.1016/J.SOARD.2017.09.005

61 Karlas T, Dietrich A, Peter V, *et al.* Evaluation of Transient Elastography, Acoustic Radiation Force Impulse Imaging (ARFI), and Enhanced Liver Function (ELF) Score for Detection of Fibrosis in Morbidly Obese Patients. *PLoS One* 2015;10:e0141649. doi:10.1371/journal.pone.0141649

62 Labenz C, Huber Y, Kalliga E, *et al.* Predictors of advanced fibrosis in non-cirrhotic non-alcoholic fatty liver disease in Germany. *Aliment Pharmacol Ther* 2018;48:1109–16. doi:10.1111/apt.14976

63 Lee MS, Bae JM, Joo SK, *et al.* Prospective comparison among transient elastography, supersonic shear imaging, and ARFI imaging for predicting fibrosis in nonalcoholic fatty liver disease. *PLoS One* 2017;12:e0188321. doi:10.1371/journal.pone.0188321

64 Lupsor M, Badea R, Stefanescu H, *et al.* Performance of unidimensional transient elastography in staging non-alcoholic steatohepatitis. *J Gastrointest liver Dis* 2010;19:53–60.

65 Mahadeva S, Mahfudz AS, Vijayanathan A, *et al.* Performance of transient elastography (TE) and factors associated with discordance in nonalcoholic fatty liver disease. *J Dig Dis* 2013;14:n/a-n/a. doi:10.1111/1751-2980.12088

66 Okajima A, Sumida Y, Taketani H, *et al.* Liver stiffness measurement to platelet ratio index predicts the stage of liver fibrosis in non-alcoholic fatty liver disease. *Hepatol Res* 2017;47:721–30. doi:10.1111/hepr.12793

67 Ooi GJ, Earnest A, Kemp WW, *et al.* Evaluating feasibility and accuracy of non-invasive tests for nonalcoholic fatty liver disease in severe and morbid obesity. *Int J Obes* 2018;42:1900–11. doi:10.1038/s41366-018-0007-3

68 Pavlides M, Banerjee R, Tunnicliffe EM, *et al.* Multiparametric magnetic resonance imaging for the assessment of non-alcoholic fatty liver disease severity. *Liver Int* 2017;37:1065–73. doi:10.1111/liv.13284

69 Petta S, Maida M, Macaluso FS, *et al.* The severity of steatosis influences liver stiffness measurement in patients with nonalcoholic fatty liver disease. *Hepatology* 2015;62:1101–10. doi:10.1002/hep.27844

70 Petta S, Vanni E, Bugianesi E, *et al.* The combination of liver stiffness measurement and NAFLD fibrosis score improves the noninvasive diagnostic accuracy for severe liver fibrosis in patients with nonalcoholic fatty liver disease. *Liver Int* 2015;35:1566–73. doi:10.1111/liv.12584

71 Petta S, Wong VW-S, Cammà C, *et al.* Serial combination of non-invasive tools improves the diagnostic accuracy of severe liver fibrosis in patients with NAFLD. *Aliment Pharmacol Ther* 2017;46:617–27. doi:10.1111/apt.14219

72 Seki K, Shima T, Oya H, *et al.* Assessment of transient elastography in Japanese patients with non-alcoholic fatty liver disease. *Hepatol Res* 2017;47:882–9. doi:10.1111/hepr.12829

73 Shen F, Zheng R-D, Shi J-P, *et al.* Impact of skin capsular distance on the performance of controlled attenuation parameter in patients with chronic liver disease. *Liver Int* 2015;35:2392–400. doi:10.1111/liv.12809

74 Staufer K, Halilbasic E, Spindelboeck W, *et al.* Evaluation and comparison of six noninvasive tests for prediction of significant or advanced fibrosis in nonalcoholic fatty liver disease. *United Eur Gastroenterol J* 2019;7:1113–23. doi:10.1177/2050640619865133

75 Wong VWS, Irles M, Wong GLH, *et al.* Unified interpretation of liver stiffness measurement by M and XL probes in non-alcoholic fatty liver disease. *Gut* 2019;68:2057–64. doi:10.1136/gutjnl-2018-317334

76 Kwok R, Choi KC, Wong GL-H, *et al.* Screening diabetic patients for non-alcoholic fatty liver disease with controlled attenuation parameter and liver stiffness measurements: a prospective cohort study. *Gut* 2016;65:1359–68. doi:10.1136/gutjnl-2015-309265

77 Loong TC-W, Wei JL, Leung JC-F, *et al.* Application of the combined FibroMeter vibration-controlled transient elastography algorithm in Chinese patients with non-alcoholic fatty liver disease. *J Gastroenterol Hepatol* 2017;32:1363–9. doi:10.1111/jgh.13671

78 Wong VW-S, Vergniol J, Wong GL-H, *et al.* Liver Stiffness Measurement Using XL Probe in Patients With Nonalcoholic Fatty Liver Disease. *Am J Gastroenterol* 2012;107:1862–71. doi:10.1038/ajg.2012.331

79 Yoneda M, Yoneda M, Mawatari H, *et al.* Noninvasive assessment of liver fibrosis by measurement of stiffness in patients with nonalcoholic fatty liver disease (NAFLD). *Dig Liver Dis* 2008;40:371–8. doi:10.1016/J.DLD.2007.10.019

80 Younes R, Rosso C, Petta S, *et al.* Usefulness of the index of NASH - ION for the diagnosis of steatohepatitis in patients with non-alcoholic fatty liver: An external validation study. *Liver Int* 2018;38:715–23. doi:10.1111/liv.13612

81 Ziol M, Kettaneh A, Ganne-Carrié N, *et al.* Relationships between fibrosis amounts assessed by morphometry and liver stiffness measurements in chronic hepatitis or steatohepatitis. *Eur J Gastroenterol Hepatol* 2009;21:1261–8. doi:10.1097/MEG.0b013e32832a20f5

82 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;7:1104–12. doi:https://doi.org/10.1016/j.cgh.2009.05.033

83 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 LP – 1269. doi:10.1136/gut.2010.216077

Supplementary Material

https://gut.bmj.com/content/71/5/1006#supplementary-materials

04

Prognostic accuracy of FIB-4, NAFLD fibrosis score, and APRI for NAFLD-related events: a systematic review

Jenny Lee
Yasaman Vali
Jerome Boursier
Rene Spijker
Quentin M. Anstee
Patrick M. Bossuyt
Hadi Zafarmand

Liver international, 2021, 41.2: 261-270.

Abstract

Background & Aims: Fibrosis is the strongest predictor for long-term clinical outcomes among patients with non-alcoholic fatty liver disease (NAFLD). There is growing interest in employing non-invasive methods for risk stratification based on prognosis. FIB-4, NFS and APRI are models commonly used for detecting fibrosis among NAFLD patients. We aimed to synthesize existing literature on the ability of these models in prognosticating NAFLD-related events.

Methods: A sensitive search was conducted in two medical databases to retrieve studies evaluating the prognostic accuracy of FIB-4, NFS and APRI among NAFLD patients. Target events were change in fibrosis, liver-related event, and mortality. Two reviewers independently performed reference screening, data extraction and quality assessment (QUAPAS tool).

Results: A total of 13 studies (FIB-4: 12, NFS: 11, APRI: 10), published between 2013 and 2019, were retrieved. All studies were conducted in a secondary or tertiary care setting, with follow-up ranging from one to 20 years. All three markers showed consistently good prognostication of liver-related events (AUC from 0.69 to 0.92). For mortality, FIB-4 (AUC of 0.67 to 0.82) and NFS (AUC of 0.70 to 0.83) outperformed APRI (AUC of 0.52 to 0.73) in all studies. All markers had inconsistent performance for predicting change in fibrosis stage.

Conclusions: FIB-4, NFS and APRI have demonstrated ability to risk stratify patients for liver-related morbidity and mortality, with comparable performance to a liver biopsy, although more head-to-head studies are needed to validate this. More refined models to prognosticate NAFLD-events may further enhance performance and clinical utility of non-invasive markers.

Introduction

In the next 20 years, non-alcoholic fatty liver disease (NAFLD) is projected to become the leading cause of liver transplantation (1, 2). The global prevalence of NAFLD is approximately 25%, among which a proportion may progress to develop non-alcoholic steatohepatitis (NASH) (3). The prevalence of NAFLD-related cirrhosis as the underlying disease among patients undergoing liver transplantation for hepatocellular carcinoma (HCC) has markedly increased in Europe and the United States (4, 5). Patients with NASH have a higher risk of progression to liver fibrosis (6, 7), and those with advanced fibrosis or cirrhosis trend towards more complications of liver failure and HCC compared to those without fibrosis (8).

Liver fibrosis is considered the strongest predictor for long-term clinical outcomes in NAFLD patients (9). Accurate assessment of NASH or fibrosis stage is resource intensive and error-prone, as a liver biopsy is currently required to confirm the diagnosis (10, 11). Moreover, biopsies carry risks for the patient such as severe complications and pain, leaving many unwilling to undergo this invasive procedure.

There is growing promise in risk stratification using non-invasive markers of NAFLD for identifying patients more likely to develop severe liver events. Using markers that are more reliable than a biopsy would circumvent the limitations of a biopsy in stratifying patients. Optimally performing prognostic markers can eventually replace a biopsy and aid clinical decision-making, as well as facilitate recruitment of patients more likely to benefit from participation in clinical trials.

Simple non-invasive panels such as the NAFLD Fibrosis Score (NFS) and Fibrosis-4 (FIB-4) are recommended by the EASL-EASD-EASO Clinical Practice Guidelines as part of the diagnostic regimen for ruling out advanced fibrosis (12). The guidelines further recommend the use of NFS and FIB-4 as prognostic markers to rule out progression to severe disease, including liver-related and all-cause mortality. Other multimarker models such as the aspartate aminotransferase (AST)/platelet ratio index (APRI) are also used for fibrosis staging and prediction of liver-related events (13). Reviewing the literature, we found other markers such as Enhanced Liver Fibrosis (ELF) test or FibroScan had limited assessment for their prognostic ability.

Despite established diagnostic performance, there is limited understanding of the relative merits of the prognostic ability of non-invasive NAFLD markers, and their comparability to a liver biopsy. While many studies have assessed diagnostic performance of these markers in reference to a biopsy, more convincing evidence would link these markers to future clinical events. In this context, we aimed to conduct a systematic review of studies on the accuracy of FIB-4, NFS and APRI in prognosis of fibrosis progression, and liver-related events including mortality.

Methods

This systematic review was conducted as part of the evidence synthesis efforts of the LITMUS project (Liver Investigation: Testing Marker Utility in Steatohepatitis), funded by the European Union’s IMI2 program. LITMUS aims to evaluate biomarkers for drug development in NAFLD. The protocol of the complete systematic review is available in PROSPERO (registration number: CRD42019136118). This study report was prepared using the PRISMA-DTA statement (Supplementary Table 1).

Search strategy

A sensitive search strategy, containing words in the title/abstract or text words across the record and the medical subject heading (MeSH), was developed in close collaboration with an experienced information specialist (RS). The full search strategy is available in Supplementary Table 2. MEDLINE (via OVID) and EMBASE (via OVID) were searched to retrieve potentially eligible studies from inception to June 2019. A search update was conducted in June 2020. Additionally, we manually screened reference lists and contacted partners within the LITMUS consortium.

Study selection

Search results of the two databases were merged and deduplicated using Endnote. Title and abstracts were screened by two independent reviewers (JL and YV), using Rayyan QCRI (http://rayyan.qcri.org). Full texts of potentially eligible studies were retrieved for evaluation against a pre-specified inclusion criterion by the same two reviewers. Any discrepancies were resolved by discussion.

Inclusion and exclusion criteria

We searched for studies published in peer-reviewed journals that had assessed the prognostic accuracy of at least one of the biomarkers of interest (FIB-4, NFS, APRI) in predicting future liver-related events, or changes in fibrosis stage at future biopsies. Publications in any language were eligible for inclusion.

Studies that included adults (≥18 years) diagnosed (based on liver histology) or clinically suspected with NAFLD, and data on either FIB-4, NFS or APRI were eligible. Studies in a mixed cohort of conditions (e.g. NAFLD and viral hepatitis patients) were only included if outcomes were separately reported for NAFLD patients.

The target events of interest were the following:

- worsening (or improvement) of fibrosis stage, evaluated preferably by using the NASH CRN score (14) and the EPoS staging system (15) for all stages of fibrosis or any dichotomized fibrosis status (e.g. F0 – F2 vs F3 – F4);
- other liver-related outcomes of interest, including model of end stage liver disease (MELD) score ≥15; liver transplant; HCC; large oesophageal/gastric varices; ascites; increase in hepatic venous pressure gradient (HVPG) >10 mmHg; histological progression to cirrhosis; hospitalization (as defined by a stay of ≥24 hours) for onset of: variceal bleed, hepatic encephalopathy, spontaneous bacterial peritonitis;
- mortality (liver-related or all-cause).

Studies that reported the area under the ROC curve (AUC) or Harrell’s C index for expressing the prognostic performance in predicting changes in fibrosis stage, liver-related events of interest, or mortality were included. Studies reporting only measures of association, such as a relative risk, hazard ratio, odds ratio, or standard deviation of change, without a direct measure of classification, were excluded.

Data extraction and quality assessment

The following data were extracted from each included study: study characteristics, clinical characteristics, index test features, target event features (if applicable), and overall performance of the test in terms of AUC or C index. Data were independently extracted and cross-checked by a second reviewer (JL and YV).

The Quality Assessment of Prognostic Accuracy Studies (QUAPAS) tool was used to assess the methodological quality and risk of bias in the included studies (16). In short, QUAPAS is a modification of the existing Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool (17), revised to account for items unique to prognostic accuracy study designs. QUAPAS follows the same domain-based framework as QUADAS-2. Two independent reviewers (JL and YV) evaluated risk of bias and concerns for applicability using the five domains (participant recruitment, index test, target event, study flow, analysis), assigning each study with a judgement of ‘low’, ‘high’, or ‘unclear’ risk. See Supplementary Table 3 for the QUAPAS tool.

Statistical Analysis

Given the anticipated heterogeneity between studies, a meta-analysis was not considered.

Results

Search results

Following deduplication, 4,510 studies were eligible for title and abstract screening, of which 126 full texts were screened. We excluded 114 studies in this phase, following the inclusion and exclusion criteria. Two studies that were identified during the search update, despite having prognostic accuracy data, did not present enough data for inclusion (18, 19). Finally, a total of 13 studies, published between 2013 and 2019, were included in the present systematic review (Figure 1).

Characteristics of included studies

The majority of studies (12/13) were comparative accuracy studies, in which two or more biomarkers were evaluated within the same cohort for a given target event. Twelve studies were identified for FIB-4, eleven for NFS and ten for APRI. The study group consisted of NASH patients in three studies (20-22), NAFLD-cirrhotic patients in one study (23), and all others were NAFLD patients. All studies were conducted in a secondary or tertiary care setting. At baseline, the prevalence of diabetic patients ranged from 9% to 78% and hypertension from 11% to 55%. Mean body mass index (BMI) spanned from 28 to 35 kg/m^2. Characteristics of the included studies are summarized in Table 1.

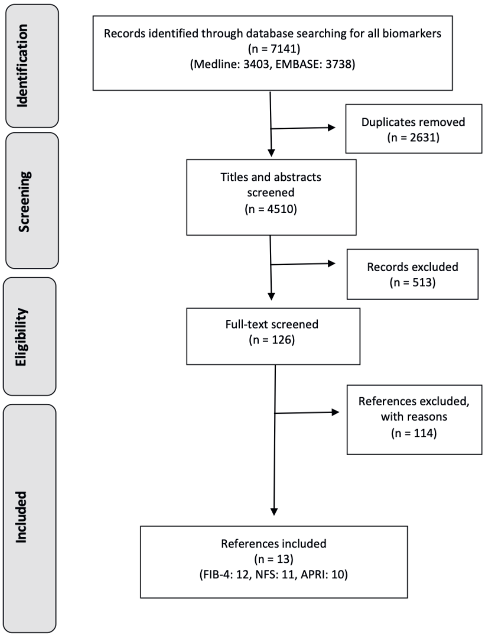

Identification
Screening
Eligibility
Included
Records identified through database searching for all biomarkers (n = 7141) (Medline: 3403, EMBASE: 3738)
Duplicates removed (n = 2631)
Titles and abstracts screened (n = 4510)
Records excluded (n = 513)
Full-text screened (n = 126)
References excluded, with reasons (n = 114)
References included (n = 13) (FIB-4: 12, NFS: 11, APRI: 10)

Figure 1. Flow diagram of included studies

participants, and the relationship between loss to follow-up and the index tests was not explored. Lastly, four studies were graded at high risk of bias for failing to apply methods to account for censoring and competing events (13, 20, 22, 27). Only one study had low risk of bias in the analysis domain (23).

Prognosis of change in fibrosis stage

Table 2 shows the AUC or C-index for the studies included in this systematic review. Change in fibrosis stage (fibrosis progression or regression) was evaluated as the event of interest in three studies (13, 20, 22). All three studies assessed the ability of FIB-4, NFS, and APRI for prognosis of fibrosis progression, defined as an increase of at least one point in fibrosis score. Two studies looked at progression into advanced fibrosis (F ≥3) (13, 28), and another at fibrosis regression (decrease of at least one point in fibrosis score) (22). The cumulative incidence (number of study participants with the target event relative to all study participants at the start of the observation period) of fibrosis spanned from 16% to 43%, with a mean follow-up period of 1 to 6.6 years.

For FIB-4, the prognostic accuracy for fibrosis progression including progression to advanced fibrosis ranged from an AUC of 0.65 (0.54-0.76) to 0.81 (0.73-0.89). The AUC for NFS ranged from 0.65 (0.56-0.73) to 0.83 (0.74-0.92), and for APRI from 0.65 (0.53-0.73) to 0.72 (0.65-0.80).

Few studies reported details regarding threshold values and corresponding sensitivity and specificity. One study used a threshold of 0.2 for all three markers (20). For NFS, suggested high and low thresholds of 0.676 (Se: 0.28, Sp: 0.9) and -1.455 (Se: 0.91, Sp: 0.46), respectively, were used in one study (29). One study also reported sensitivity and specificity data, but with no reporting of threshold (13).

Prognosis of liver-related events

Six studies evaluated liver-related events among NAFLD patients (21, 23, 25, 30-32). Liver-related events were defined as a combination of clinical outcomes, consisting of but not limited to ascites, esophageal varices, encephalopathy, variceal bleeding, decompensated liver disease, HCC, and liver transplantation. Each study assessed a different cluster of events (see Table 2 for details). Two studies included more severe clinical outcomes such as liver failure or death (21, 30). One study evaluated solely HCC (23). The mean follow-up was 1.9 to 19.9 years, with cumulative incidence ranging from 6% to 56%.

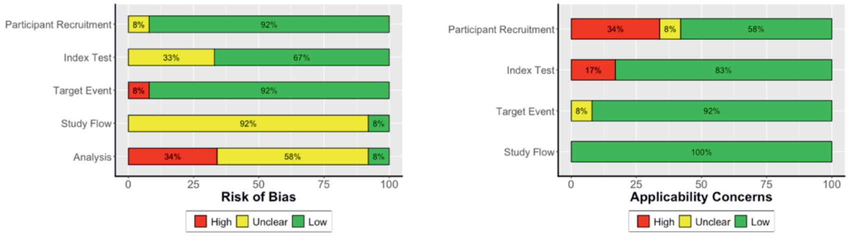


Figure 2. Graphical summary of the risk of bias and applicability concerns of the included studies using the QUAPAS tool

Table 1. Characteristics of the included studies

Author	No. of centers	Country(s)	*N*	Females, n (%)	Mean Age, ±SD	Mean BMI (kg/m2), ±SD	Mean ALT (IU/L), ±SD	Mean AST (IU/L), ±SD	Diabetes (%)	Hypertension (%)
Angulo (2013) (25)	7	USA, Australia, UK, Iceland, Thailand, Italy	309	182 (57)	52 (43 - 61)	33 (29.4 - 36)	61 (38 - 85)	50 (37 - 78)	116 (36)	152 (48)
Treeprasertsuk (2013) (27)	NR	USA	302	169 (56)	47 ± 13	33.6 ± 6.2	61.5 ± 43.3	41.4 ± 21.9	48 (16)	124 (41)
Xun (2014) (24)	1	China	180	84 (47)	39 (30 - 49)	26.0 ± 3.1	129 ± 103	83.7 ± 98.2	17 (9)	20 (11)
Sebastiani (2015) (21)ƚ	1	Canada	148	55 (30)	49.5 ± 10.5	31.3 ± 5.4	NR	NR	49 (33)	58 (39)
McPherson (2015) (28)	1	UK	108	48 (44)	48 ± 12	33.9 ± 5.0	112 ± 80	73 ± 48	52 (48)	NR
Boursier (2016) (26)	1	France	360	124 (34)	59.3 ± 14.3	NR	50 ± 45	40 ± 33	NR	NR
Vilar-Gomez (2017) (22)ƚ	1	Cuba	261	159 (61)	48.5 ± 9.6	31.3 ± 5.3	52.4 ± 34.5	35.2 ± 20.7	90 (35)	NR
Chalasani (2018) (20)ƚ	8	NR	191	NR	NR	NR	NR	NR	NR	NR
Peleg (2018) (32)	1	Israel	153	85 (56)	49.5	NR	NR	NR	97 (63)	64 (41)
Ioannou (2019) (23)ǂ	1221	USA	7068	318 (4.5)	67.1 ± 9.7	33.0 ± 6.6	NR	NR	5506 (78)	NR
Siddiqui (2019) (13)	NR	NR	292	186 (64)	48.9 ± 11.7	34.7 ± 6.3	75.8 ± 50.5	53.9 ± 36.8	113 (39)	160 (55)
Onnerhag (2019) (31)	1	Sweden	144	61 (42.4)	53.2 ± 13.4	28.0 ± 4.6	79.1 ± 64.5	51.9 ± 41.8	32 (22)	66 (46)
Hagstrom (2019) (30)	2	Sweden	646	244 (38)	50 (38 - 58)	28.0 (25.7 - 30.8)	73 (49 - 106)	40 (31 - 59)	93 (14)	196 (30)

ƚ Non-alcoholic steatohepatitis patients; ǂ NAFLD-cirrhotic patients; NR: Not Reported, DM: diabetes mellitus, HTN: hypertensio

The AUC for prognosis of liver-related events ranged from 0.71 to 0.89 for FIB-4, 0.72 to 0.92 for NFS, and 0.69 to 0.89 (0.82-0.96) for APRI (Table 2). In the two studies that conducted statistical testing, both showed significant differences ($p < 0.005$) between the three markers and the null hypothesis (AUC of 0.5) (25, 31). In one study that compared non-invasive methods

to a liver biopsy, FIB-4 and APRI had higher AUC than histologic fibrosis (21). Length of follow-up period did not seem to influence the performance of any biomarker in a consistent pattern.

In prognosticating liver-related events, most studies reported using either one or both the suggested high and low thresholds for NFS (low: -1.45, high: 0.676) and APRI (low: 0.5, high: 1.5). For FIB-4, two studies used a single threshold of 3.25 (one study finding a sensitivity and specificity of 0.59 and 0.92, respectively) (21, 23), while the rest adhered to the suggested low threshold of 1.3 and/or high threshold of 2.67. In the sole study that reported paired point accuracy data, the high threshold showed a sensitivity and specificity of 0.50 and 0.90 for NFS, and 0.50 and 0.92, for APRI, respectively (21).

Prognosis of mortality (liver-related and all-cause)

All-cause mortality was the most frequently investigated event, evaluated in seven studies (24-27, 30-32). One study additionally looked at liver-related mortality (26). The cumulative incidence was between 5% and 59%; mean follow-up ranged from 1.9 to 19.9 years.

The prognostic accuracy of FIB-4, expressed as the AUC, ranged from 0.67 (0.58-0.76) to 0.82 (0.75-0.90) (Table 2). The AUC reported for NFS ranged from 0.70 (0.62-0.78) to 0.83 (0.73-0.93). The accuracy of APRI was lower compared to FIB-4 and NFS in all seven studies, with AUC ranging from 0.52 to 0.73 (0.60-0.86). Four out of four studies showed significant results ($p < 0.05$) (24-26, 31). Here also, length of follow-up did not seem to influence the performance of any biomarker.

Of the studies that reported the threshold values used for prognosticating mortality, all used either or both the suggested high and low thresholds for FIB-4 and APRI. For NFS,

Table 2. Accuracy of biomarkers FIB-4, NFS and APRI in prognosticating change in fibrosis stage, liver-related events, and mortality among NAFLD patients

Author	Target event	No. of cases(%)ǂ	Time horizon (years)	AUC/C-index		
				FIB-4	NFS	APRI
Fibrosis						
Vilar-Gomez (2017)	Fibrosis progression[1]	45 (17)	1	0.65 (0.54-0.76)	0.69 (0.58-0.79)	0.65 (0.53-0.73)
Chalasani (2018)	Fibrosis progression[1]	NA	1.4	0.68 (0.60-0.76)	0.65 (0.56-0.73)	0.72 (0.65-0.80)
Siddiqui (2019)	Fibrosis progression[1]	92 (32)	2.6	0.73 (0.67-0.79)	0.66 (0.59-0.73)	0.70 (0.63-0.77)
McPherson (2015)	Progression to fibrosis stage ≥3	46 (43)	6.6	NA	0.83 (0.74-0.92)*	0.72 (0.62-0.82)*
Siddiqui (2019)	Progression to fibrosis stage ≥3	35 (16)	2.6	0.81 (0.73-0.89)	0.80 (0.71-0.88)	0.82 (0.74-0.89)
Vilar-Gomez (2017)	Fibrosis regression[2]	51 (20)	1	0.57 (0.51-0.68)	0.63 (0.58-0.75)	0.59 (0.52-0.70)
Liver-related events						
Ioannou (2019)	HCC[3]	407 (6)	3.7	0.71	NA	NA
Peleg (2018)	Liver-related events[4]	86 (56)	1.9	0.89	0.92	0.73
Angulo (2013)	Liver-related events[5]	60 (19)	8.7	0.86 (0.80-0.92)*	0.81 (0.76-0.87)*	0.80 (0.73-0.86)*
Onnerhag (2019)	Liver-related events[6]	20 (14)	17.7	0.81 (0.69-0.93)*	0.77 (0.64-0.89)*	0.82 (0.72-0.92)*
Hagstrom (2019)	Severe liver disease[7]	76 (12)	19.9	0.72	0.72	0.69
Sebastiani (2015)	Clinical outcomes[8]	25 (17)	5	0.79 (0.69-0.91)	0.89 (0.83-0.95)	0.89 (0.82-0.96)
Mortality						
Boursier (2016)	Liver-related mortality	17 (5)	6.4	0.78 (0.66-0.88)*	NA	0.69 (0.49-0.84)*
Peleg (2018)	All-cause mortality	19 (12)	1.9	0.78	0.80	0.63

Boursier (2016)	All-cause mortality	83 (23)	6.4	0.70 (0.64-0.75)	NA	0.54 (0.46-0.61)
Xun (2014)	All-cause mortality	12 (7)	6.6	0.81 (0.70-0.91)**	0.83 (0.73-0.93)**	0.73 (0.60-0.86)**
Angulo (2013)	All-cause mortality[9]	41 (13)	8.7	0.67 (0.58-0.76)**	0.70 (0.62-0.78)*	0.63 (0.53-0.72)**
Treeprasertsuk (2013)	All-cause mortality	39 (13)	11.9	NA	0.70	NA
Onnerhag (2019)	All-cause mortality	85 (59)	17.7	0.82 (0.75-0.90)*	0.82 (0.74-0.90)*	0.59 (0.50-0.68)
Hagstrom (2019)	All-cause mortality	214 (33)	19.9	0.72	0.72	0.52

ɫ Cumulative incidence: number of new cases/number of persons at start of the observation period

[1] Increase of at least 1 point in fibrosis score

[2] Decrease of at least 1 point in fibrosis score

[3] Hepatocellular carcinoma, defined as ICD-9 code 155.0 and ICD-10 code C22.0

[4] Ascites, esophageal varices, hepatic encephalopathy, liver transplantation, TIPS or hospitalizations

[5] Ascites, gastroesophageal varices/bleeding, portosystemic encephalopathy, spontaneous bacterial peritonitis, hepatocellular cancer, hepatopulmonary syndrome, or hepatorenal syndrome

[6] Ascites, encephalopathy, variceal bleeding, or hepatocellular carcinoma

[7] Cirrhosis, decompensated liver disease, liver failure, or hepatocellular carcinoma

[8] Death, liver transplantation and end-stage hepatic complications defined as hepatocellular carcinoma, ascites, spontaneous bacterial peritonitis, hepatic encephalopathy, de novo varices or significant worsening of varices

[9] Including liver transplant

* p-value < 0.001, ** p-value < 0.0

thresholds of -0.9 and -1.836 were also studied in addition to the suggested thresholds. We again found sparse reporting of sensitivity and specificity. One study found that at the high threshold, FIB-4, NFS and APRI showed sensitivity and specificity of 0.70 and 0.72, 0.69 and 0.76, and 0.55 and 0.89, respectively (32).

Discussion

Non-invasive markers with comparable ability to prognosticate severe liver-related outcomes may be valuable tools for stratifying patients with higher risk of complication, in place of a liver biopsy. In this systematic review, we aimed to summarize the evidence on the prognostic performance of three multimarker models in identifying those at risk of developing worsening of NAFLD-related outcomes. We found that FIB-4, NFS and APRI have limited performance in predicting changes in fibrosis, as evaluated by future biopsies, but consistently demonstrated the ability to predict liver-related morbidity and mortality, with a level of performance that met or exceeded that of a liver biopsy.

Strengths and limitations

While many studies have synthesized data on the diagnostic accuracy of non-invasive NAFLD markers, to our knowledge, this is the first systematic review conducted on the prognostic context of use. In collaboration with a search specialist, we developed a highly sensitive search strategy, including abstracts, to minimize bias that may arise from selective inclusion. For robust evaluation of bias in individual studies, we used a new risk of bias tool developed specifically for systematic reviews of prognostic accuracy (16). All screening phases, data extraction and quality assessment were independently conducted by two experienced methodologists.

Our work comes with limitations, some inherent to the nature of prognostic research. Several studies had a relatively short follow-up period. This can be problematic for assessing outcomes of a chronic condition such as NAFLD, where patients have a median survival period of >10 years (33, 34). The results should be interpreted with caution, given the limited and heterogeneous follow-up periods, which ranged from one to 20 years. The variability in study designs prohibited meta-analysis to produce summary estimates of performance.

In the scheme of disease management, risk stratification may be most beneficial in a primary care setting, in which the purpose is to identify patients who require expedited referral to tertiary care centers. All identified studies evaluated the markers prognostic performance in a secondary or tertiary care setting. Thus, data from these studies cannot necessarily be extrapolated to a primary care setting.

Furthermore, we observed that very few studies reported data on both threshold values and corresponding sensitivity and specificity, which are more informative and clinically relevant than the AUC alone. Sparse reporting may be attributed to the relatively new and therefore less established nature of prognostic accuracy studies in general, in comparison to diagnostic accuracy studies. Given the increased volume of prognostic accuracy research, reporting guidelines and quality assessment tools specific for this area of research should be further developed.

In the context of current evidence

A 2015 editorial illustrated the prognostic value of histological features of NAFLD, in the form of a hierarchical model (34). This model ranked fibrosis as the most important histologic lesion associated with long term outcomes in NAFLD, and many studies support biopsy-confirmed fibrosis to be a major prognostic marker for mortality (35, 36). However, growing literature highlights the limitations of a liver biopsy (11), particularly for detection of fibrosis (37). Aside from the risk of complications and invasive nature, sampling variability is a big concern. In a study by Ratziu et al., where two biopsy samples were compared, fibrosis stage was different in 41% of patients (38). This may not be surprising, as only 1/50,000 of a whole liver tissue is sampled during a biopsy (39). Even for NASH, histological lesions are unevenly distributed throughout the liver tissue. Further problems with pathological diagnosis arise with inter- and intra-observer variability. Therefore, evaluating test accuracy with an imperfect reference standard such as a liver biopsy poses the risk of underestimating NASH and fibrosis severity.

While histologic fibrosis predicts disease progression, prognostication of NAFLD-related events using non-invasive markers is an appealing alternative, especially if performance of these markers approximates or equals that of histology-confirmed fibrosis. In comparing the performance of non-invasive methods to histologic fibrosis (F3-F4) in prognosticating liver-related events, APRI and FIB-4 had higher AUC compared to a biopsy, and the overall percent of accurate prognosis was higher for all three multimarker models (models had 84% to 86% accuracy compared to 76% with a liver biopsy) (21). The AUC found in this study were consistent with others identified in this systematic review.

This direct comparison illustrated the ability of non-invasive markers to risk stratify patients with comparable, or even better performance than a liver biopsy. Another study supported this finding for the ELF test (40). However, studies evaluating head-to-head comparisons of non-invasive markers and a liver biopsy are limited, and future studies should aim to validate these findings and build a stronger evidence-base for non-invasive tests, particularly for the simple multimarker models that contain components readily evaluated in routine laboratories.

In addition to FIB-4, NFS and APRI, other NAFLD markers have been studied for their prognostic ability. The ELF test is recognized by guidelines as a diagnostic marker for liver fibrosis. For predicting progression to cirrhosis and liver related events, the AUC for ELF was 0.79 and 0.68, respectively, out-performing histological assessment for both outcomes (40). Vibration-controlled transient elastography (VCTE), a imaging technique validated for liver fibrosis, had an AUC of 0.73 (0.66-0.78) for all-cause mortality, significantly outperforming APRI ($p = 0.001$) but not FIB-4 (26). Liver stiffness measurement, by transient elastography (FibroScan) had an AUC of 0.86 (0.82-0.95) in prognosticating liver-related mortality in one study (26), and an AUC of 0.911 (0.82-0.99) in prognosticating liver-related events (41). Fibroscan significantly outperformed APRI for predicting all-cause mortality. FibroTest, another marker for determining stages of NAFLD-related fibrosis, had an AUC of 0.94 (0.91-0.98) in prognostication of liver-related death (42). The same study conducted a post hoc analysis comparing FibroTest and FIB-4 and found no significant difference in performance (p = 0.32). In this study, FIB-4 had an AUC of 0.87 (0.74-0.99). Longitudinal assessment of magnetic resonance elastography (MRE) showed prognostic accuracy of 0.62 (0.46-0.78) for predicting fibrosis improvement and magnetic resonance imaging (MRI-PDFF) had an AUC of 0.70 (0.57-0.83) for predicting steatosis reduction (43). While some of these markers show promising results, more studies are needed to validate the findings.

Implications for current practice

In clinical practice, FIB-4 and NFS can be used in regular intervals to detect disease progression, offering a less invasive, and perhaps a more accurate alternative to a biopsy. The annual change of NFS in patients who died was two-fold that of survivors and, for fibrosis progression, four-fold higher in progressors than in those who were stable (27). Another study found that FIB-4 and NFS were significantly higher among fibrosis progressors compared to non-progressors, despite no significant difference in histological grading (28). Patients who underwent serial measurements of FIB-4 within five years and had high-risk in both occurrences had significantly increased risk of severe liver disease with an adjusted hazard ratio of 17.04 (11.67-24.88), and an accuracy of 98% (44).

The costs and time invested into drug development has become increasingly exhaustive (45). Given the volume of ongoing clinical trials for the treatment of NASH and fibrosis, and the understood complexities and required resources, prognostic markers can be an integral measure for expediting clinical trials. A marker linked to a clinical trial endpoint can improve efficiency for late stage clinical trials by identifying patients more likely to develop the outcome, ultimately reducing the number of participants recruited to a study

(46). For clinical trials targeting patients with cirrhosis, long term events that characterize clinical decompensation (ascites, encephalopathy, HCC, variceal hemorrhage) are of interest (47). We observed that all three markers showed consistently good prognostic performance for events indicating clinical decompensation.

In conclusion, this systematic review shows that FIB-4, NFS and APRI can risk stratify patients for liver-related morbidity and mortality, with comparable performance to a liver biopsy. If confirmed in future comparative studies with sufficient length of follow-up, the strong prognostic performance of these multimarker models could position them at the cornerstone for risk stratification and risk management among NAFLD patients.

References

1. 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 Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association. Hepatology. 2012;55(6):2005-23.
2. Wong RJ, Cheung R, Ahmed A. Nonalcoholic steatohepatitis is the most rapidly growing indication for liver transplantation in patients with hepatocellular carcinoma in the U.S. Hepatology. 2014;59(6):2188-95.
3. Younossi Z, Anstee QM, Marietti M, Hardy T, Henry L, Eslam M, et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol. 2018;15(1):11-20.
4. Pais R, Barritt ASt, Calmus Y, Scatton O, Runge T, Lebray P, et al. NAFLD and liver transplantation: Current burden and expected challenges. Journal of hepatology. 2016;65(6):1245-57.
5. Anstee QM, Reeves HL, Kotsiliti E, Govaere O, Heikenwalder M. From NASH to HCC: current concepts and future challenges. Nat Rev Gastroenterol Hepatol. 2019;16(7):411-28.
6. Powell EE, Cooksley WG, Hanson R, Searle J, Halliday JW, Powell LW. The natural history of nonalcoholic steatohepatitis: a follow-up study of forty-two patients for up to 21 years. Hepatology. 1990;11(1):74-80.
7. Fassio E, Alvarez E, Dominguez N, Landeira G, Longo C. Natural history of nonalcoholic steatohepatitis: a longitudinal study of repeat liver biopsies. Hepatology (Baltimore, Md). 2004;40(4):820-6.
8. Adams LA, Lymp JF, St Sauver J, Sanderson SO, Lindor KD, Feldstein A, et al. The natural history of nonalcoholic fatty liver disease: a population-based cohort study. Gastroenterology. 2005;129(1):113-21.
9. Angulo P, Kleiner DE, Dam-Larsen S, Adams LA, Bjornsson ES, Charatcharoenwitthaya P, 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(2):389-97.e10.
10. Williams CD, Stengel J, Asike MI, Torres DM, Shaw J, Contreras M, et al. Prevalence of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis among a largely middle-aged population utilizing ultrasound and liver biopsy: a prospective study. Gastroenterology. 2011;140(1):124-31.
11. Lee DH. Noninvasive Evaluation of Nonalcoholic Fatty Liver Disease. Endocrinol Metab. 2020;35(2):243-59.
12. European Association for the Study of the L, European Association for the Study of D, European Association for the Study of O. EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease. J Hepatol. 2016;64(6):1388-402.
13. Siddiqui M, S., Yamada G, Vuppalanchi R, Van Natta M, Loomba R, et al. Diagnostic Accuracy of Noninvasive Fibrosis Models to Detect Change in Fibrosis Stage. Clinical Gastroenterology & Hepatology.4:04.
14. 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.
15. Bedossa P, Arola J, Susan D, Gouw A, Maria G, Lackner K, et al. The EPoS staging system is a reproducible 7-tierfibrosis score for NAFLD adapted both to glass slides and digitized images (e-slides). Journal of Hepatology. 2018;68:S553.
16. Lee J, Vali Y, Zafarmand M, Bossuyt P. Quality Assessment of Prognostic Accuracy Studies (QUAPAS): an extension of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool for systematic reviews of prognostic test accuracy studies Abstracts of the 26th Cochrane Colloquium, Santiago, Chile. : Cochrane Database of Systematic Reviews 2020; 2020.
17. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.

18. Hagström H, Talbäck M, Andreasson A, Walldius G, Hammar N. Ability of Noninvasive Scoring Systems to Identify Individuals in the Population at Risk for Severe Liver Disease. Gastroenterology. 2020;158(1):200-14.
19. Munteanu M, Pais R, Peta V, Deckmyn O, Moussalli J, Ngo Y, et al. Long-term prognostic value of the FibroTest in patients with non-alcoholic fatty liver disease, compared to chronic hepatitis C, B, and alcoholic liver disease. Aliment Pharmacol Ther. 2018;48(10):1117-27.
20. Chalasani N, Abdelmalek M, F., Loomba R, Kowdley K, V., et al. Relationship between three commonly used non-invasive fibrosis biomarkers and improvement in fibrosis stage in patients with non-alcoholic steatohepatitis. Liver International.39(5):924-32.
21. Sebastiani G, Alshaalan R, Wong P, Rubino M, Salman A, Metrakos P, et al. Prognostic Value of Non-Invasive Fibrosis and Steatosis Tools, Hepatic Venous Pressure Gradient (HVPG) and Histology in Nonalcoholic Steatohepatitis. PLoS ONE [Electronic Resource]. 2015;10(6):e0128774.
22. Vilar-Gomez E, Calzadilla-Bertot L, Friedman S, L., Gra-Oramas B, Gonzalez-Fabian L, et al. Serum biomarkers can predict a change in liver fibrosis 1 year after lifestyle intervention for biopsy-proven NASH. Liver International.37(12):1887-96.
23. Ioannou G, N., Green P, Kerr K, F., Berry K. Models estimating risk of hepatocellular carcinoma in patients with alcohol or NAFLD-related cirrhosis for risk stratification. Journal of Hepatology.27:27.
24. Xun Y, H., Guo J, C., Lou G, Q., et al. Non-alcoholic fatty liver disease (NAFLD) fibrosis score predicts 6.6-year overall mortality of Chinese patients with NAFLD. Clinical & Experimental Pharmacology & Physiology.41(9):643-9.
25. Angulo P, Bugianesi E, Bjornsson E, S., Charatcharoenwitthaya P, Mills P, et al. Simple noninvasive systems predict long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology.145(4):782-9.e4.
26. Boursier J, Vergniol J, Guillet A, Hiriart J, B., Lannes A, et al. Diagnostic accuracy and prognostic significance of blood fibrosis tests and liver stiffness measurement by FibroScan in non-alcoholic fatty liver disease. Journal of Hepatology.65(3):570-8.
27. Treeprasertsuk S, Bjornsson E, Enders F, Suwanwalaikorn S, Lindor K, D. NAFLD fibrosis score: a prognostic predictor for mortality and liver complications among NAFLD patients. World Journal of Gastroenterology.19(8):1219-29.
28. McPherson S, Hardy T, Henderson E, Burt AD, Day CP, Anstee QM. Evidence of NAFLD progression from steatosis to fibrosing-steatohepatitis using paired biopsies: Implications for prognosis and clinical management. Journal of Hepatology. 2015;62(5):1148-55.
29. McPherson S, Stewart SF, Henderson E, Burt AD, Day CP. Simple non-invasive fibrosis scoring systems can reliably exclude advanced fibrosis in patients with non-alcoholic fatty liver disease. Gut. 2010;59(9):1265-9.
30. Hagstrom H, Nasr P, Ekstedt M, Stal P, Hultcrantz R, Kechagias S. Accuracy of Noninvasive Scoring Systems in Assessing Risk of Death and Liver-Related Endpoints in Patients With Nonalcoholic Fatty Liver Disease. Clinical Gastroenterology & Hepatology.17(6):1148-56.e4.
31. Onnerhag K, Hartman H, Nilsson P, M., Lindgren S. Non-invasive fibrosis scoring systems can predict future metabolic complications and overall mortality in non-alcoholic fatty liver disease (NAFLD). Scandinavian Journal of Gastroenterology.54(3):328-34.
32. Peleg N, Sneh Arbib O, Issachar A, Cohen-Naftaly M, Braun M, Shlomai A. Noninvasive scoring systems predict hepatic and extra-hepatic cancers in patients with nonalcoholic fatty liver disease. PLoS ONE [Electronic Resource]. 2018;13(8):e0202393.
33. Calzadilla Bertot L, Adams LA. The Natural Course of Non-Alcoholic Fatty Liver Disease. International journal of molecular sciences. 2016;17(5):774.
34. Loomba R, Chalasani N. The Hierarchical Model of NAFLD: Prognostic Significance of Histologic Features in NASH. Gastroenterology. 2015;149(2):278-81.
35. Angulo P, Kleiner D, E., Dam-Larsen S, Adams L, A., et al. Liver Fibrosis, but No Other Histologic Features, Is Associated With Long-term Outcomes of Patients With Nonalcoholic Fatty Liver Disease. Gastroenterology.149(2):389-97.e10.
36. Taylor RS, Taylor RJ, Bayliss S, Hagström H, Nasr P, Schattenberg JM, et al. Association Between Fibrosis Stage and Outcomes of Patients With Nonalcoholic Fatty Liver Disease: A Systematic Review and Meta-Analysis. Gastroenterology. 2020;158(6):1611-25.e12.

37. Sumida Y, Nakajima A, Itoh Y. Limitations of liver biopsy and non-invasive diagnostic tests for the diagnosis of nonalcoholic fatty liver disease/nonalcoholic steatohepatitis. World journal of gastroenterology. 2014;20(2):475-85.

38. 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.

39. Goldstein NS, Hastah F, Galan MV, Gordon SC. Fibrosis heterogeneity in nonalcoholic steatohepatitis and hepatitis C virus needle core biopsy specimens. Am J Clin Pathol. 2005;123(3):382-7.

40. Sanyal A, J., Harrison S, A., Ratziu V, Abdelmalek M, et al. The Natural History of Advanced Fibrosis Due to Nonalcoholic Steatohepatitis: Data From the Simtuzumab Trials. Hepatology.16:16.

41. Shili-Masmoudi S, Wong GL-H, Hiriart J-B, Liu K, Chermak F, Shu SS-T, et al. Liver stiffness measurement predicts long-term survival and complications in non-alcoholic fatty liver disease. Liver International. 2020;40(3):581-9.

42. Munteanu M, Pais R, Peta V, Deckmyn O, Moussalli J, Ngo Y, et al. Long-term prognostic value of the FibroTest in patients with non-alcoholic fatty liver disease, compared to chronic hepatitis C, B, and alcoholic liver disease. Alimentary pharmacology & therapeutics. 2018;48(10):1117-27.

43. Jayakumar S, Middleton M, S., Lawitz E, J., Mantry P, et al. Longitudinal correlations between MRE, MRI-PDFF, and liver histology in patients with non-alcoholic steatohepatitis: Analysis of data from a phase II trial of selonsertib. Journal of Hepatology.70(1):133-41.

44. Hagström H, Talbäck M, Andreasson A, Walldius G, Hammar N. Repeated FIB-4 measurements can help identify individuals at risk of severe liver disease. J Hepatol. 2020.

45. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nature reviews Drug discovery. 2004;3(8):711-6.

46. Bakhtiar R. Biomarkers in drug discovery and development. Journal of Pharmacological and Toxicological Methods. 2008;57(2):85-91.

47. Sanyal AJ, Brunt EM, Kleiner DE, Kowdley KV, Chalasani N, Lavine JE, et al. Endpoints and clinical trial design for nonalcoholic steatohepatitis. Hepatology (Baltimore, Md). 2011;54(1):344-53.

Supplementary Material

https://onlinelibrary.wiley.com/doi/10.1111/liv.14669

05

QUAPAS: an adaptation of the QUADAS-2 tool to assess prognostic accuracy studies

Jenny Lee
Frits Mulder
Mariska Leeflang
Robert Wolff
Penny Whiting
Patrick M Bossuyt

Annals of Internal Medicine, 2022, 175.7: 1010-1018.

Abstract

Whereas diagnostic tests help detect the cause of signs and symptoms, prognostic tests assist in evaluating the probable course of the disease and future outcome. Studies to evaluate prognostic tests are longitudinal, which introduces sources of bias different from those for diagnostic accuracy studies. At present, systematic reviews of prognostic tests often use the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) tool to assess risk of bias and applicability of included studies because no equivalent instrument exists for prognostic accuracy studies.

QUAPAS (Quality Assessment of Prognostic Accuracy Studies) is an adaptation of QUADAS-2 for prognostic accuracy studies. Questions likely to identify bias were evaluated in parallel and collated from QUIPS (Quality in Prognosis Studies) and PROBAST (Prediction Model Risk of Bias Assessment Tool) to paired to the coresponding question (or domain) in QUADAS-2. A steering group conducted and reviewed 3 rounds of modifications before arriving at the final set of domains and signaling questions.

QUAPAS follows the same steps as QUADAS-2: Specify the review question, tailor the tool, draw a flow diagram, judge risk of bias, and identify applicability concerns. Risk of bias is judged across the following 5 domains: participants, index test, outcome, flow and timing, and analysis. Signaling questions assist the final judgment for each domain. Applicability concerns are assessed for the first 4 domains.

The authors used QUAPAS in parallel with QUADAS-2 and QUIPS in a systematic review of prognostic accuracy studies. QUAPAS improved the assessment of the flow and timing domain and flagged a study at risk of bias in the new analysis domain. Judgment of risk of bias in the analysis domain was challenging because of sparse reporting of statistical methods.

Introduction

Informed decision making in clinical care relies on accuracy in both diagnosis and prognosis. Medical tests can support such decisions: Diagnostic tests help clinicians detect the cause of a patient's signs and symptoms, and prognostic tests assist in evaluating the probable course of the disease and future outcome.

The 2 different contexts of use make the interpretation and evaluation of the performance of a test for a diagnostic purpose distinctly different from that for a prognostic purpose (1). An evaluation of the diagnostic accuracy of a test aims to answer a cross-sectional question: How good is this test at detecting the target condition in patients presenting with signs and symptoms, here and now? Diagnostic accuracy studies evaluate 1 or more index tests by comparing the results with the reference standard outcome in the same patient (2). The findings are usually expressed as estimates of the test's sensitivity or specificity. An example is the evaluation of D-dimer in patients with suspected pulmonary embolism, using computed tomography scans as the reference standard (3).

In contrast, evaluations of the prognostic accuracy of a test address longitudinal questions: How good is this test at predicting a future patient outcome, such as an event or a specific functional status (4)? This can be evaluated by comparing results of the same test in patients who developed versus did not develop the future outcome. A prognostic accuracy study might, for example, assess whether a scoring system for sepsis among emergency department patients can prognosticate in-hospital death (5), or evaluate C-reactive protein for early risk assessment of patients with acute pancreatitis (6).

A prognostic test can be based on the measurement of a single biomarker, a multimarker score, an imaging modality, or other methods. The test result can be dichotomous or expressed on an ordinal or quantitative scale, such as calculated risk. The performance of the test—here generically called prognostic accuracy, but also known as predictive accuracy or discrimination—can be expressed in terms of sensitivity and specificity for dichotomized results, as the area under the receiver-operating characteristic curve, as the c-index, or in other ways.

Most diagnostic accuracy studies are cross-sectional in design, although some, known as delayed cross-sectional, rely on follow-up as the reference standard (7). Studies to evaluate prognostic tests, on the other hand, are always longitudinal in nature: cohort studies in tested patients, for example, with the outcome captured during a follow-up period, or nested case–control studies, with test results obtained in previously collected samples (8).

Systematic reviews have become the preferred method for synthesizing the available evidence on the performance of a test. A key component of a systematic review is an assessment of methodological quality of the included studies, and several tools have been developed to assist in this evaluation. In systematic reviews of diagnostic accuracy studies, QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) is commonly used to assess risk of bias and express concerns about the applicability of the study findings to the review question (9).

No specific tool exists to assist quality assessment of prognostic accuracy studies. Reviewers sometimes use QUADAS-2, which usually requires tailoring because it does not address some of the issues in longitudinal studies. Reviewers also rely on quality assessment tools for prognostic questions, such as QUIPS (Quality in Prognosis Studies) (10) and PROBAST (Prediction Model Risk of Bias Assessment Tool) (11, 12), neither of which is ideally suited for evaluating prognostic accuracy studies. Because QUIPS was developed for prognostic factor studies, it also emphasizes confounding, an element less relevant when evaluating the accuracy of prognostic tests. PROBAST was specifically developed for prediction models, but not all prognostic tests are model-based, and questions about model development do not apply. To alleviate some of these problems, some review authors use 2 or 3 tools in parallel when evaluating studies of prognostic tests (13, 14).

We believed that it would be helpful to have a tool for assessing the methodological quality of prognostic accuracy studies that retains the widely accepted structure and easy-to-communicate flow of QUADAS-2 while addressing some of the specific features of longitudinal studies, as QUIPS and PROBAST do.

We propose QUAPAS (Quality Assessment of Prognostic Accuracy Studies), an adaptation of QUADAS-2 to assess risk of bias and applicability in systematic reviews of prognostic accuracy studies. QUAPAS combines the structure and process of QUADAS-2 with elements from QUIPS and PROBAST.

Development of QUAPAS

Six experienced reviewers and methodologists (J.L., P.W., M.L., R.W., P.M.B., and Jill Hayden, PhD), experts in the area of test evaluation or development of risk-of-bias tools, participated in the development of QUAPAS. The tool was developed with the intention of assessing risk of bias and applicability concerns for the evaluation of the performance of a single index test against 1 future outcome.

We reviewed risk-of-bias tools suggested for systematic reviews by the Cochrane Prognosis Methods Group and Diagnostic Test Accuracy Working Group (15). We relied on these existing tools because they have demonstrated ability to assess bias and applicability in the related research areas of test accuracy and prognosis. Published systematic reviews were screened at random to identify any other relevant tools. Although we did not identify other unique tools, we observed that reviewers sometimes used a combination of QUADAS-2 and either QUIPS or PROBAST.

Considering the similarities between diagnostic and prognostic accuracy studies, we used the 4 domains of QUADAS-2 (patient selection, index tests, reference standard, and flow and timing) as the starting point. From there, we evaluated in parallel the 6 domains of QUIPS (study participation, study attrition, prognostic factor measurement, outcome measurement, study confounding, and statistical analysis and reporting) and the 4 domains of PROBAST (participants, predictors, outcome, and analysis) to identify and pair to corresponding QUADAS-2 domains.

This tool relies on signaling questions—factual questions to assist the final judgment for each domain—as do QUADAS-2 and other tools. These questions were collated from the 3 tools. We kept intact the original QUADAS-2 signaling questions as much as possible. Duplicate questions from QUIPS and PROBAST were eliminated on the basis of style and clarity of wording and sometimes revised to be consistent with QUADAS-2. Additional signaling questions and sources of bias, relevant for prognostic accuracy studies, were included on the basis of discussions from the steering committee meetings.

The steering committee conducted and reviewed 3 rounds of modifications before arriving at the final set of domains and signaling questions. Table 1 lists the key changes made to QUADAS-2 in its transformation to QUAPAS.

Explanation of Major Changes From QUADAS-2 to QUAPAS

The proposed QUAPAS tool covers the following 5 domains: participants, index test, outcome, flow and timing, and analysis (Table 2). The complete user template is available in Part 1 of the Supplement (available at Annals.org), with footnotes to guide reviewers in answering the signaling questions.

QUAPAS follows the same structure as QUADAS-2 and is applied in 4 phases. Details about the phases can be found in the Appendix (available at Annals.org) and in the QUADAS-2 publication (9).

Part 2 of the Supplement includes examples for answering the signaling questions in published prognostic accuracy studies. The following section describes the major changes from QUADAS-2 to QUAPAS.

Domain 1: Participants

Risk of bias: Could the selection of participants have introduced bias?

This domain covers methods for participant enrollment and avoidance of inappropriate exclusions at the point of entry to the study; exclusions after enrollment are covered in the flow and timing domain.

Applicability: Are there concerns that the participants do not match the review question?

Concerns about applicability arise if the study group does not reflect the population in the review question in terms of characteristics, disease severity, and clinical setting. This matters because the performance of a test typically varies across demographic and clinical groups.

Changes

No major changes were made in this domain; Table 1 shows minor changes.

Table 1. Key Changes Made From QUADAS-2 to QUAPAS

Signaling Question	Changes From QUADAS-2 and Source (If Applicable)	Explanation
Domain 1: Participants		
	Domain name change Source: QUIPS, PROBAST	The name of this domain was modified from *patient selection* (QUADAS-2) to *participants*.
S1.1: Was a consecutive or random sample of participants enrolled?	Modified QUADAS-2 signaling question	*Patients* was modified to *participants*, here and throughout the tool.
S1.3: Did the study avoid inappropriate selection criteria?	Modified QUADAS-2 signaling question	This signaling question from QUADAS-2 ("Did the study avoid inappropriate exclusions?") was modified because the selection criteria consider both appropriate inclusion and exclusion of participants.
Domain 2: Index test		
S2.1: Was the method used to perform the index test valid and reliable?	Added signaling question Source: QUIPS	This prompting item from QUIPS ("Method of PF measurement is … valid and reliable") was included as a signaling question to flag any sources of measurement error in index test measurements.
S2.2: Was the method for performing the index test the same for all participants?	Added signaling question Source: QUIPS, PROBAST	This signaling question was a combination of a prompting item from QUIPS ("The method and setting of measurement of PF is the same for all study participants") and a signaling question from PROBAST ("Were predictors defined and assessed in a similar way for all participants?").
S2.3: Were the index test results interpreted without knowledge of the outcome?	Modified QUADAS-2 signaling question	*Reference standard* was modified to *outcome,* here and throughout the tool.
Domain 3: Outcome		
	Domain name change	The domain name was modified from *reference standard* to *outcome* because most prognostic tests focus on an event or state in the future.
S3.1: Was the method used to measure the outcome valid and reliable?	Modified QUADAS-2 signaling question	We rephrased a question from QUADAS-2 ("Is the reference standard likely to correctly classify the target condition?") to better emphasize sources of outcome misclassification in longitudinal studies and to be consistent with wording in S2.1.
S3.2: Was the method used to measure the outcome the same for all participants?	Redirected and modified QUADAS-2	This signaling question is a modification of one in the flow and timing domain of QUADAS-2 ("Did all patients receive the same reference standard?"). We redirected this question to the current domain

	signaling question	because the tool is structured in such a way that all signaling questions addressing bias related to outcome are in the outcome domain.
S3.3: Was the outcome measured without knowledge of the index test results?	Modified QUADAS-2 signaling question	*Reference standard* was modified to *outcome.*
Domain 4: Flow and timing		
S4.1: Did all participants receive the index test?	Added signaling question Source: PROBAST	This signaling question was adopted from a PROBAST question (“Are all predictors available at the time the model is intended to be used?”). We included it to account for any participants who did not undergo the index test.
S4.2: Was treatment avoided after the index test was performed?	New signaling question	This signaling question was introduced to highlight potential for treatment selection bias. Differences in baseline clinical features can result in a subset of participants receiving treatment, which may affect the likelihood of the outcome thereafter.
S4.3: Was the time horizon sufficient for capturing the outcome?	Modified QUADAS-2 signaling question	We modified an existing QUADAS-2 question (“Was there an appropriate interval between index tests and reference standard?”) to keep replacement of the term *reference standard* with *outcome* consistent throughout the tool.
S4.4: Was information on the outcome available for all participants?	Modified QUADAS-2 signaling question	This question was modified from a QUADAS-2 question (“Did all patients receive a reference standard?”) to better account for any participants lost during follow-up and missing outcome data.
Domain 5: Analysis		
	New domain Source: QUIPS, PROBAST	This domain is a new addition from QUADAS-2. Given the complexities of analyzing longitudinal studies, we included this in QUAPAS, adopting signaling questions from QUIPS/PROBAST.
S5.1: Were all enrolled participants included in the analysis?	Modified QUADAS-2 signaling question	This signaling question was redirected from the flow and timing domain of QUADAS-2 to the current domain because the question addresses an analytic issue.
S5.2: If data were missing, were appropriate methods used?	Added signaling question Source: PROBAST	We added a signaling question to address appropriate handling of missing data, adopted from a signaling question in PROBAST (“Were participants with missing data handled appropriately?”).
S5.3: Were appropriate methods used to account for censoring?	Added signaling question Source: PROBAST	This signaling question was based on a question from PROBAST (“Were complexities in the data [e.g., censoring, competing risks, sampling of control participants] accounted for appropriately?”). Censoring can occur in longitudinal data for various reasons, and appropriate methods should be applied.
S5.4: In case of competing events, were appropriate	Added signaling question Source: PROBAST	The same signaling question from PROBAST used for S5.3 was split into a second question to address competing events. Competing events occur when

methods used to account for them?	participants can experience other events, preventing occurrence of the outcome of interest.

PF = prognostic factor; PROBAST = Prediction Model Risk of Bias Assessment Tool; QUADAS-2 = Quality Assessment of Diagnostic Accuracy Studies 2; QUAPAS = Quality Assessment of Prognostic Accuracy Studies; QUIPS = Quality in Prognosis Studies.

Domain 2: Index Test

Risk of bias: Could the conduct or interpretation of the index test have introduced bias?

This domain covers potential sources of bias related to the definition, measurement, or interpretation of the index test.

Applicability: Are there concerns that the index test or its conduct, interpretation, or threshold differs from the review question?

Applicability concerns can arise if there are variations in the test version, method of measurement, or interpretation. Factors related to the index test should be consistent with the review question because variability may influence performance estimates.

Changes

Signaling question 2.1: Was the method used to perform the index test valid and reliable?
This signaling question, based on a prompting item from QUIPS (“Method of [prognostic factor] measurement is … valid and reliable” [10]), was added to flag bias that may arise from index test measurement error. A study should use an adequately validated assay or other measurement method that is reliable, reproducible, and fit for the intended use. Consistent and valid measurement methods can minimize unwanted heterogeneity in index test measurements (16).

Signaling question 2.2: Was the method for performing the index test the same for all participants?
We added this signaling question based on a prompting item from QUIPS (“The method and setting of measurement of [prognostic factors] is the same for all study participants” [10]) and signaling question from PROBAST (“Were predictors defined and assessed in a similar way for all participants?” [11]). The index test should be executed consistently for each participant. Accuracy may fluctuate on the basis of the index test procedures—for

example, related to variability between imaging technologies or with insufficiently standardized assays for blood-based biomarkers (17).

Domain 3: Outcome

Risk of bias: Could measurement of the outcome have introduced bias?

This domain refers to bias that may arise from the definition, method of measurement, or interpretation of the outcome.

Applicability: Are there concerns that the outcome does not match the review question?

Concerns about applicability arise when the outcome definition or measurement methods differ between the study and the review question.

Changes

The signaling questions in this domain had only minor modifications, all of which were based on questions from QUADAS-2 (Table 1).

Domain 4: Flow and Timing

Risk of bias: Could the study flow have introduced bias?

This domain focuses on the inclusion of participants in the analysis, any participants who received treatment between the index test and occurrence of the outcome, and the time horizon.

Applicability: Are there concerns that the time horizon does not match the review question?

Concerns about applicability surface if the time horizon (period between index test measurement and occurrence of the event) differs between the primary study and the review question.

Table 2. The QUAPAS Tool*

Domain	Participants	Index Test	Outcome	Flow and Timing	Analysis
Description	Describe methods for recruiting participants Describe participants (previous testing, presentation, intended use of index test, and setting)	Describe the index test (definition, context of use, method of measurement, and interpretation)	Describe the outcome (definition, method of measurement, and interpretation)	Describe any participants lost to follow-up or excluded from the analysis Describe the time horizon from the index test to the outcome	Describe the statistical methods
Signaling questions (yes, no, unclear)	S1.1: Was a consecutive or random sample of participants enrolled?† S1.2: Was a case–control design avoided?‡ S1.3: Did the study avoid inappropriate selection criteria?†	S2.1: Was the method used to perform the index test valid and reliable?§ S2.2: Was the method for performing the index test the same for all participants?§‖ S2.3: Were the index test results interpreted without knowledge of the outcome?† S2.4: If a threshold was used, was it prespecified?‡	S3.1: Was the method used to measure the outcome valid and reliable?† S3.2: Was the method for measuring the outcome the same for all participants?† S3.3: Was the outcome measured without knowledge of the index test results?†	S4.1: Did all participants receive the index test?‖ S4.2: Was treatment avoided after the index test was performed?¶ S4.3: Was the time horizon sufficient to capture the outcome?† S4.4: Was information on the outcome available for all participants?†	S5.1: Were all enrolled participants included in the analysis?† S5.2: If data were missing, were appropriate methods used?‖ S5.3: Were appropriate methods used to account for censoring?‖ S5.4: In case of competing events, were appropriate methods used to account for them?‖
Risk of bias (high, low, unclear)	Could the selection of participants have introduced bias?	Could the conduct or interpretation of the index test have introduced bias?	Could measurement of the outcome have introduced bias?	Could the study flow have introduced bias?	Could the analysis have introduced bias?
Concerns about applicability (high, low, unclear)	Are there concerns that the participants do not match the review question?	Are there concerns that the index test or its conduct, interpretation, or threshold differs from the review question?	Are there concerns that the outcome does not match the review question?	Are there concerns that the time horizon does not match the review question?	–

QUAPAS = Quality Assessment of Prognostic Accuracy Studies.

* QUAPAS is a tool for evaluating risk of bias in and applicability of prognostic accuracy studies. It follows the domain-based structure of QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2). Five key domains are rated in terms of risk of bias; 4 are also assessed for concerns about applicability to the review question. Each domain has a set of signaling questions to help judge risk of bias as low, high, or unclear. All signaling questions are phrased in such a way that “yes” indicates low risk of bias. Answer “unclear” if relevant information was not reported. Any signaling question answered as “no” flags potential for bias.

† Signaling question from QUADAS-2 redirected to a different domain and/or modified to better reflect evaluation of tests used in prognosis settings, ‡ Signaling question intact from QUADAS-2, § Prompting item and consideration from QUIPS (Quality in Prognosis Studies) added as a signaling question, || Signaling question added from PROBAST (Prediction Model Risk of Bias Assessment Tool), ¶ New signaling question

Changes

Signaling question 4.1: Did all participants receive the index test?
We included this question, adopted from PROBAST ("Are all predictors available at the time the model is intended to be used?" [11]), to account for any participants who did not receive the index test. Systematic differences between those who did and did not receive the index test indicate potential bias in longitudinal studies (12, 18).

Signaling question 4.2: Was treatment avoided after the index test was performed?
This signaling question was newly added to highlight potential for treatment selection bias. Using routinely collected data is sometimes the pragmatic approach for prognostic accuracy studies. Baseline clinical differences can result in a subset of participants receiving treatment, affecting the likelihood of the outcome thereafter (19). Clinical management should ideally be identical for all study participants during the follow-up period.

Domain 5: Analysis

Risk of bias: Could the analysis have introduced bias?

We incorporated a fifth domain, as in QUIPS and PROBAST, to capture complexities introduced from time-dependent analysis in longitudinal studies. This domain does not exist in the QUADAS-2 tool. Here, the aim is to guide the reviewer in judging whether results are likely to be biased by analytic decisions. There is no applicability assessment for this domain.

Signaling question 5.1: Were all enrolled participants included in the analysis?
This signaling question was redirected from the flow and timing domain of QUADAS-2. Participants may be excluded from the analysis for various reasons, such as inconvenient index test or outcome results (for example, uninterpretable or intermediate results). Omitting eligible participants who are systematically different from those included in the analysis can introduce biased performance estimation (12).

Signaling question 5.2: If data were missing, were appropriate methods used?
We added this signaling question, adopted from a signaling question in PROBAST ("Were participants with missing data handled appropriately?" [11]), to flag bias that arises from neglecting missing data when appropriate methods should be applied. The risk of bias due to missing data increases with the magnitude of missingness (12, 20). Therefore, small percentages of missing data, or no systematic difference between those with missing versus complete data, would indicate low risk of bias (21).

Multiple imputation of missing data is widely advocated as an improvement over complete-case analysis because it has been shown to minimize bias and produce correct SEs (22). Test results can be missing for structural reasons—for example, because of a failure to complete the testing procedure, or incomplete analysis of samples. Such cases should not be ignored, but how they should be handled will depend on context. Depending on the clinical consequences, they may be analyzed as negative results, positive results, or a distinct test result category, without imputation (23).

Signaling question 5.3: Were appropriate methods used to account for censoring?
This signaling question was based on a question from PROBAST ("Were complexities in the data [e.g., censoring, competing risks, sampling of control participants] accounted for appropriately?" [11]). Censoring can occur in longitudinal data for various reasons: The follow-up period may end without a participant having the outcome, or a participant may be lost to follow-up or experience another event (24). Analysis of such data requires appropriate methods (25), such as time-specific versions of sensitivity and specificity with cumulative cases and dynamic controls, time-dependent receiver-operating characteristic curve analysis, or time-varying hazard ratios (26, 27).

Signaling question 5.4: In case of competing events, were appropriate methods used to account for them?
We included an additional question to address handling of competing events, based on a PROBAST question ("Were complexities in the data [e.g., censoring, competing risks, sampling of control participants] accounted for appropriately?" [11]). Competing events may preclude the occurrence of the main event or outcome of interest (28). For example, investigators evaluating a test for predicting cancer should consider other incidents that may occur before a participant can receive a cancer diagnosis, such as death from other causes. Ignoring this in the analysis can produce overestimations of prognostic

performance (29).

Such methods as a Cox proportional hazards (cause-specific) model, cumulative incidence function, or other related competing risk model may be applied (30-32). Simple Kaplan–Meier and Cox regression methods may overestimate the probabilities of outcomes and should be avoided (29, 31, 33).

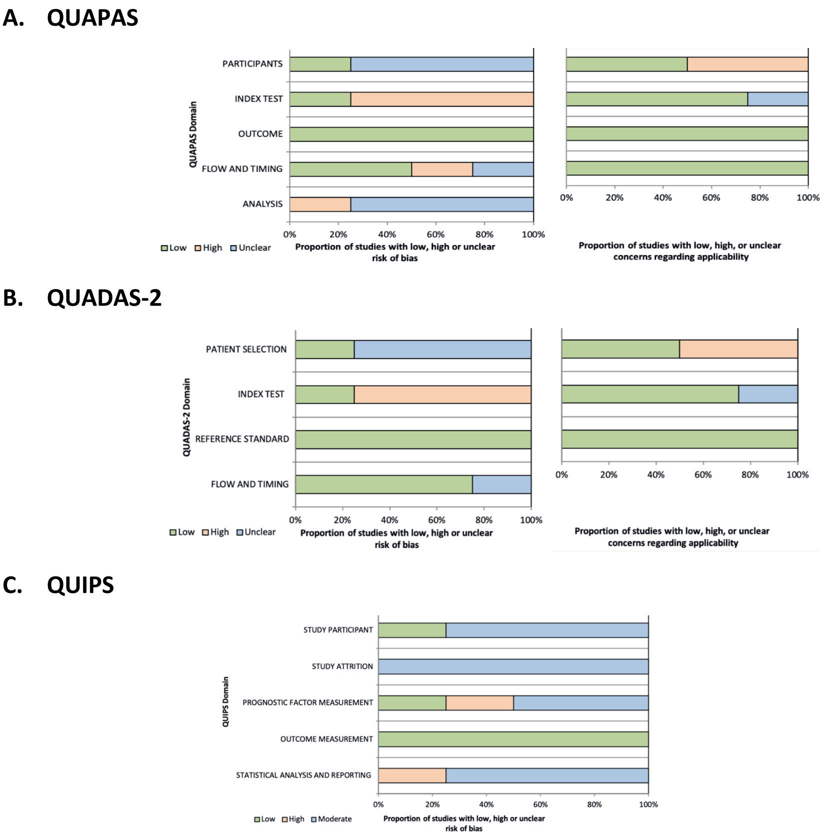
Application of QUAPAS

Two reviewers (J.L. and F.M.) independently assessed studies included in a published systematic review of prognostic accuracy studies using the final version of QUAPAS (Table 2), after which no further changes were deemed necessary. This evaluation was done in parallel with QUADAS-2 and QUIPS.

The systematic review aimed to evaluate the ability of a blood-based score, called the Fibrosis-4 score, to prognosticate changes in liver fibrosis stage or mortality. The review included 4 studies for predicting changes in fibrosis (34-37) and 6 for predicting mortality (38-43). We note that the version of QUAPAS used in the published systematic review of prognostic accuracy was a pilot version, before finalization of the tool (44).

The reviewers created figures to summarize the judgments by modifying the QUADAS-2 template, available at www.quadas.org, which includes the tabular presentation of judgments for each study and graphical summary of all studies. The Figure shows the fibrosis example assessment with QUAPAS, QUADAS-2, and QUIPS. Part 3 of the Supplement illustrates assessment of mortality as the outcome.

In contrast to QUADAS-2, the addition of signaling question 4.2 (“Was treatment avoided after the index test was performed?”) in QUAPAS flagged a study at risk of treatment selection bias, influencing the overall judgment of bias in the flow and timing domain for this study. In the new analysis domain, 1 study had high risk of bias because missing data issues were not addressed, in addition to absence of methods for censoring or accounting for competing events. Three remaining studies had unclear risk (34-36) due to sparse reporting of analytic methods. Using the new analysis domain, QUAPAS was able to flag 1 study at high risk of bias in that domain, which would otherwise have gone unnoticed by QUADAS-2.



A. QUAPAS

B. QUADAS-2

C. QUIPS

Figure 1. Graphical display of risk of bias and applicability judgments for index test (Fibrosis-4) in prognosticating changes in liver fibrosis stages in 4 studies using QUAPAS (top), QUADAS-2 (middle), and QUIPS (bottom). QUADAS-2 = Quality Assessment of Diagnostic Accuracy Studies 2; QUAPAS = Quality Assessment of Prognostic Accuracy Studies; QUIPS = Quality in Prognosis Studies.

We further shared QUAPAS with and invited written feedback from 10 researchers and potential end users with varying levels of experience to assess if the items of the tool were appropriate and clear to the targeted objective of highlighting sources of potential bias and applicability concerns.

Users found the steps for using the tool and signaling questions clear. They further found footnotes in the user template helpful for answering signaling questions. On the basis of the response, we implemented minor changes to improve the clarity of the footnotes.

Discussion

Systematic reviews of prognostic accuracy research synthesize empirical evidence and enable clinicians to make informed decisions on risk stratification and disease management. Assessment of risk of bias and concerns about the applicability of primary studies is an essential step because it may influence the interpretation of a review’s findings. The QUAPAS tool was developed specifically for evaluating prognostic accuracy studies, using the widely used QUADAS-2 as a starting point and mapping relevant items from QUIPS and PROBAST. We introduce a systematically tailored tool, which can minimize any challenges or variability that can arise when leaving tailoring up to the discretion of each user.

Prognostic accuracy research is still a young and rapidly evolving field. Critical assessment of prognostic accuracy studies remains a challenge, even for researchers with expertise, when studies do not clearly report all necessary information. We encourage primary study investigators to adhere to existing reporting checklists, such as REMARK (Reporting Recommendations for Tumor Marker Prognostic Studies) (45), TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) (46), or STARD (Standards for the Reporting of Diagnostic Accuracy Studies) (47). Adherence to these guidelines will enable reproducibility and facilitate assessment of risk of bias and applicability concerns until a specific reporting guideline is developed. Such a guideline for prognostic accuracy studies could be developed, because primary study investigators will likely encounter challenges that are analogous to reviewers using QUIPS, PROBAST, or QUADAS-2 for evaluating prognostic accuracy studies.

The process of developing QUAPAS presented some limitations. We did not use a Delphi approach for developing this tool and instead followed a more pragmatic approach because our aim was to tailor QUADAS-2 for use in prognostic accuracy studies. The items of QUAPAS were selected from 3 existing risk-of-bias tools on the basis of a consensus procedure. We relied on the knowledge and experience of experts in the field to arrive at the final set of signaling questions, in the absence of methodological systematic reviews.

QUAPAS is intended to be used for a single pair of index test and outcome, as is the case for QUADAS-2. Prognostic accuracy studies, much like their diagnostic counterpart, are often done in the form of comparative accuracy studies, where the performance of several tests is evaluated simultaneously. Such designs can introduce their own set of biases when conducted improperly. QUADAS-C (Quality Assessment of Diagnostic Accuracy Studies–Comparative) was recently introduced as an extension for assessing comparative diagnostic accuracy studies (48). A similar addition could be developed for QUAPAS for comparative prognostic accuracy studies.

We believe that QUAPAS can help future systematic reviewers and readers in assessing risk of bias and applicability in prognostic accuracy studies. In providing the reviewing community with a systematically tailored tool, we hope to improve the quality assessment process and help produce a more robust evidence base for prognostic tests.

Acknowledgment: The authors thank Jill Hayden, PhD (Dalhousie University, Faculty of Medicine), for her valuable advice in the earlier stages of planning the development of QUAPAS. They further thank the volunteers for reviewing the tool.

References

1. Group F-NBW. BEST (Biomarkers, EndpointS, and other Tools) resource. 2016.
2. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ. 1986;134(6):587-94.
3. Raja AS, Greenberg JO, Qaseem A, Denberg TD, Fitterman N, Schuur JD. Evaluation of patients with suspected acute pulmonary embolism: best practice advice from the Clinical Guidelines Committee of the American College of Physicians. Annals of internal medicine. 2015;163(9):701-11.
4. Hemingway H. Prognosis research: why is Dr. Lydgate still waiting? Journal of clinical epidemiology. 2006;59(12):1229-38.
5. Freund Y, Lemachatti N, Krastinova E, Van Laer M, Claessens Y-E, Avondo A, et al. Prognostic Accuracy of Sepsis-3 Criteria for In-Hospital Mortality Among Patients With Suspected Infection Presenting to the Emergency Department. JAMA. 2017;317(3):301-8.
6. Cardoso FS, Ricardo LB, Oliveira AM, Canena JM, Horta DV, Papoila AL, et al. C-reactive protein prognostic accuracy in acute pancreatitis: timing of measurement and cutoff points. European Journal of Gastroenterology & Hepatology. 2013;25(7):784-9.
7. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. Journal of clinical epidemiology. 2003;56(11):1118-28.
8. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. Journal of the National Cancer Institute. 2008;100(20):1432-8.
9. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36.
10. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing Bias in Studies of Prognostic Factors. Annals of Internal Medicine. 2013;158(4):280-6.
11. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170(1):51-8.
12. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Annals of internal medicine. 2019;170(1):W1-W33.
13. Halligan S, Boone D, Bhatnagar G, Ahmad T, Bloom S, Rodriguez-Justo M, et al. Prognostic biomarkers to identify patients destined to develop severe Crohn's disease who may benefit from early biological therapy: protocol for a systematic review, meta-analysis and external validation. Systematic reviews. 2016;5(1):1-9.
14. Reynolds JC, Issa MS, Nicholson TC, Drennan IR, Berg KM, O'Neil BJ, et al. Prognostication with point-of-care echocardiography during cardiac arrest: a systematic review. Resuscitation. 2020;152:56-68.
15. Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HC. Prognosis research: toward evidence-based results and a Cochrane methods group. J Clin Epidemiol. 2007;60(8):863-5; author reply 5-6.
16. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? Bmj. 2009;338.
17. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, et al. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK). JNCI: Journal of the National Cancer Institute. 2005;97(16):1180-4.
18. Tin Tin S, Woodward A, Ameratunga S. Estimating bias from loss to follow-up in a prospective cohort study of bicycle crash injuries. Injury Prevention. 2014;20(5):322.
19. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of Observational Studies in the Presence of Treatment Selection BiasEffects of Invasive Cardiac Management on AMI Survival Using Propensity Score and Instrumental Variable Methods. JAMA. 2007;297(3):278-85.
20. Altman DG, Bland JM. Missing data. BMJ. 2007;334(7590):424-.

21. Mack C, Su Z, Westreich D. Managing missing data in patient registries: addendum to registries for evaluating patient outcomes: a user's guide. 2018.
22. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International journal of methods in psychiatric research. 2011;20(1):40-9.
23. Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. Bmj. 2013;346:f2778.
24. Leung K-M, Elashoff RM, Afifi AA. CENSORING ISSUES IN SURVIVAL ANALYSIS. Annual Review of Public Health. 1997;18(1):83-104.
25. Altman DG, Bland JM. Time to event (survival) data. BMJ. 1998;317(7156):468-9.
26. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Medical Research Methodology. 2017;17(1):53.
27. Bansal A, Heagerty PJ. A comparison of landmark methods and time-dependent ROC methods to evaluate the time-varying performance of prognostic markers for survival outcomes. Diagnostic and Prognostic Research. 2019;3(1):14.
28. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016;133(6):601-9.
29. Wolbers M, Koller MT, Witteman JC, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology. 2009:555-61.
30. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. Int J Epidemiol. 2012;41(3):861-70.
31. Verduijn M, Grootendorst DC, Dekker FW, Jager KJ, le Cessie S. The analysis of competing events like cause-specific mortality--beware of the Kaplan-Meier method. Nephrol Dial Transplant. 2011;26(1):56-61.
32. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;94(446):496-509.
33. Noordzij M, Leffondré K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? Nephrology Dialysis Transplantation. 2013;28(11):2670-7.
34. Vilar-Gomez E, Calzadilla-Bertot L, Friedman SL, Gra-Oramas B, Gonzalez-Fabian L, Lazo-Del Vallin S, et al. Serum biomarkers can predict a change in liver fibrosis 1 year after lifestyle intervention for biopsy-proven NASH. Liver Int. 2017;37(12):1887-96.
35. Siddiqui MS, Yamada G, Vuppalanchi R, Van Natta M, Loomba R, Guy C, et al. Diagnostic Accuracy of Noninvasive Fibrosis Models to Detect Change in Fibrosis Stage. Clin Gastroenterol Hepatol. 2019;17(9):1877-85 e5.
36. McPherson S, Hardy T, Henderson E, Burt AD, Day CP, Anstee QM. Evidence of NAFLD progression from steatosis to fibrosing-steatohepatitis using paired biopsies: implications for prognosis and clinical management. J Hepatol. 2015;62(5):1148-55.
37. Chalasani N, Abdelmalek MF, Loomba R, Kowdley KV, McCullough AJ, Dasarathy S, et al. Relationship between three commonly used non-invasive fibrosis biomarkers and improvement in fibrosis stage in patients with non-alcoholic steatohepatitis. Liver Int. 2019;39(5):924-32.
38. Angulo P, Bugianesi E, Bjornsson ES, Charatcharoenwitthaya P, Mills PR, Barrera F, et al. Simple noninvasive systems predict long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology. 2013;145(4):782-9 e4.
39. Boursier J, Vergniol J, Guillet A, Hiriart JB, Lannes A, Le Bail B, et al. Diagnostic accuracy and prognostic significance of blood fibrosis tests and liver stiffness measurement by FibroScan in non-alcoholic fatty liver disease. J Hepatol. 2016;65(3):570-8.
40. Hagstrom H, Nasr P, Ekstedt M, Stal P, Hultcrantz R, Kechagias S. Accuracy of Noninvasive Scoring Systems in Assessing Risk of Death and Liver-Related Endpoints in Patients With Nonalcoholic Fatty Liver Disease. Clin Gastroenterol Hepatol. 2019;17(6):1148-56 e4.
41. Onnerhag K, Hartman H, Nilsson PM, Lindgren S. Non-invasive fibrosis scoring systems can predict future metabolic complications and overall mortality in non-alcoholic fatty liver disease (NAFLD). Scand J Gastroenterol. 2019;54(3):328-34.

42. Peleg N, Sneh Arbib O, Issachar A, Cohen-Naftaly M, Braun M, Shlomai A. Noninvasive scoring systems predict hepatic and extra-hepatic cancers in patients with nonalcoholic fatty liver disease. PLoS One. 2018;13(8):e0202393.
43. Xun YH, Guo JC, Lou GQ, Jiang YM, Zhuang ZJ, Zhu MF, et al. Non-alcoholic fatty liver disease (NAFLD) fibrosis score predicts 6.6-year overall mortality of Chinese patients with NAFLD. Clin Exp Pharmacol Physiol. 2014;41(9):643-9.
44. Lee J, Vali Y, Boursier J, Spijker R, Anstee QM, Bossuyt PM, et al. Prognostic accuracy of FIB-4, NAFLD fibrosis score, and APRI for NAFLD-related events: a systematic review. Liver International. 2019;n/a(n/a).
45. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. BMC Medicine. 2012;10(1):51.
46. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Annals of Internal Medicine. 2015;162(1):W1-W73.
47. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. Clinical Chemistry. 2015;61(12):1446-52.
48. Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF, et al. QUADAS-C: A Tool for Assessing Risk of Bias in Comparative Diagnostic Accuracy Studies. Annals of Internal Medicine. 2021.

Supplementary Material

https://www.acpjournals.org/doi/suppl/10.7326/M22-0276

06

Biomarkers for staging fibrosis and non-alcoholic steatohepatitis in non-alcoholic fatty liver disease (the LITMUS project): a comparative diagnostic accuracy study

§Jenny Lee §Yasaman Vali Jerome Boursier
Salvatore Petta Kristy Wonders Dina Tiniakos
Pierre Bedossa Andreas Geier Sven Francque
Mike Allison Georgios Papatheodoridis
Helena Cortez-Pinto Raluca Pais Jean-Francois Dufour
Diana Julie Leeming Stephen Harrison Yu Chen
Jeremy Cobbold Michael Pavlides Adriaan G. Holleboom
Hannele Yki-Jarvinen Javier Crespo Morten Karsdal
Rachel Ostroff Mohammad Hadi Zafarmand
Richard Torstenson Kevin Duffin Carla Yunis Clifford Brass
Mattias Ekstedt Guruprasad P Aithal Jörn M. Schattenberg
Elisabetta Bugianesi Manuel Romero-Gomez Vlad Ratziu
*Quentin M. Anstee *Patrick M. Bossuyt

§ Joint First Authors
* Joint Senior & Corresponding Authors

A slightly modified version was accepted for publication
The Lancet Gastroenterology & Hepatology, 2023.

Abstract

Background: We evaluated the accuracy of seventeen biomarkers and multi-marker scores to detect non-alcoholic steatohepatitis (NASH) or stage liver fibrosis in the LITMUS Metacohort, an international cohort of biopsy-confirmed NAFLD patients.

Methods: Accuracy was expressed as the area under the ROC curve (AUC), using liver histology as the reference standard, and compared against the Fibrosis-4 Index for Liver Fibrosis (FIB-4) in the same subgroup. Target conditions were at-risk NASH (NAS≥4 and F≥2) and advanced fibrosis (F≥3). We identified thresholds for each biomarker for reducing the number of liver biopsy-based screen failures when recruiting patients with at-risk NASH for future trials.

Findings: Data from 966 adult patients were included; 335 (35%) had at-risk NASH and 271 (28%) advanced fibrosis. For at-risk NASH, no single biomarker or multi-marker score significantly reached the predefined AUC 0.80 acceptability threshold, with accuracy mostly comparable to that of FIB-4. Performance in detecting advanced fibrosis was better; SomaSignal, ADAPT and liver stiffness measurement significantly reached acceptable accuracy. With several markers, histological screen failure rates could be reduced to one-third in future trials if only marker-positive patients underwent biopsy for evaluating eligibility. Best screening performance was observed for SomaSignal, followed by ADAPT, MACK-3, and PRO-C3.

Interpretation: None of the single markers or multi-marker scores achieved an acceptable AUC for replacing biopsy in detecting patients with at-risk NASH. Several biomarkers could be applied in a pre-screening strategy for identifying at-risk NASH patients in clinical trial recruitment. The performance of promising markers will be further evaluated in the ongoing prospective LITMUS study cohort.

Introduction

Non-alcoholic fatty liver disease (NAFLD), a leading cause of chronic liver disease, spans a histological spectrum from steatosis to non-alcoholic steatohepatitis (NASH) with progressive hepatic fibrosis leading to cirrhosis and/or hepatocellular carcinoma in a subset of patients. [1] This progressive disorder is predicted to become more prevalent in the next decade, consuming substantial healthcare resources and posing a growing public health challenge. [1]

Despite its high prevalence, accurate diagnosis and delivery of effective management of NAFLD remain challenging, with generally poor public health readiness internationally. This failure is, at least in part, due to a lack of clarity on biomarker performance in detecting at-risk NASH, deterring their adoption by clinicians.[2]

NAFLD patients with advanced fibrosis (F≥3) are at higher risk of adverse liver-related outcomes, liver transplantation, and death. [3] International guidelines recommend assessment of NAFLD patients for early identification of high stages of liver fibrosis (F≥3). [4] In the absence of approved pharmacological therapies, identification of patients with NASH and clinically significant fibrosis (F≥2) is essential to support recruitment into therapeutic clinical trials.

The current reference standard for detecting NASH and staging fibrosis is liver histology. Liver biopsy sampling is invasive, resource-intensive, prone to sampling error, and carries a small but appreciable risk of complications. [5] Despite debates regarding its limitations, participants recruited to NAFLD trials require biopsy to qualify for enrolment. [5,6] There is, therefore, an urgent need for non-invasive biomarkers to support clinical care and facilitate the evaluation of new therapies.

Several non-invasive biomarkers have been proposed, with variable performance in detecting fibrosis and NASH. Few studies have compared these biomarkers in a single cohort to evaluate their relative performances. The LITMUS (Liver Investigation: Testing Marker Utility in Steatohepatitis) consortium [7] is evaluating the performance of seventeen non-invasive single biomarkers, multi-marker scores and vibration-controlled transient elastography (VCTE) in identifying patients with at-risk NASH and advanced fibrosis, with liver biopsy as the reference standard. We also aimed to increase the efficiency of future drug trial enrolment by selecting thresholds for each marker that lead to an acceptable screen failure rate in identifying eligible participants.

Methods

Study design and participants

This was a comparative diagnostic accuracy study in biopsy-confirmed NAFLD patients from thirteen countries across Europe. Data were collected in the LITMUS Metacohort of the European NAFLD Registry, an international cohort of NAFLD patients prospectively recruited following standardized procedures and monitoring; see Hardy and Wonders et al. for details. [8] The recruitment period was from 2010 to 2019. Patients were required to provide informed consent prior to inclusion. Studies contributing to the Metacohort were approved by the relevant Ethical Committees in the participating countries and conform to the guidelines of the Declaration of Helsinki.

Adults aged ≥18 years with suspected NAFLD and paired liver biopsy and serum samples were eligible for inclusion in this analysis. All patients met pre-defined inclusion/exclusion criteria [8] and had undergone a liver biopsy as part of the routine diagnostic workup for presumed NAFLD, for example, having originally been identified due to abnormal biochemical tests (ALT and/or gamma-glutamyltransferase) and/or an ultrasonographically detected bright liver, associated with features of the metabolic syndrome. Patients with excessive alcohol consumption (>20-30 g/day) or evidence of other chronic liver diseases, such as viral hepatitis B or C, were excluded.

Clinical assessment

Detailed clinical data were collected from all participants by a trained investigator and entered directly into a central registry. Body mass index (BMI) was calculated by dividing weight (kg) by height (meters) squared.

Clinical laboratory blood assays were performed in laboratories of the respective recruitment centres. Lipid (LDL, HDL, cholesterol, triglyceride (TG)) and liver profiles (platelet count, alanine aminotransferase (ALT), aspartate aminotransferase (AST), and gamma glutamyl transferase (GGT)) were collected. Comorbidities such as dyslipidaemia (fasting TG level ≥150 mg/dL [1.7 mmol/L]; or fasting HDL <40 mg/dL [1.03 mmol/L] in males and <50 mg/dL [1.29 mmol/L] in females; or on treatment), hypertension (systolic blood pressure ≥130 or diastolic pressure ≥85 mmHg), and diabetes (fasting glucose >7.0 mmol/L) were also captured.

Liver biopsy

Liver biopsy samples were considered adequate and were histologically examined locally in each centre by expert liver pathologists, prospectively following clinical work-up. [9] NAFLD activity was assessed according to the NASH Clinical Research Network, steatosis and lobular inflammation were scored on four-point scales (0 to 3); ballooning was scored on a three-point scale (0 to 2). [10] Liver fibrosis was graded on a 5 point scale (0 to 4) according to Kleiner et al.[10].

Biomarker measurements

All serum samples were collected in standardized collection kits and processed locally before storage at -80°C, according to prespecified biobanking standard operating procedures. Samples were shipped in batches on dry ice from recruitment sites to the LITMUS Central Biobank, where serum samples were catalogued and subsequently sent for central analysis at Nordic Biosciences, a College of American Pathologists accredited laboratory. Only serum samples collected within six months of liver biopsy were eligible for this analysis.

The following biomarkers were measured in the central LITMUS laboratory: CK-18 M30 (M30 Apoptosense ELISA no. 10011, VLVbio), CK-18 M65 (M65 EpiDeath ELISA no. 10040, VLVbio), PRO-C3 (ELISA based) [11], PRO-C4 (ELISA based) [12], and PRO-C6 (ELISA based) [13]. All measurements were performed blinded to all clinical data associated with the samples. Due to differences in available sample volumes, not all biomarkers could be measured in every participant.

In addition, liver stiffness measurement (LSM) and controlled attenuation parameter (CAP) by VCTE (FibroScan, Echosens, Paris, France™) collected within six months of liver biopsy were evaluated. Probe sizes were selected as advised by device guidelines.

Multi-marker scores

Using the available clinical laboratory data, NFS, FIB-4, and APRI were calculated using their originally published formulas [14-16]. The following nine previously reported multi-marker scores were also calculated: MACK-3 (HOMA, AST, CK-18 M30), [17] Cao 2013 (ALT, platelet count, CK-18 M30 and TG), [18] ADAPT (age, platelet count, diabetes, PRO-C3), [19] FIBC3 (age, BMI, diabetes, platelet count, PRO-C3), [20] ABC3D (age, BMI, diabetes, platelet count, PRO-C3), [20] NFS (age, BMI, IFG/diabetes, AST/ALT ratio, platelet count, albumin), [14] and APRI (AST, platelet count), [16]

ELF test scores, based on hyaluronic acid, tissue inhibitor of matrix metalloproteinase-1, and aminoterminal propeptide of procollagen type III, were measured in the Central Laboratory (Siemens Advia Centaur). The SomaSignal serum tests for fibrosis, steatosis, inflammation, and ballooning, were assayed at the UK SomaLogic facility.

The SomaSignal NASH tests are modified aptamer-based elastic net logistic regression models trained and validated against biopsy for each component (steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis) and contain 12, 14, 5, and 8 protein analytes, respectively. The tests were developed as dichotomized protein-phenotype models for clinically relevant severity of steatosis (NAS score 0 vs 1-3), hepatocellular ballooning (0 vs 1-2), lobular inflammation (0-1 vs 2-3) and fibrosis (stages 0-1 vs 2-4). [21,22] We multiplied the model probabilities for steatosis, inflammation, and ballooning as a SomaSignal marker for NASH.

Target Conditions

Significant fibrosis was defined as F≥2 and advanced fibrosis as F≥3. The NAFLD activity score (NAS), the sum of the steatosis, lobular inflammation, and ballooning, scored according to the NASH CRN system, ranged from 0 to 8. NASH was defined as the presence of steatosis, lobular inflammation, and hepatocellular ballooning. This was operationalized in accordance with standard clinical trial practice as a NAS score of ≥4 with at least one point in each component. [23-26]

The main target condition was the combination of significant fibrosis and NASH, referred to as at-risk NASH. This combination has been defined by Health Authorities (FDA, EMA, CDE) as the critical inclusion criterion in phase 3 drug development for treatment of noncirrhotic NASH with liver fibrosis. In addition, we evaluated the performance in detecting advanced fibrosis.

Statistical analysis

Non-parametric, empirical receiver operating characteristic (ROC) curves were constructed for each biomarker and multi-marker score. Diagnostic accuracy was expressed in terms of the area under this receiver operating characteristic curve (AUC) with its 95% confidence interval (95% CI), calculated using the DeLong method. [27] To be considered as a diagnostic marker of acceptable accuracy, an AUC of at least 0.80 in detecting at-risk NASH was expected. [28].

Recruiting 966 participants, of whom an anticipated 35% have the target condition, would give us at least 80% power to reject the null hypothesis that the AUC does not exceed the minimally acceptable value of 0.80 if the actual AUC is 0.85 or more, and at least 99% power if the actual AUC is 0.87 or more, at a 5% type I error rate. [29]

The performance of the FIB-4 score was calculated for comparison. Due to the absence of any existing validated non-invasive test for NASH and the high collinearity between NASH and fibrosis stage, performance of FIB-4, a widely used simple fibrosis test, was adopted as a comparator in all target conditions. To account for differences in the subgroups for which marker results were available, we recalculated the AUC for FIB-4 in each of the corresponding marker subgroups.

In an additional head-to-head direct comparison, we evaluated the performance of the most clinically available biomarkers and scores (PRO-C3, CK-18 M30 and M65, ELF, NFS, APRI, ADAPT, FIBC3, ABC3D, and FIB-4) in a subgroup in which results for all ten were available. We additionally evaluated the subgroup specific performance of all markers according to diabetes status (R-package ROCnReg).

We aimed to identify an optimal cut-off level for each biomarker as a screening test to identify patients with at-risk NASH. The optimal cut-off would allow a hypothetical trial enrolment screen failure rate, based on liver biopsy, not exceeding 33% (one in three). This value was selected based on a survey conducted among clinicians and drug developers within the LITMUS consortium. The biopsy screen failure rate corresponds to one minus the positive predictive value (PPV).

Given the differences between the subgroups in which biomarker results were available, we calculated the minimally required likelihood ratio for a positive biomarker result to yield the corresponding screen failure rate and PPV, based on a single measure of the prevalence, using the following formula:

$$LR = \frac{1-prev}{prev} \times \frac{PPV}{1-PPV}$$

in which *prev* denotes the prevalence of at-risk NASH in the population undergoing testing.

Based on the proportion with at-risk NASH in the Metacohort study group, we estimated the prevalence of at-risk NASH to be 35%. To achieve a screen failure rate not exceeding 33%, the likelihood ratio of a positive biomarker result would then have to be at least 3.77. Of all positivity thresholds with a likelihood ratio exceeding 3.77, we selected the one with the highest sensitivity, thereby maximizing efficiency at the preselected acceptable screen failure rate.

We report sensitivity, specificity, proportion of positive biomarker results (at the 35% prevalence), true positive fraction (proportion of potential study participants with a biomarker positive result found to have at-risk NASH at biopsy) and number needed to test to find one eligible trial participant after liver biopsy in test positives (the inverse of the true positive fraction). Confidence intervals were based on 10,000 bootstrap samples. All statistical analyses were performed using R statistical computing software version 4.2.1 (Vienna, Austria).

The manuscript is reported according to STARD guidelines (Supplementary Table S1).

Results

Study group characteristics

Data from 966 Metacohort participants were included in the analysis (see Figure 1). Their mean age was 51 years, 58% were men, and the majority were of white ethnicity (90%). In the group of participants with at-risk NASH, we observed higher liver enzymes, higher proportion of patients with hypertension (69%) and diabetes (64%) at baseline compared to those without at-risk NASH (Table 1).

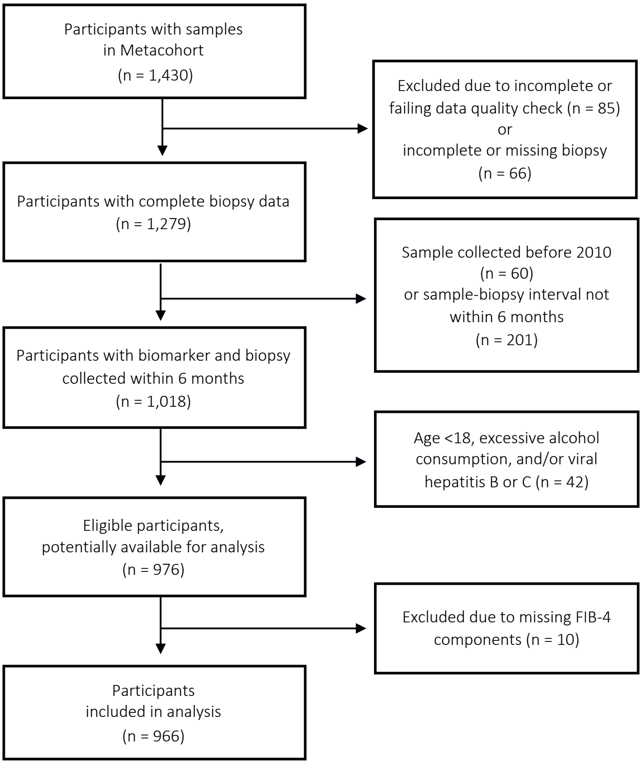

Participants with samples in Metacohort (n = 1,430)
Excluded due to incomplete or failing data quality check (n = 85) or incomplete or missing biopsy (n = 66)
Participants with complete biopsy data (n = 1,279)
Sample collected before 2010 (n = 60) or sample-biopsy interval not within 6 months (n = 201)
Participants with biomarker and biopsy collected within 6 months (n = 1,018)
Age <18, excessive alcohol consumption, and/or viral hepatitis B or C (n = 42)
Eligible participants, potentially available for analysis (n = 976)
Excluded due to missing FIB-4 components (n = 10)
Participants included in analysis (n = 966)

Figure 1. Flow diagram of participants included in the analysis

In this study group, 35% had at-risk NASH (335/966) and 28% (271/966) had advanced fibrosis. The distribution of the NAS score was: 0 (2%), 1 (7%), 2 (12%), 3 (19%), 4 (23%), 5 (19%), 6 (12%), 7 (4%), 8 (<1%). For fibrosis the distribution was: F0 (32%), F1 (19%), F2 (20%), F3 (20%), F4 (9%).

Diagnostic performance in detecting at-risk NASH

The mean time interval between biopsy and blood sampling was less than one week. We evaluated the performance of each biomarker in its respective subgroup for which results were available (Table 2; Supplementary Table S2).

Table 1. Characteristics of the Metacohort study group

	Overall **(n=966)**	**At-risk NASH** **(n=335)**	**No at-risk NASH** **(n=631)**
Age, years	51.2 (13.0)	55.0 (12.1)	49.2 (13.0)
Female	403 (43%)	152 (45%)	251 (40%)
BMI, kg/m²	34.12 (8.26)	33.86 (6.56)	34.25 (9.04)
Diabetes	406 (42%)	215 (64%)	191 (30%)
ALT, U/L	62.67 (42.54)	70.57 (44.97)	58.47 (40.62)
AST, U/L	42.88 (26.01)	51.89 (30.11)	38.10 (22.13)
GGT, U/L	108.08 (154.61)	121.77 (142.86)	101.07 (159.95)
Albumin, g/L	43.8 (4.2)	43.5 (3.9)	44.0 (4.3)
Platelet count, 10^9/L	238.96 (73.45)	228.41 (69.41)	244.57 (74.95)
Glucose, mmol/L	6.42 (2.47)	7.15 (2.99)	6.06 (2.08)
Triglycerides, mg/L	2.08 (1.18)	2.31 (1.37)	1.95 (1.04)
Fibrosis stage			
0	309 (32%)	-	309 (49%)
1	185 (19%)	-	185 (29%)
2	199 (21%)	131 (39%)	68 (11%)
3	188 (19%)	144 (43%)	44 (7%)
4	85 (9%)	60 (18%)	25 (4%)

All continuous variables are expressed as mean (SD)

Single biomarkers

No single marker reached the prespecified 0.80 AUC threshold. Performance comparable to that of FIB-4 was observed for most biomarkers.

Multi-marker scores

The best performing multi-marker score was the SomaSignal test, with an AUC of 0.81, but the confidence interval still included 0.80. FIB-4 had an AUC of 0.66 in the corresponding subgroup. MACK-3 and ADAPT had an AUC around 0.77 versus 0.69 and 0.73 for FIB-4.

Diagnostic screening for recruiting at-risk NASH trial participants

Table 3 presents the optimal threshold for each marker, corresponding to a failure rate not exceeding 33% while maximizing sensitivity. For most evaluated biomarkers and scores, it was possible to define such a threshold, but the corresponding proportion testing positive varied widely, as a result of the differences in the underlying distributions of marker results in those with and those without at-risk NASH. No acceptable threshold could be identified for the ABC3D, APRI, ELF, NFS, FIB-4, or CAP-VCTE.

Single biomarkers

The single biomarker with the highest proportion testing positive and, consequently, the highest proportion of patients with a true positive result was PRO-C3. At a threshold of 24.05 ng/ml, 17 per 100 would test positive and qualify for a screening biopsy, of which 11 would then be true positives. This means that for every nine patients undergoing PRO-C3 testing, one eligible patient with biopsy-proven at-risk NASH would be recruited.

In contrast, the optimal diagnostic screening threshold for PRO-C4 was 433.35 ng/ml. At that threshold, 6 per 100 undergoing testing would have a positive test result and could be selected for biopsy to evaluate trial eligibility. Of these six, four would have at-risk NASH while two would not. This also corresponds to a 33% failure rate, but at a lower efficiency: 23 patients would have to undergo PRO-C4 testing to find one eligible trial participant with biopsy-confirmed at-risk NASH.

Table 2. Diagnostic Accuracy of single biomarkers and multi-marker scores compared to FIB-4 in the same subgroup.

Marker	n	At-risk NASH			Advanced Fibrosis		
		Perc	AUC Marker	AUC FIB-4	Perc	AUC Marker	AUC FIB-4
CK-18 M30	795	35%	0.69 (0.65-0.73)	0.70 (0.66-0.73)	28%	0.70 (0.66-0.74)	0.79 (0.75-0.82)
CK-18 M65	817	34%	0.70 (0.66-0.74)	0.69 (0.65-0.73)	28%	0.70 (0.66-0.74)	0.79 (0.75-0.82)
PRO-C3	444	36%	0.68 (0.63-0.74)	0.73 (0.68-0.78)	28%	0.75 (0.70-0.80)	0.76 (0.71-0.81)
PRO-C6	229	41%	0.68 (0.61-0.75)	0.70 (0.63-0.77)	36%	0.71 (0.63-0.78)	0.73 (0.66-0.80)
PRO-C4	391	40%	0.63 (0.57-0.68)	0.72 (0.67-0.77)	31%	0.66 (0.60-0.71)	0.75 (0.70-0.81)
NFS	933	35%	0.66 (0.62-0.69)	0.69 (0.66-0.73)	28%	0.75 (0.72-0.79)	0.77 (0.74-0.81)
APRI	966	35%	0.68 (0.64-0.71)	0.69 (0.66-0.73)	28%	0.72 (0.68-0.75)	0.77 (0.74-0.81)
ELF	919	33%	0.67 (0.63-0.71)	0.68 (0.65-0.72)	27%	0.80 (0.76-0.83)	0.77 (0.74-0.81)
SomaSignal	264	46%	0.81 (0.75-0.86)	0.66 (0.60-0.73)	36%	0.90 (0.86-0.94)	0.72 (0.66-0.79)
MACK-3	538	34%	0.76 (0.71-0.80)	0.69 (0.64-0.73)	24%	0.74 (0.69-0.79)	0.76 (0.71-0.80)
Cao 2013	635	37%	0.67 (0.63-0.72)	0.69 (0.65-0.73)	30%	0.68 (0.64-0.73)	0.79 (0.75-0.83)
ADAPT	444	36%	0.77 (0.73-0.81)	0.73 (0.68-0.78)	28%	0.85 (0.81-0.89)	0.76 (0.71-0.81)
FIBC3	440	36%	0.74 (0.69-0.79)	0.73 (0.68-0.78)	28%	0.82 (0.78-0.87)	0.76 (0.71-0.81)
ABC3D	440	36%	0.74 (0.69-0.79)	0.73 (0.68-0.78)	28%	0.81 (0.76-0.85)	0.76 (0.71-0.81)
LSM-VCTE	632	40%	0.74 (0.70-0.78)	0.66 (0.62-0.71)	30%	0.83 (0.80-0.86)	0.73 (0.70-0.78)
CAP-VCTE	263	48%	0.61 (0.54-0.67)	0.66 (0.60-0.73)	35%	0.61 (0.54-0.69)	0.71 (0.65-0.78)

At-risk NASH defined as (NASH and F≥2); advanced fibrosis as F≥3; cells indicate estimated area under the Receiver Operating Characteristic curve (AUC) and corresponding 95% confidence interval; Perc: percentage with target condition in corresponding subgroup.

Thresholds correspond to a liver biopsy screen failure rate of 33% at a 35% prevalence. Markers are ranked based on the number of patients with biopsy-confirmed at-risk NASH found per 100 patients tested with the marker, if liver biopsy is restricted to marker positives only. Confidence intervals based on bootstrapping. No acceptable threshold was found for ABC3D, APRI, ELF, NFS, FIB-4, or CAP-VCTE.

Multi-marker scores

The best performing screening tests were the SomaSignal test, ADAPT and MACK-3. With these tests, 35, 24, and 21 per 100 patients, respectively, would test positive at the selected thresholds and would undergo biopsy, and 24, 16, and 14 eligible patients would be true positives for at-risk NASH. The highest sensitivity was observed for the SomaSignal test (0.67).

At a different prevalence of at-risk NASH, the optimal thresholds and proportions will be different. Figure 2 shows the number of test positives and the number of true positives for four of the markers and scores included in the head-to-head comparison at various levels of prevalence.

Table 3. Thresholds for identifying at-risk NASH in diagnostic screening.

Marker	Threshold	Sensitivity	Specificity	Number of positive patients undergoing biopsy (Per 100)	Number of eligible patients found (Per 100)	Number needed to test
SomaSignal	0.06	0.67 (0.59 – 0.75)	0.82 (0.59 – 0.75)	35 (30 – 40)	24 (20 – 26)	4 (4 – 5)
ADAPT	6.91	0.47 (0.39 – 0.55)	0.88 (0.83 – 0.91)	24 (21 – 28)	16 (14 – 19)	6 (5 – 7)
MACK-3	0.53	0.41 (0.34 – 0.48)	0.89 (0.85 – 0.92)	21 (19 – 25)	14 (12 – 17)	7 (6 – 8)
PRO-C3	24.05 ng/ml	0.33 (0.25 – 0.40)	0.92 (0.88 – 0.94)	17 (14 – 20)	11 (9 – 14)	9 (7 – 11)
FIBC-3	0.84	0.28 (0.21 – 0.35)	0.93 (0.89 – 0.96)	14 (11 – 18)	10 (7 – 12)	10 (8 – 14)
LSM-VCTE	16.4 kPa	0.26 (0.21 – 0.32)	0.93 (0.90 – 0.95)	14 (11 – 16)	9 (7 – 11)	11 (9 – 14)
CK-18 M30	573.80 IU/L	0.25 (0.20 – 0.30)	0.93 (0.91 – 0.95)	13 (11 – 15)	9 (7 – 11)	11 (9 – 14)
Cao 2013	1.74	0.22 (0.17 – 0.28)	0.94 (0.92 – 0.96)	12 (9 – 14)	8 (6 – 10)	13 (10 – 16)
PRO-C6	14.25 ng/ml	0.18 (0.11 – 0.26)	0.96 (0.91 – 0.98)	9 (6 – 13)	6 (4 – 9)	16 (11 – 26)
PRO-C4	433.35 ng/ml	0.12 (0.08 – 0.18)	0.97 (0.94 – 0.99)	6 (4 – 9)	4 (3 – 6)	23 (16 – 37)
CK-18 M65	1283.55 IU/L	0.12 (0.09 – 0.16)	0.97 (0.95 – 0.98)	6 (5 – 8)	4 (3 – 6)	24 (17 – 33)
No marker	-	-	-	100	35	-

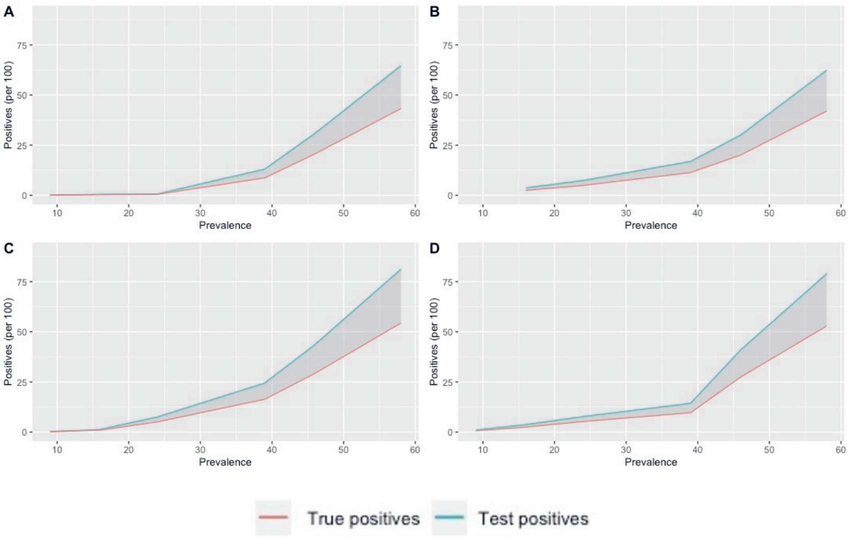


Figure 2. Proportion of true positives and test positives at varying levels of prevalence, at a threshold corresponding to a 33% screen failure rate (A) CK18M30 (B) PROC-3, (C) ADAPT, and (D) FIBC-3. Blue line: patients testing positive; Red line: Proportion of patients with at-risk NASH confirmed by biopsy.

Diagnostic performance in detecting advanced fibrosis

Single biomarkers

For detecting advanced fibrosis, only LSM-VCTE significantly reached the predefined 0.80 threshold, with an AUC of 0.83, compared to 0.73 for FIB-4. (Figure 3B)

Multi-marker scores

Five different multi-marker scores exceeded the 0.80 AUC threshold in detecting advanced fibrosis, but only two did so significantly. The SomaSignal test had an AUC of 0.90, versus 0.72 for FIB-4. The 0.85 AUC for the ADAPT score was also significantly higher than the threshold. LSM, FIBC3, ABC3D, and ELF had AUC of 0.83, 0.82, 0.81, and 0.80, respectively. (Figure 3B)

Results from ten circulating biomarker and multi-marker scores (PRO-C3, CK-18 M30 and M65, ELF, NFS, APRI, ADAPT, FIBC3, ABC3D, and FIB-4) were used for a direct head-to-head comparison in 335 participants. In this subgroup, 38% had at-risk NASH and 29%

advanced fibrosis (see Supplementary Table S3). We observed AUCs for detecting at-risk NASH and advanced fibrosis to be similar in this subgroup to the ones in the main analysis. (Supplementary Figure S1, S2, and Table S4).

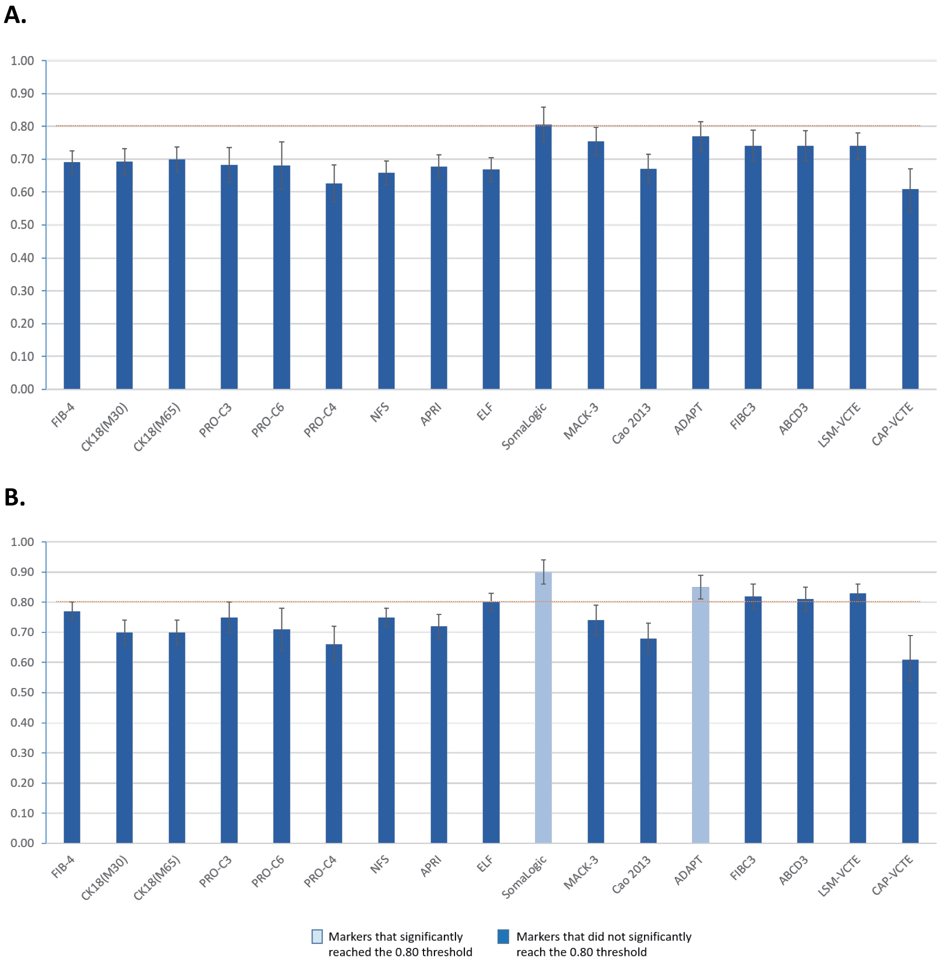


Figure 3. Diagnostic accuracy of single biomarkers and multi-marker scores in detecting (A) at-risk NASH and (B) advanced fibrosis.

Subgroup analysis

The performance of each marker was evaluated separately in those with and without diabetes. In our study group 42% had diabetes. In detecting at-risk NASH, performance was marginally lower in for those with diabetes, although only significantly so for the multi-marker score ADC3D, with AUCs of 0.74 (no diabetes) versus 0.56 (diabetes) (Supplementary Table S5). In detecting advanced fibrosis, there were no significant differences between the two subgroups, with comparable AUC.

Discussion

In this comparative diagnostic accuracy study, we used data and samples collected in the LITMUS Metacohort to evaluate the performance of several markers in identifying NAFLD patients with at-risk NASH (NASH & F≥2) or those with advanced fibrosis (F≥3), using liver histology as the reference standard. Based on the ROC analyses, none of the evaluated single biomarkers met our prespecified 0.80 threshold in detecting at-risk NASH. Of the multi-marker scores, best performance was observed for the SomaSignal test, comprised of 35 different proteins. AUC values were higher for detecting advanced fibrosis. Here the SomaSignal test, the ADAPT score, and LSM-VCTE significantly exceeded our prespecified 0.80 AUC threshold.

Recruitment for clinical trials is at present based on liver biopsy, and screening for patients with at-risk NASH is limited due to high screen failure rates for histological assessment. A successful screening biomarker would be expected to identify most of the at-risk NASH patients while significantly reducing the number of patients requiring biopsy. We proposed a strategy for pre-selecting those who would undergo liver biopsy by targeting a screen failure rate not exceeding 33%. Using this strategy, we observed that some tests would substantially reduce the number needed to undergo liver biopsy with acceptable sensitivity, as only marker positives would require further evaluation.

Without a screening biomarker, all 966 patients would require biopsy to identify the 335 at-risk NASH patients. So, patient selection efficiency would be 335 out of 966 (35%). The best performing biomarker assessed in this study, SomaSignal, would reduce the number of patients requiring biopsy by 65%, from 966 tested to 338 biopsied, resulting in 232 identified at-risk NASH patients. Many of the other biomarkers measured in this study would similarly increase patient selection efficiency but with lower sensitivity, resulting in a lower identification rate compared to the SomaSignal test. We note that the 33% screen failure rate was defined based on expert opinion, and others may arrive at a different acceptable proportion based on factors such as feasibility and costs. The thresholds identified here should be externally validated, as several factors such as disease spectrum may affect the performance of the tests in diagnostic screening for trial recruitment.

A major strength of the study was the centralized measurement of all novel biomarkers instead of the use of local, historical measurements, although measured in batches. The analysis was performed by an independent group of expert epidemiologists with no vested interest in demonstrating superior performance of any test. We provide comparative accuracy data for a wide selection of both staple fibrosis tests and newer developments proposed for NAFLD, which can supplement guideline development for their suggested use in the future. Limitations also need to be acknowledged. Stability of these markers is not well understood, which is why we did not include samples collected before 2010. Histological scoring was not centralized and variability in recognition of elementary lesions or composite diagnoses might have occurred [30]. Histology-based semi-quantitative scoring is an imperfect reference standard, limiting the accuracy of grading necroinflammatory activity and staging fibrosis [31]. Another limitation was the retrospective collection of biological samples with inherent exhaustion of sample material, restricting the measurement of most biomarkers in different subsets of patients.

Some of these biomarkers were not originally proposed for identifying at-risk NASH. Since many are available to clinicians, we decided, pragmatically, to explore their diagnostic performance for this key histological aspect as well. Various markers [32] and multi-marker scores [24,33] have been specifically developed for the diagnosis of at-risk NASH and will need to be compared with the best performing biomarkers from the current study. The analysis presented here focuses on serum-based biomarkers and multi-marker scores. Although, we included additional analysis on VCTE, a non-invasive technology proposed to evaluate liver aetiologies, other non-invasive imaging technologies to evaluate liver aetiologist have been proposed and should be further studied. We further note the influence of recruiting patients from mostly tertiary care centres in multiple countries. Factors such as differences in prevalence, epidemiology, referral patterns, and clinical work-up leading up to biopsy may affect the generalizability of our findings to other settings.

The performance of many markers in this study was comparable to findings recently published. Collagen-based markers and scores analysed from participants enrolled in the CENTAUR phase IIb trial showed that the single marker PRO-C3 performed marginally worse than FIB-4 in detecting advanced fibrosis, while the ADAPT score had a higher AUC. [34] Incorporating a direct marker of fibrosis in the algorithm resulted in an improvement from simple scores, such as APRI. The ELF test had performance levels consistent with those presented in a meta-analysis, which reported a summary AUC of 0.83 for identifying advanced fibrosis. [35] For at-risk NASH, Chuah et al. concluded MACK-3 (AUC of 0.80) had comparable performance to FIB-4 (0.82) and outperformed single markers like CK-18 (0.72). [36] In comparison to another large meta-analysis, the CK-18 M30 antigens demonstrated consistent AUCs for at-risk NASH (0.73). [37]

When interpreting contrasting results between studies or subgroups, spectrum effect should be considered. [38] Test performance often varies across population subgroups, as can be seen in the varying AUC estimates for FIB-4 in the partially overlapping subgroups in our analysis. The performance of NAFLD markers to correctly identify patients with advanced fibrosis will vary with the relative proportions of patients with F0 fibrosis and F4 fibrosis in the study group. Having a higher number of patients with F0 fibrosis and/or F4 fibrosis will increase the performance of markers in discriminating between those with and without advanced fibrosis. There is a clear difference between the distribution in our study group, which probably represents the one typically seen in secondary and tertiary centres, and that in the recently reported NIMBLE stage 1-NASH CRN study, which had equal numbers in the five fibrosis stage subgroups. [39]

The limited performance of biomarkers in detecting at-risk NASH provides a mandate for further study of novel biomarker algorithms, adopting both hypothesis-driven approaches founded in pathophysiology and machine learning approaches. Such approaches should be tailored to a specific target condition and context of use.

The ultimate utility of these or any other biomarkers would be their ability to predict clinical outcomes. The longitudinal outcome data currently being generated by LITMUS within the Europe NAFLD Registry will be an important asset for evaluating their prognostic value. [5,40]

We conducted one of the largest comparative diagnostic accuracy studies, with seventeen different non-invasive markers for NAFLD. The results from the present study showed that none of the single biomarkers achieved the desired level of performance to replace liver histology in detecting patients with at-risk NASH. However, some multi-marker scores, such as the SomaSignal test and ADAPT, are promising tools for identifying advanced fibrosis. Of note, no biomarkers have been approved by FDA or EMA, which further highlights the urgency of the LITMUS consortium's aim to validate and advance toward regulatory qualification markers for NAFLD and NASH. The LITMUS project will continue to collect data in the prospective LITMUS Study cohort and will perform analysis of blood-based and imaging biomarkers to further facilitate the evaluation of new and existing interventions in trials and to improve the clinical care and outcomes of NAFLD patients.

References

1. Riazi K, Azhari H, Charette JH, et al. The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. *The Lancet Gastroenterology & Hepatology* 2022.
2. Lazarus JV, Anstee QM, Hagström H, et al. Defining comprehensive models of care for NAFLD. *Nature reviews Gastroenterology & hepatology* 2021; 18(10): 717-29.
3. Taylor RS, Taylor RJ, Bayliss S, et al. Association Between Fibrosis Stage and Outcomes of Patients with Non-Alcoholic Fatty Liver Disease: a Systematic Review and Meta-Analysis. *Gastroenterology* 2020.
4. Francque SM, Marchesini G, Kautz A, et al. Non-alcoholic fatty liver disease: A patient guideline. *JHEP Reports* 2021; 3(5): 100322.
5. Davison BA, Harrison SA, Cotter G, et al. Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. *Journal of hepatology* 2020; 73(6): 1322-32.
6. Food U, Administration D. Nonalcoholic steatohepatitis with compensated cirrhosis: developing drugs for treatment guidance for industry. 2019. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/nonalcoholic-steatohepatitis-compensated-cirrhosis-developing-drugs-treatment-guidance-industry (accessed August 2022).
7. LITMUS: Liver Investigation: Testing Marker Utility in Steatohepatitis consortium (European Union IMI2-funded under Grant Agreement 777377). https://www.imi.europa.eu/projects-results/project-factsheets/litmus. (accessed August 2022).
8. Hardy T, Wonders K, Younes R, et al. The European NAFLD Registry: A real-world longitudinal cohort study of nonalcoholic fatty liver disease. *Contemp Clin Trials* 2020; 98: 106175.
9. Bedossa P, Consortium FP. Utility and appropriateness of the fatty liver inhibition of progression (FLIP) algorithm and steatosis, activity, and fibrosis (SAF) score in the evaluation of biopsies of nonalcoholic fatty liver disease. *Hepatology* 2014; 60(2): 565-75.
10. Kleiner DE, Brunt EM, Van Natta M, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* 2005; 41(6): 1313-21.
11. Nielsen MJ, Nedergaard AF, Sun S, et al. The neo-epitope specific PRO-C3 ELISA measures true formation of type III collagen associated with liver and muscle parameters. *American journal of translational research* 2013; 5(3): 303.
12. Leeming DJ, Nielsen MJ, Dai Y, et al. Enzyme-linked immunosorbent serum assay specific for the 7S domain of collagen type IV (P4NP 7S): a marker related to the extracellular matrix remodeling during liver fibrogenesis. *Hepatology Research* 2012; 42(5): 482-93.
13. Sun S, Henriksen K, Karsdal MA, et al. Collagen type III and VI turnover in response to long-term immobilization. *PLoS One* 2015; 10(12): e0144525.
14. 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(4): 846-54.
15. 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(6): 1317-25.
16. Wai C-T, 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(2): 518-26.
17. Boursier J, Anty R, Vonghia L, et al. Screening for therapeutic trials and treatment indication in clinical practice: MACK-3, a new blood test for the diagnosis of fibrotic NASH. *Alimentary Pharmacology & Therapeutics* 2018; 47(10): 1387-96.
18. Cao W, Zhao C, Shen C, Wang Y. Cytokeratin 18, alanine aminotransferase, platelets and triglycerides predict the presence of nonalcoholic steatohepatitis. *PLoS One* 2013; 8(12): e82092.
19. Daniels SJ, Leeming DJ, Eslam M, et al. ADAPT: An Algorithm Incorporating PRO-C3 Accurately Identifies Patients With NAFLD and Advanced Fibrosis. *Hepatology* 2019; 69(3): 1075-86.
20. Boyle M, Tiniakos D, Schattenberg JM, et al. Performance of the PRO-C3 collagen neo-epitope biomarker in non-alcoholic fatty liver disease. *JHEP Reports* 2019; 1(3): 188-98.
21. Rachel O, Leigh A, Stephen W. A liquid liver biopsy: Serum protein patterns of liver steatosis, inflammation, hepatocyte ballooning and fibrosis in NAFLD and NASH. 2020. https://assets.website-

files.com/5f3d77cd56d46907a50fb8d9/5f9d9c2057efc43f55b78db7_2020%20TLMdX%20Late-breaking%20Abstracts-%20Oct%2030.pdf 2022).

22. Williams SA, Ostroff R, Hinterberg MA, et al. A proteomic surrogate for cardiovascular outcomes that is sensitive to multiple mechanisms of change in risk. *Science Translational Medicine* 2022; 14(639): eabj9625.

23. Woreta TA, Van Natta ML, Lazo M, et al. Validation of the accuracy of the FAST™ score for detecting patients with at-risk nonalcoholic steatohepatitis (NASH) in a North American cohort and comparison to other non-invasive algorithms. *PloS one* 2022; 17(4): e0266859.

24. Newsome PN, Sasso M, Deeks JJ, et al. FibroScan-AST (FAST) score for the non-invasive identification of patients with non-alcoholic steatohepatitis with significant activity and fibrosis: a prospective derivation and global validation study. *The lancet Gastroenterology & hepatology* 2020; 5(4): 362-73.

25. Corey KE, Pitts R, Lai M, et al. ADAMTSL2 protein and a soluble biomarker signature identify at-risk non-alcoholic steatohepatitis and fibrosis in adults with NAFLD. *Journal of Hepatology* 2022; 76(1): 25-33.

26. Ratziu V, Magnanensi J, Deledicque S, et al. IN TYPE 2 DIABETIC PATIENTS, THE IDENTIFICATION OF AT-RISK NASH IS IMPACTED BY AGE: A COMPARISON OF SERUM-BASED NITS INCLUDING NIS4®. JOURNAL OF HEPATOLOGY; 2021: ELSEVIER RADARWEG 29, 1043 NX AMSTERDAM, NETHERLANDS; 2021. p. S586-S7.

27. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988: 837-45.

28. LITMUS LITMUiS. LITMUS Grant Agreement No. 777377, 2018.

29. Obuchowski NA, Lieber ML, Wians Jr FH. ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clinical chemistry* 2004; 50(7): 1118-25.

30. Brunt EM, Clouston AD, Goodman Z, et al. Complexity of ballooned hepatocyte feature recognition: Defining a training atlas for artificial intelligence-based imaging in NAFLD. *Journal of hepatology* 2022; 76(5): 1030-41.

31. Mehta SH, Lau B, Afdhal NH, Thomas DL. Exceeding the limits of liver histology markers. *Journal of hepatology* 2009; 50(1): 36-41.

32. Harrison SA, Ratziu V, Boursier J, et al. A blood-based biomarker panel (NIS4) for non-invasive diagnosis of non-alcoholic steatohepatitis and liver fibrosis: a prospective derivation and global validation study. *The lancet Gastroenterology & hepatology* 2020; 5(11): 970-85.

33. Noureddin M, Truong E, Gornbein JA, et al. MRI-based (MAST) score accurately identifies patients with NASH and significant fibrosis. *Journal of hepatology* 2021.

34. Nielsen MJ, Leeming DJ, Goodman Z, et al. Comparison of ADAPT, FIB-4 and APRI as non-invasive predictors of liver fibrosis and NASH within the CENTAUR screening population. *J Hepatol* 2021; 75(6): 1292-300.

35. Vali Y, Lee J, Boursier J, et al. Enhanced liver fibrosis test for the non-invasive diagnosis of fibrosis in patients with NAFLD: A systematic review and meta-analysis. *Journal of hepatology* 2020; 73(2): 252-62.

36. Chuah KH, Wan Yusoff WNI, Sthaneshwar P, Nik Mustapha NR, Mahadeva S, Chan WK. MACK-3 (combination of hoMa, Ast and CK18): A promising novel biomarker for fibrotic non-alcoholic steatohepatitis. *Liver Int* 2019; 39(7): 1315-24.

37. Lee J, Vali Y, Boursier J, et al. Accuracy of cytokeratin 18 (M30 and M65) in detecting non-alcoholic steatohepatitis and fibrosis: A systematic review and meta-analysis. *Plos one* 2020; 15(9): e0238717.

38. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med* 2002; 137(7): 598-602.

39. Sanyal AJ, Shankar SS, Yates K, et al. Primary results of the nimble stage 1-NASH CRN study of circulating biomarkers for nonalcoholic steatohepatitis and its activity and fibrosis stage. AASLD Meeting, November 12–15; 2021; USA: EMJ Hepatology 10 [Supplement 1] 2022 2021. p. 1383A-A.

40. Brunt EM, Clouston AD, Goodman Z, et al. Complexity of ballooned hepatocyte feature recognition: Defining a training atlas for artificial intelligence-based imaging in NAFLD. *Journal of Hepatology* 2022.

Supplementary Material

https://www.thelancet.com/journals/langas/article/PIIS2468-1253(23)00017-1/fulltext#supplementaryMaterial

07

Machine learning algorithm improves detection of NASH (NAS-based) and at-risk NASH, a development and validation study

Jenny Lee Max Westphal Yasaman Vali
Jerome Boursier Salvatore Petta Rachel Ostroff
Leigh Alexander Yu Chen Celine Fournier
Andreas Geier Sven Francque Kristy Wonders
Dina Tiniakos Pierre Bedossa Mike Allison
Georgios Papatheodoridis Helena Cortez-Pinto
Raluca Pais Jean-Francois Dufour Diana Julie Leeming
Stephen Harrison Jeremy Cobbold Adriaan G. Holleboom
Hannele Yki-Järvinen Javier Crespo Mattias Ekstedt
Guruprasad P Aithal Elisabetta Bugianesi
Manuel Romero-Gomez. Richard Torstenson Morten Karsdal
Carla Yunis Jörn M. Schattenberg Detlef Schuppan
Vlad Ratziu Clifford Brass Kevin Duffin Koos Zwinderman
Michael Pavlides §Quentin M. Anstee §Patrick M Bossuyt

§ Joint Senior Authors

Slightly modified version accepted for publication
Hepatology, 2023, 78.1: 258-271.

Abstract

Background & Aims: Detecting non-alcoholic steatohepatitis (NASH) remains challenging, while at-risk NASH (steatohepatitis and F≥2) tends to progress and is of interest for drug development and clinical application. We developed prediction models by supervised machine learning (ML) techniques, with clinical data and biomarkers to stage and grade non-alcoholic fatty liver disease (NAFLD) patients.

Approach & Results: Learning data were collected in the LITMUS Metacohort (966 biopsy-proven NAFLD adults), staged and graded according to NASH-CRN. Conditions of interest were clinical trial definition of NASH (NAS≥4;53%), at-risk NASH (NASH with F≥2;35%), significant (F≥2;47%) and advanced fibrosis (F≥3;28%). Thirty-five predictors were included. Missing data were handled by multiple imputation. Data were randomly split into training/validation (75/25) sets. Gradient boosting machine (GBM) was applied to develop two models for each condition: clinical versus extended (clinical and biomarkers). Two variants of the NASH and at-risk NASH models were constructed: direct and composite models.

Clinical GBM models for steatosis/inflammation/ballooning had AUCs of 0.94/0.79/0.72. There were no improvements when biomarkers were included. The direct NASH model produced AUCs (clinical/extended) of 0.61/0.65. The composite NASH model performed significantly better (0.71) for both variants. The composite at-risk NASH model had an AUC of 0.83 (clinical and extended), an improvement over the direct model. Significant fibrosis models had AUCs (clinical/extended) of 0.76/0.78. The extended advanced fibrosis model (0.86) performed significantly better than the clinical version (0.82).

Conclusions: Detection of NASH and at-risk NASH can be improved by constructing independent ML models for each component, using only clinical predictors. Adding biomarkers only improved accuracy for fibrosis.

Background

Non-alcoholic fatty liver disease (NAFLD) is characterized by fat accumulation in hepatocytes. There is a need for more robust and accessible non-invasive tests (NITs), as NAFLD affects nearly 25% of the global population (1, 2). As a progressive condition, NAFLD ranges from isolated steatosis (liver fat content ≥5%) to non-alcoholic steatohepatitis (NASH) with or without fibrosis and cirrhosis (3, 4). NASH is associated with progression to liver fibrosis and hepatocellular carcinoma (5). “At-risk” NASH (NASH with at least significant fibrosis) is an important target for drug development and the focus of health authorities, as it carries an increased risk of liver-related mortality and contributes significantly to the total burden of hepatocellular carcinoma (6). In a prospective cohort study, the population with fibrosis stage 3 and higher had the greatest risk to develop liver endpoints, while fibrosis stage 2 and higher was linked to increased hepatic and extrahepatic morbidity (7).

Liver biopsy remains the reference standard for a definitive NASH diagnosis, however, the procedure carries risks to the patient and has several inherent limitations including sampling error and reader variability (8, 9). Even so, no NITs for NASH that match similar standards are available. This unmet clinical need has been the driving force for a marathon of research to develop and validate novel NITs that can distinguish patients with a greater likelihood of disease progression than those with comparable liver biopsy performance. Identifying those at higher risk is critical for risk-stratification, monitoring, and expediting recruitment for NASH clinical trials.

The list of NITs for NAFLD fibrosis has rapidly grown, with the Liver Stiffness Measurement by Vibration-Controlled Transient Elastography (LSM by VCTE), Enhanced Liver Fibrosis (ELF) test and Fibrosis-4 (FIB-4) score recommended to rule out advanced fibrosis (10, 11). However, the EASL Clinical Practice Guidelines currently do not recommend NITs for diagnosis of NASH (10). Extensively studied biomarkers such as caspase-cleaved cytokeratin-18 (CK-18) fragments and full length soluble CK-18 show suboptimal performance, although combining CK-18 with synergistic markers showed some improvement (12). Multivariable models developed using regression-based techniques, such as FIC-22 (13), the NAFLD diagnostic panel (14), or the NASH test (15), have either proved to be less effective in more extensive multicenter studies or have not undergone sufficient external validation. More recently, the MACK-3, FAST, and NIS-4 scores were developed specifically for detecting at-risk NASH (16-18).

While the list of NITs for NAFLD grows, few were developed based on machine-learning algorithms, which are probably more suitable for handling complicated diseases with multifaceted etiology. Simple regression-based methods rely heavily on statistical assumptions, which do not always hold true for real-world data, whereas model-free machine learning algorithms adapt to data characteristics with fewer assumptions.

Machine learning uses algorithms to learn associations, identify patterns, and create predictions from complex data structures, which can provide opportunities for improving the diagnosis or prognosis of diseases. More recently, machine learning has been applied to develop diagnostic scores across multiple disciplines, offering a potential solution for developing tools for conditions that prove more difficult to detect (19-21). Our aim was to employ machine learning to develop diagnostic models for detecting clinical trial definition of NASH, at-risk NASH, and significant and advanced fibrosis, first by utilizing only routinely collected clinical data and second by adding biomarkers.

Material and Methods

This manuscript was prepared using the TRIPOD guidelines (Supplementary Table 2) (22).

Study participants (LITMUS Metacohort)

We analyzed data from 966 participants in the LITMUS Metacohort. These participants were recruited from 12 centers in 9 countries across Europe, between 2010 and 2019 and include adults with biopsy-confirmed NAFLD with available clinical, laboratory and biomarker data within 6 months of biopsy. Serum samples drawn within 6 months of biopsy and stored at -80^{0}C were also available. Details of the study can be found elsewhere (23). All participants provided informed consent prior to inclusion; the cohort studies were approved by the relevant ethics committees in the participating countries.

Liver biopsy

Biopsy samples were examined prospectively in each center by expert liver pathologists. NAFLD activity was graded according to the NASH Clinical Research Network (NASH CRN) (24). Liver fibrosis was graded on a 5-point scale (0 to 4), denoted as *F* in the following.

NASH is comprised of three components: steatosis, lobular inflammation, scored on four-point scales (0-3), and ballooning, on a three-point scale (0-2) according to the NASH CRN classification (24). The NAFLD activity score (NAS), the unweighted sum of steatosis, lobular inflammation and ballooning scores thus ranges from 0 to 8.

Target conditions

This study addressed four target conditions:

i. **Significant fibrosis**: Defined as F≥2;

ii. **Advanced fibrosis**: Defined as F≥3;

iii. **Clinical Trial NASH**: Steatohepatitis is a histopathological diagnosis based on the presence of steatosis, lobular inflammation and hepatocyte ballooning (25). For inclusion in therapeutic trials, the FDA and EMA mandate steatohepatitis is defined as a NAS≥4 with at least a score of 1 point for each histological component (26), thus selecting patients with greater disease activity that are considered more likely to exhibit disease progression (27); and

iv. **"At-risk" NASH:** Like the above, "at-risk" NASH is defined as the presence of steatohepatitis (NAS≥4 with at least one point in each component) plus significant fibrosis (F≥2) (2, 17, 18). This defines the population commonly recruited into phase 3 trials of novel therapeutics for noncirrhotic NASH.

Predictors

Clinical assessment

Clinical data, including anthropometric, lifestyle/activity, dietary, comorbidity, pharmacotherapy, clinical biochemistry, and incident disease/events, were collected in the respective recruitment centers, with blood assays performed in local laboratories. The list of 25 clinical predictors used is shown in Supplementary Table 3.

Biomarker measurements

Additional serum samples were collected in standardized collection kits within 6 months of liver biopsy and stored at -80^{0}C. Samples were centrally analyzed at Nordic Biosciences (Herlev, Denmark), a CLIA certified laboratory, blinded to clinical data. The following markers were measured and included as predictors: caspase-cleaved CK-18 fragments and full length soluble CK18 (M30 and M65 antigens), serum peptides that represent the aminoterminal propeptide of procollagentype III (PRO-C3), and the carboxyterminal propeptides of procollagen type IV (PRO-C4) and VI (PRO-C6). We further include the components of the Siemen's ELF test: tissue inhibitor of metalloproteinases 1 (TIMP-1), amino-terminal propeptide of type III procollagen (P3NP) and hyaluronic acid (HA).

LSM and controlled attenuation parameter (CAP) by VCTE (FibroScan, Echosens, Paris, France™) were collected within 6 months of liver biopsy and also included as predictors. Probe sizes were selected as advised by device guidelines.

Machine learning algorithm

A variation of gradient boosting machine (GBM) was used to develop the models. GBM is an ensemble machine learning technique for regression and classification to produce prediction models of multiple base-learners (or decision trees). This algorithm involves three elements: optimization of a loss function, predictions made by a base-learner, and an additive model to add base-learners to minimize the loss function successively.

As GBM methods are known for overfitting, we applied stochastic GBM to reduce the correlation between trees in the sequence of GBM. Each iteration uses a sub-sample of the full training dataset, drawn at random, used in place of the full dataset to fit the base-learner and compute the model update for the current iteration (28). The randomized approach improves model robustness and reduces overfitting. Other GBM variations, such as XGBoost, were tested but were not an improvement from the stochastic method.

We explored alternative algorithms (logistic regression, k-nearest neighbors, support vector machine, decision tree algorithms). Only the results for GBM were further evaluated as it produced the best performing models in the preliminary analyses.

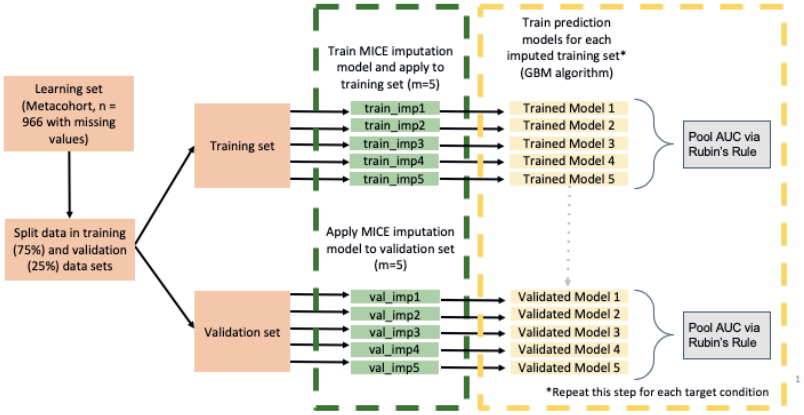

Learning set (Metacohort, n = 966 with missing values)
Split data in training (75%) and validation (25%) data sets
Training set
Validation set
Train MICE imputation model and apply to training set (m=5)
train_imp1
train_imp2
train_imp3
train_imp4
train_imp5
Apply MICE imputation model to validation set (m=5)
val_imp1
val_imp2
val_imp3
val_imp4
val_imp5
Train prediction models for each imputed training set* (GBM algorithm)
Trained Model 1
Trained Model 2
Trained Model 3
Trained Model 4
Trained Model 5
Validated Model 1
Validated Model 2
Validated Model 3
Validated Model 4
Validated Model 5
Pool AUC via Rubin's Rule
Pool AUC via Rubin's Rule
*Repeat this step for each target condition
1

Figure 1. Model development workflow

Dataset preprocessing

The original Metacohort dataset underwent a lengthy preprocessing phase to convert raw data to the optimal structure for training and testing GBM models. As the original dataset included over 200 clinical variables, we isolated those relevant for NAFLD based on clinical accessibility and established association guided by experienced hepatologists.

A pairwise Pearson correlation matrix was used to visualize the predictors' relationships and test for high intercorrelation. No variables were removed in this process, resulting in a final working dataset of 35 predictors.

Missing data were handled by first assessing the degree of missingness and if data were missing at random. Variables with missing values for more than 80% of participants were excluded entirely. For the remaining variables, missing data were replaced by multiple imputations (m=5) using the multivariate imputation by chain equations (MICE) approach (29). As data were split prior to this step, the training and validation data were imputed separately, resulting in 5 imputed training and validation sets. By purpose, outcome variables were excluded from the predictor matrix in the validation set as we aimed to mimic a scenario were the model is used in a ‘new’ patient, where outcome data is obviously not available. In support of our strategy, simulation studies have shown that including outcomes when imputing the validation set leads to over-optimistic predictions (30, 31). We further excluded variables with over 60% missing from the predictor matrix.

Continuous variables were centered and scaled to a mean of 0 and standard deviation of 1 to improve model stability and fit.

Model development

Figure 1 provides an overview of the model training and validation workflow. The learning data were randomly split into training (75%) and validation (25%) sets. The training set (n=742) was used to develop models using the GBM algorithm for each target condition. A grid search strategy was applied to tune hyperparameters (boosting iterations, max tree depth, shrinkage, minimum terminal node size) using 5 repeats of 10-fold cross-validation. Agreement between the model prediction and the observed outcome was inspected visually using calibration plots for each of the constructed models. Two sets of models were developed for each target condition, one using only routinely available clinical predictors (clinical model), and a second employing these same routinely available clinical predictors plus additional biomarkers (extended model).

.

As discussed above, NASH is established by the presence of three histological features (steatosis, lobular inflammation and ballooning). To address how these may best be combined, we developed two variants of the NASH models: one directly including these three histological features, the other by building a composite model that aggregated the calculated probabilities from models for steatosis, lobular inflammation and ballooning (Figure 2):

i. **The "direct" NASH model** was trained to a NAS≥4 with at least one point in each of the three components (S≥1 + B≥1 + LI≥1, with the sum being ≥4) (32, 33).

ii. **The "composite" NASH model** was similarly trained to a NAS ≥4 but with an additional, more stringent, liver inflammation threshold of 2 points (S≥1 + B≥1 + LI≥2, with the sum being ≥4). In the composite model, separate models were built for each histological feature, steatosis (0 vs. 1-3), lobular inflammation (0-1 vs. 2-3), and ballooning (0 vs 1-2), and the respective probabilities for each component were multiplied to yield a NASH prediction index.

In the same way, two different “at-risk” NASH model variants were developed: a “direct” at-risk NASH model and a “composite” at-risk NASH model, the later built by aggregating the calculated probabilities from the steatosis, lobular inflammation, ballooning and significant fibrosis models as discussed above.

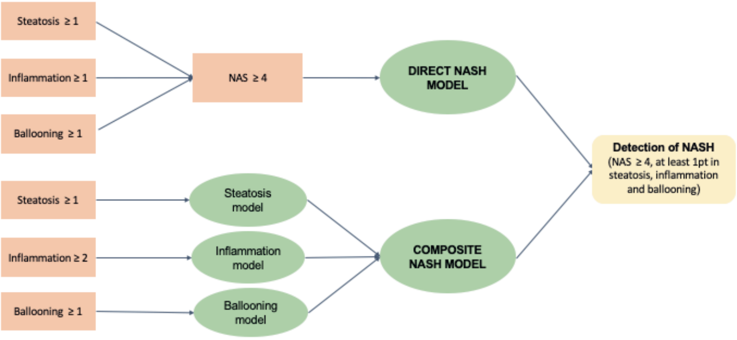


Figure 2. Construction of the direct and composite NASH models

Statistical analysis

The performance of each model was evaluated in the validation data (n=242), which were untouched and isolated from the model training process. The area under the receiver operating characteristic curve (AUC) in detecting the respective target conditions was calculated to express the accuracy of classifications against liver biopsy as the reference standard. As depicted in Figure 1, the model development and validation steps were repeated for each of the five imputed datasets, for each condition, and AUCs were pooled following Rubin’s Rule (34, 35).

Irrespective of how the model had been trained, to ensure that the full spectrum of NAS-defined steatohepatitis was captured, the target definition of NASH (or “at risk” NASH) used in the validation analyses were NAS of ≥4 with at least one point in each of the three components (S≥1 + B≥1 + LI≥1, in any permutation where the sum is ≥4).

The extended GBM models were compared to other tests: CK-18 and CAP by VCTE for NASH, FAST score (17) and ADAPT (36) for at-risk NASH, and PRO-C3, LSM by VCTE, the FIB-4 score (37) and the ELF test (38) for fibrosis, according to their original formula.

Table 1. Characteristics of the study group in the training and validation sets

	Overall Study group	Training set	Validation set
n	966	724	242
Age, years	51.19 (12.97)	51.80 (12.70)	49.37 (13.61)
Male, n (%)	563 (58.3)	416 (57.5)	147 (60.7)
BMI	34.08 (8.25)	34.10 (8.36)	34.03 (7.93)
ALT, U/L	62.67 (42.54)	62.25 (42.20)	63.92 (43.63)
AST, U/L	42.88 (26.01)	43.09 (26.80)	42.25 (23.55)
GGT, U/L	110.05 (160.19)	113.97 (172.46)	98.34 (115.50)
Albumin, g/L	4.39 (0.42)	4.38 (0.42)	4.41 (0.42)
Platelet, 10^9/L	238.96 (73.45)	237.36 (73.66)	243.76 (72.74)
Glucose, mmol/L	6.50 (2.57)	6.57 (2.65)	6.30 (2.31)
Triglyceride, mg/L	2.07 (1.21)	2.10 (1.26)	1.99 (1.05)
Diabetes, n (%)	406 (42.0)	318 (43.9)	88 (36.4)
FIB-4	1.38 (1.02)	1.41 (1.04)	1.29 (0.96)
VCTE-CAP	312.85 (73.19)	314.13 (71.71)	309.04 (77.46)
VCTE-LSM	11.47 (9.29)	11.15 (8.72)	12.44 (10.78)
Steatosis grade, % (0/1/2/3)	7/33/35/24/1	8/32/35/25/1	5/35/35/23/0
Steatosis, n (%)	898 (93.0)	670 (92.5)	228 (94.2)
Inflammation grade, % (0/1/2/3)	20/57/21/2	19/57/22/2	21/58/18/3
Inflammation, n (%)	223 (23.1)	172 (23.8)	51 (21.1)
Ballooning grade, % (0/1/2)	26/50/24	26/50/25	27/51/22
Ballooning, n (%)	715 (74.0)	539 (74.4)	176 (72.7)
NASH, n (%)	512 (53.0)	385 (53.2)	127 (52.5)
At-risk NASH, n (%)	335 (34.7)	260 (35.9)	75 (31.0)
Fibrosis stage, % (0/1/2/3/4)	32/19/21/20/9	32/18/22/20/8	34/24/15/17/10
Significant fibrosis, n (%)	471 (48.8)	368 (50.8)	103 (42.6)
Advanced fibrosis, n (%)	273 (28.3)	207 (28.6)	66 (27.3)

Continuous values are shown as mean (SD).

Steatosis is defined as 0 vs 1-3, inflammation as 0-1 vs 2-3, ballooning as 0 vs 1-2, NASH as NAS ≥4 (with at least one point in each component), significant fibrosis as F≥2, advanced fibrosis as F≥3, and at-risk NASH is the combination of NASH and significant fibrosis.

Variable importance scores were calculated for the GBM models to rank selected predictors based on their relative importance (scaled between 0 and 100) for making more accurate predictions. This was determined based on selection of variables in the tree building process and improvement for each boosting iteration (39).

All statistical analysis was performed using R software version 4.0.3. Multiple imputation was applied using the MICE package (29); GBM models were trained using the caret package (39).

Results

The study group had a mean age of 51 and mean body mass index of 34; 58% were men and 42% had diabetes. Based on liver biopsy, 53% had NASH, 35% had at-risk NASH, 49% had significant fibrosis, and 28% had advanced fibrosis (including 7% patients with histological cirrhosis). Details are summarized in Table 1. The flow of participants included in the LITMUS (Liver Investigation: Testing Marker Utility in Steatohepatitis) Metacohort can be seen in Supplementary Figure 1.

In comparison, CAP by VCTE and CK-18 M30 had AUCs of 0.64 (0.59, 0.70) and 0.61 (0.57, 0.65), respectively, for the detection of NASH (Figure 4). The composite NASH model was a significant improvement over CK-18.

At-risk NASH model

Two different at-risk NASH models were evaluated. The direct at-risk NASH model had an AUC of 0.79 (0.76, 0.82) using clinical variables, and 0.78 (0.75, 0.82) for the extended model (Table 2).

For the composite at-risk NASH model, the AUCs were 0.83 (0.80, 0.86) for both the clinical and extended versions. See Figure 3 for the predictors selected for each model and aggregated to calculate the composite models.

The composite GBM models performed well compared to other multi-marker scores
: FAST had an AUC of 0.77 (0.73, 0.81), and ADAPT had 0.77 (0.73, 0.80) (Figure 4).

Significant and advanced fibrosis model

The significant fibrosis models had AUCs of 0.76 (0.73, 0.80) and 0.78 (0.75, 0.82) for the clinical and extended version, respectively. Both fibrosis model probabilities were very consistent with the observed event rates (see calibration plots in Figure S3). Tuning parameters for each imputed dataset are shown in Supplementary Table 4.

In comparison, PRO-C3 had an AUC of 0.67 (0.63, 0.71), LSM by VCTE had 0.77 (0.73, 0.81), FIB-4 had 0.70 (0.66, 0.73), and ELF had 0.70 (0.66, 0.73) (Figure 4).

For advanced fibrosis, the AUC for the clinical model was 0.82 (0.79, 0.84). Adding biomarkers significantly improved the detection of advanced fibrosis, with an AUC of 0.86 (0.85, 0.87).

For advanced fibrosis, PRO-C3 had an AUC of 0.77 (0.73, 0.81), LSM by VCTE had 0.83 (0.80, 0.87), FIB-4 had 0.76 (0.73, 0.80), and ELF had 0.80 (0.76, 0.83) (Figure 4).

Table 2. Performance of the clinical and extended GBM models for detecting stages of NAFLD in the validation set

Outcome, model variant	Definition	Prevalence (%)	Clinical GBM model	Extended GBM model
Steatosis	0 vs. 1-3	93	0.94 (0.93, 0.96)	0.94 (0.92, 0.96)
Inflammation	0-1 vs. 2-3	23	0.79 (0.76, 0.81)	0.79 (0.76, 0.82)
Ballooning	0 vs 1-2	74	0.72 (0.69, 0.76)	0.74 (0.70, 0.77)
NASH, composite	S * I * B	53	0.71 (0.67, 0.74)	0.71 (0.68, 0.77)
NASH, direct	NAS≥4		0.61 (0.57, 0.66)	0.65 (0.60, 0.69)
At-risk NASH, composite	S * I * B * F	35	0.83 (0.80, 0.86)	0.83 (0.80, 0.86)
At-risk NASH, direct	NAS≥4 and F≥2		0.79 (0.76, 0.82)	0.78 (0.75, 0.82)
Significant fibrosis	F≥2	47	0.76 (0.73, 0.80)	0.78 (0.75, 0.82)
Advanced fibrosis	F≥3	28	0.82 (0.79, 0.84)	0.86 (0.85, 0.87)

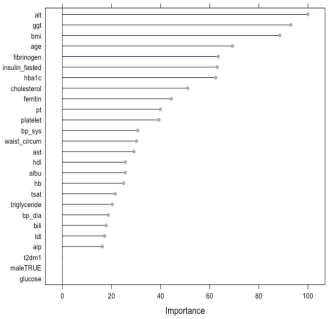
Clinical GBM models include only clinical predictors, extended GBM models include clinical predictors and biomarkers.

Composite NASH model was constructed by aggregating the three NASH components: steatosis, lobular inflammation and ballooning, which were dichotomized according to the definition as described. At-risk NASH was constructed similarly, including significant fibrosis (F≥2).

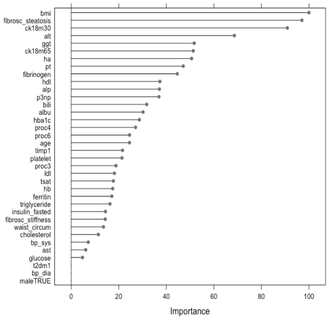
Direct NASH model was constructed using the standard dichotomization of NAS score (≥4), with at least one point in each component of steatosis, lobular inflammation and ballooning.

Direct at-risk NASH is the combination of NAS score (≥4) and significant fibrosis (F≥2).

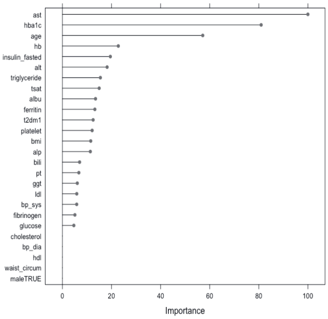
Clinical variable importance for steatosis



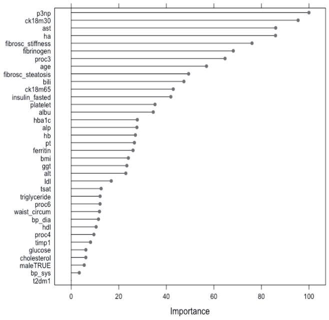
Variables of importance for steatosis



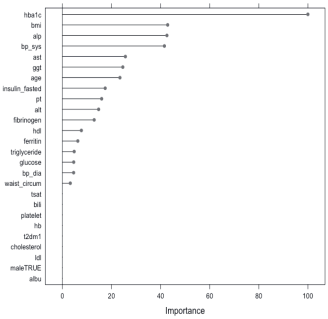
Clinical variable importance for inflammation



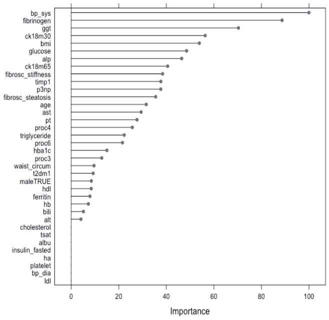
Variables of importance for inflammation



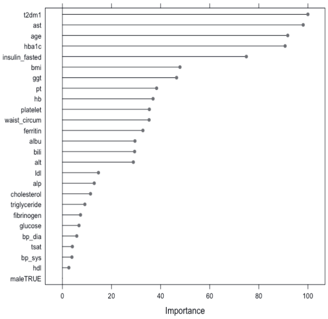
Clinical variable importance with for ballooning



Variables of importance for ballooning



Clinical variables of importance for significant fibrosis



Variables of importance for significant fibrosis

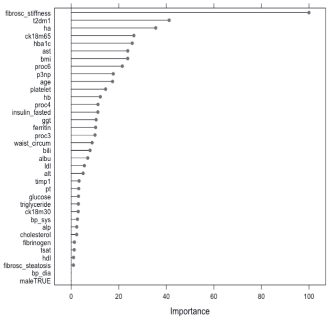


Figure 3. **Variables of importance for steatosis, inflammation, ballooning and significant fibrosis for the clinical and extended GBM models. systolic blood pressure (bp_sys), diastolic blood pressure (bp_dia), type 2 diabetes (t2dm), high density lipoprotein (hdl), low density lipoprotein (ldl), alanine aminotransferase (alt), aspartate aminotransferase (ast), gamma-glutamyl transferase (ggt), alkaline phosphatase (alp), hemoglobin (hb), transferrin saturation (tsat), albumin (albu), clotting (pt), bilirubin (bili), glycosylated hemoglobin A1c(hba1c), cytokeratin 18 (ck18, m30 and m65 antigens), plasma propeptides of procollagen type III (proc3, proc4, proc6), tissue inhibitor of metalloproteinases 1 (timp1), amino-terminal propeptide of type III procollagen (p3np) and hyaluronic acid (ha).**

Discussion

Several diagnostic scores have been studied to identify patients with advanced stages of fibrosis. However, those for detecting active NASH (with or without fibrosis) have been less successful. The present study utilized a large histologically-characterized NAFLD cohort in Europe with a rich selection of novel biomarkers to develop diagnostic models using the GBM algorithm. Two sets of models were developed for each condition, one using only clinical features, and a second by adding biomarkers, such as CK-18, PRO-C3/4/6, and LSM and CAP by VCTE.

We explored the added value of fitting the GBM algorithm for steatosis, inflammation, and ballooning separately and creating an aggregate model combining the three components. The purpose was to enhance classifications, as NASH models are generally developed solely based on the NAS score and, so far, none are suggested for use by clinical guidelines (10). Our results showed that aggregating the probabilities for each component to arrive at the composite NASH score significantly improved the accuracy for detecting NASH. The same strategy also improved detection of at-risk NASH.

The performance of the models for NASH and at-risk NASH were comparable between the clinical and extended models. The fibrosis models benefited the most from the additional biomarkers, with significant improvement in detecting advanced fibrosis.

The study presents some limitations, mostly related to the retrospective nature of the LITMUS Metacohort. Liver biopsy serves as the reference standard in our analysis as it remains the recommended technique for evaluating NASH, despite caveats such inter and intra- reader variability (8, 40). We further note that our definition of NASH

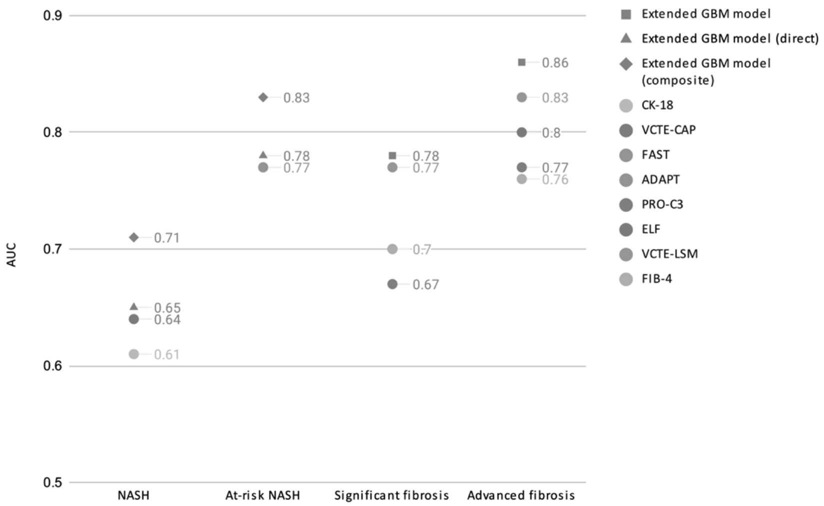


Figure 4. AUC of extended GBM models in the validation set compared to existing non-invasive scores for detecting NASH, at-risk NASH, significant and advanced fibrosis.

corresponds to the efficacy endpoint defined by health authorities and clinical trials for NAFLD drug development, which may differ from clinical diagnosis of NASH that considers other inputs in addition to histopathologic diagnosis (33). We relied on locally read biopsies and local lab results for standard markers, which may introduce differences across study sites. While blood-based biomarkers were centrally analyzed, they were measured in retrospectively collected samples and reserved in a biobank. They were also analyzed in batches.

Due to limited sample volume, not all biomarkers were measured in all patients. We avoided complete case analysis by imputing missing values by multiple imputation. Five imputed datasets were produced and analyzed as simulation studies have shown a required number of repeated imputations to be as low as three for datasets with 20% missingness (41). The retrospective nature of this study also meant that some samples were older than others. As the stability of these samples is largely unknown, we excluded a handful of samples collected before 2010 from the analysis.

Other machine learning models have been developed using only clinical predictors to detect NASH. Using a variant of GBM, one study found out-of-sample AUCs of 0.82 and 0.76, using data from the National Institute of Diabetes, Digestive and Kidney Diseases and Optum Analytics (21). Another study applied machine learning algorithms to predict NASH using data from Optum Analytics and found the highest AUC (0.88) using XGBoost (42). This study, however, included healthy participants without any liver-related diseases in the model development phase. Both studies relied on data from Optum, which, in the absence of histological diagnosis of NASH, relied on several different ICD codes for NASH or NAFLD.

Other scores have been developed using regression-based methods, such as NIS4 and MACK-3. NIS4, which includes four components (miR-34a-5p, alpha-2 macroglobulin, YKL-40, and glycated hemoglobin), had an AUC of 0.80 from three validation cohorts for detecting at-risk NASH (18). An external validation study found that MACK-3 (fasting glucose and insulin, AST and CK-18) also had an AUC of 0.80 for detecting at-risk NASH. More recently, the SomaSignal test developed based on elastic net produced an AUC of 0.76 (43). All of these multi-marker scores include more novel biomarkers, which come with the cost of additional testing. Our at-risk NASH model performed well, relying only on clinical data, highlighting a potential advantage of utilizing machine learning. This warrants further evaluation in an external cohort.

Constructing separate models for each component of NASH further allowed us to observe that different predictors were selected as most informative for each component. The most influential predictors had strong biological plausibility or an established position in the disease pathway (44) (45) (12, 46). However, the ranking of markers was variable across imputed datasets and should be interpreted with caution.

In the future, we plan to finalize the models using complete data from the on-going prospective LITMUS Study Cohort, focusing on the aggregated approach for constructing the models for NASH and at-risk NASH. A single model for each outcome will be converted to a user-friendly interface, in the form of a Shiny-app. Such tools would allow clinicians, including those in a primary care setting, to enter values of clinical parameters to detect NASH or at-risk NASH with greater ease.

Machine learning approaches are sometimes perceived as too complicated compared to classic regression-based tools. Some studies have demonstrated superior performance of machine learning algorithms over logistic regression, such as Feng et al., who found machine learning models outperformed regression-based models for detecting significant fibrosis across different subgroups (47). However, a large meta-analysis found no benefit of machine learning over logistic regression (48). Given the vast selection of available algorithms, heterogeneous study designs, sparse reporting, and conflicting conclusions in the literature, more work is needed to understand which tools and study design elements are optimal for developing diagnostic models for NAFLD. This should be paired with a clear emphasis on the tools desired clinical context of use, whether to triage patients in clinical practice or select participants most likely to benefit from therapeutic interventions in clinical trials.

Our study found promising results to explore machine learning algorithms further to improve the diagnosis of NASH and at-risk NASH, using readily available clinical data. The inherent ability to adapt to new data positions machine learning as a valuable tool for rapidly evolving healthcare settings and conditions with a complex etiology such as NAFLD. While the move towards machine learning to detect NAFLD is still in its infancy, concerted effort to robust methodology, biomarker discovery, and quality data can improve the clinical management of NAFLD. Importantly, this is most needed outside expert centers, where the vast majority of patients do not have access to specialists focusing on liver disease.

References

1. Younossi Z, Henry L. The Burden of NAFLD Worldwide. Non-Alcoholic Fatty Liver Disease: Springer; 2020. p. 15-24.

2. Noureddin M, Truong E, Gornbein JA, Saouaf R, Guindi M, Todo T, et al. MRI-based (MAST) score accurately identifies patients with NASH and significant fibrosis. Journal of Hepatology. 2022;76(4):781-7.

3. Satapathy SK, Sanyal AJ, editors. Epidemiology and natural history of nonalcoholic fatty liver disease. Seminars in liver disease; 2015: Thieme Medical Publishers.

4. Younossi Z, Anstee QM, Marietti M, Hardy T, Henry L, Eslam M, et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nature reviews Gastroenterology & hepatology. 2018;15(1):11-20.

5. Younossi Z, Stepanova M, Ong JP, Jacobson IM, Bugianesi E, Duseja A, et al. Nonalcoholic steatohepatitis is the fastest growing cause of hepatocellular carcinoma in liver transplant candidates. Clinical Gastroenterology and Hepatology. 2019;17(4):748-55. e3.

6. Ascha MS, Hanouneh IA, Lopez R, Tamimi TAR, Feldstein AF, Zein NN. The incidence and risk factors of hepatocellular carcinoma in patients with nonalcoholic steatohepatitis. Hepatology. 2010;51(6):1972-8.

7. Sanyal AJ, Van Natta ML, Clark J, Neuschwander-Tetri BA, Diehl A, Dasarathy S, et al. Prospective study of outcomes in adults with nonalcoholic fatty liver disease. New England Journal of Medicine. 2021;385(17):1559-69.

8. 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.

9. Brunt EM, Clouston AD, Goodman Z, Guy C, Kleiner DE, Lackner C, et al. Complexity of ballooned hepatocyte feature recognition: Defining a training atlas for artificial intelligence-based imaging in NAFLD. Journal of Hepatology. 2022;76(5):1030-41.

10. Berzigotti A, Tsochatzis E, Boursier J, Castera L, Cazzagon N, Friedrich-Rust M, et al. EASL Clinical Practice Guidelines on non-invasive tests for evaluation of liver disease severity and prognosis–2021 update. Journal of Hepatology. 2021;75(3):659-89.

11. Mózes FE, Lee JA, Selvaraj EA, Jayaswal ANA, Trauner M, Boursier J, et al. Diagnostic accuracy of non-invasive tests for advanced fibrosis in patients with NAFLD: an individual patient data meta-analysis. Gut. 2022;71(5):1006-19.

12. Lee J, Vali Y, Boursier J, Duffin K, Verheij J, Brosnan MJ, et al. Accuracy of cytokeratin 18 (M30 and M65) in detecting non-alcoholic steatohepatitis and fibrosis: A systematic review and meta-analysis. Plos one. 2020;15(9):e0238717.

13. Tada T, Kumada T, Toyoda H, Saibara T, Ono M, Kage M. New scoring system combining the FIB-4 index and cytokeratin-18 fragments for predicting steatohepatitis and liver fibrosis in patients with nonalcoholic fatty liver disease. Biomarkers. 2018;23(4):328-34.

14. Younossi ZM, Page S, Rafiq N, Birerdinc A, Stepanova M, Hossain N, et al. A biomarker panel for non-alcoholic steatohepatitis (NASH) and NASH-related fibrosis. Obes Surg. 2011;21(4):431-9.

15. Anty R, Iannelli A, Patouraux S, Bonnafous S, Lavallard V, Senni-Buratti M, et al. A new composite model including metabolic syndrome, alanine aminotransferase and cytokeratin-18 for the

diagnosis of non-alcoholic steatohepatitis in morbidly obese patients. Alimentary pharmacology & therapeutics. 2010;32(11-12):1315-22.

16. Boursier J, Anty R, Vonghia L, Moal V, Vanwolleghem T, Canivet C, et al. Screening for therapeutic trials and treatment indication in clinical practice: MACK-3, a new blood test for the diagnosis of fibrotic NASH. Alimentary pharmacology & therapeutics. 2018;47(10):1387-96.

17. Newsome PN, Sasso M, Deeks JJ, Paredes A, Boursier J, Chan W-K, et al. FibroScan-AST (FAST) score for the non-invasive identification of patients with non-alcoholic steatohepatitis with significant activity and fibrosis: a prospective derivation and global validation study. The Lancet Gastroenterology & Hepatology. 2020;5(4):362-73.

18. Harrison SA, Ratziu V, Boursier J, Francque S, Bedossa P, Majd Z, et al. A blood-based biomarker panel (NIS4) for non-invasive diagnosis of non-alcoholic steatohepatitis and liver fibrosis: a prospective derivation and global validation study. Lancet Gastroenterol Hepatol. 2020;5(11):970-85.

19. Adamichou C, Genitsaridi I, Nikolopoulos D, Nikoloudaki M, Repa A, Bortoluzzi A, et al. Lupus or not? SLE Risk Probability Index (SLERPI): a simple, clinician-friendly machine learning-based model to assist the diagnosis of systemic lupus erythematosus. Annals of the rheumatic diseases. 2021;80(6):758-66.

20. Karaglani M, Gourlia K, Tsamardinos I, Chatzaki E. Accurate blood-based diagnostic biosignatures for Alzheimer's disease via automated machine learning. Journal of clinical medicine. 2020;9(9):3016.

21. Docherty M, Regnier SA, Capkun G, Balp M-M, Ye Q, Janssens N, et al. Development of a novel machine learning model to predict presence of nonalcoholic steatohepatitis. Journal of the American Medical Informatics Association. 2021;28(6):1235-41.

22. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of internal medicine. 2015;162(1):W1-W73.

23. Hardy T, Wonders K, Younes R, Aithal GP, Aller R, Allison M, et al. The European NAFLD Registry: A real-world longitudinal cohort study of nonalcoholic fatty liver disease. Contemp Clin Trials. 2020;98:106175.

24. 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.

25. Bedossa P. Diagnosis of non-alcoholic fatty liver disease/non-alcoholic steatohepatitis: Why liver biopsy is essential. Liver International. 2018;38(S1):64-6.

26. Anania FA, Dimick-Santos L, Mehta R, Toerner J, Beitz J. Nonalcoholic Steatohepatitis: Current Thinking From the Division of Hepatology and Nutrition at the Food and Drug Administration. Hepatology. 2021;73(5):2023-7.

27. Ratziu V, Harrison SA, Francque S, Bedossa P, Lehert P, Serfaty L, et al. Elafibranor, an Agonist of the Peroxisome Proliferator−Activated Receptor−α and −δ, Induces Resolution of Nonalcoholic Steatohepatitis Without Fibrosis Worsening. Gastroenterology. 2016;150(5):1147-59.e5.

28. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis. 2002;38(4):367-78.

29. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2011;45(1):1-67.

30. Hoogland J, van Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGW, et al. Handling missing predictor values when validating and applying a prediction model to new patients. Statistics in Medicine. 2020;39(25):3591-607.

31. Musoro JZ, Zwinderman AH, Puhan MA, ter Riet G, Geskus RB. Validation of prediction models based on lasso regression with multiply imputed data. BMC Med Res Methodol. 2014;14(1):1-13.

32. Sanyal AJ, Brunt EM, Kleiner DE, Kowdley KV, Chalasani N, Lavine JE, et al. Endpoints and clinical trial design for nonalcoholic steatohepatitis. Wiley Online Library; 2011.

33. Noncirrhotic nonalcoholic steatohepatitis with liver fibrosis: developing drugs for treatment. In: Services DoHaH, (CDER) CfDEaR, editors.

34. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009;9:57-.

35. Heymans M, Eekhout I. Applied missing data analysis with SPSS and (R) Studio. Heymans and Eekhout: Amsterdam, The Netherlands: 20Available online: https://bookdown org/mwheymans/bookmi/[accessed 23 May 2020]. 2019.

36. Daniels SJ, Leeming DJ, Eslam M, Hashem AM, Nielsen MJ, Krag A, et al. ADAPT: An Algorithm Incorporating PRO-C3 Accurately Identifies Patients With NAFLD and Advanced Fibrosis. Hepatology. 2019;69(3):1075-86.

37. Vallet-Pichard A, Mallet V, Nalpas B, Verkarre V, Nalpas A, Dhalluin-Venier V, et al. FIB-4: an inexpensive and accurate marker of fibrosis in HCV infection. comparison with liver biopsy and fibrotest. Hepatology. 2007;46(1):32-6.

38. Day JW, Rosenberg WM. The enhanced liver fibrosis (ELF) test in diagnosis and management of liver fibrosis. Br J Hosp Med (Lond). 2018;79(12):694-9.

39. Kuhn M. Building predictive models in R using the caret package. Journal of statistical software. 2008;28(1):1-26.

40. Davison BA, Harrison SA, Cotter G, Alkhouri N, Sanyal A, Edwards C, et al. Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. Journal of Hepatology. 2020;73(6):1322-32.

41. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in medicine. 1999;18(6):681-94.

42. Fialoke S, Malarstig A, Miller MR, Dumitriu A. Application of Machine Learning Methods to Predict Non-Alcoholic Steatohepatitis (NASH) in Non-Alcoholic Fatty Liver (NAFL) Patients. AMIA Annu Symp Proc. 2018;2018:430-9.

43. Vali Y, Lee J, Schattenberg J, Gomez MR, Tiniakos D, Bedossa P, et al. Comparative diagnostic accuracy of blood-based biomarkers for diagnosing NASH: phase 1 results of the LITMUS project. International Liver Congress2021.

44. Tanwar S, Trembling PM, Guha IN, Parkes J, Kaye P, Burt AD, et al. Validation of terminal peptide of procollagen III for the detection and assessment of nonalcoholic steatohepatitis in patients with nonalcoholic fatty liver disease. Hepatology. 2013;57(1):103-11.

45. Darweesh SK, AbdElAziz RA, Abd-ElFatah DS, AbdElazim NA, Fathi SA, Attia D, et al. Serum cytokeratin-18 and its relation to liver fibrosis and steatosis diagnosed by FibroScan and controlled attenuation parameter in nonalcoholic fatty liver disease and hepatitis C virus patients. European Journal of Gastroenterology & Hepatology. 2019;31(5):633-41.

46. Feldstein AE, Alkhouri N, De Vito R, Alisi A, Lopez R, Nobili V. Serum cytokeratin-18 fragment levels are useful biomarkers for nonalcoholic steatohepatitis in children. Am J Gastroenterol. 2013;108(9):1526-31.

47. Feng G, Zheng KI, Li Y-Y, Rios RS, Zhu P-W, Pan X-Y, et al. Machine learning algorithm outperforms fibrosis markers in predicting significant fibrosis in biopsy-confirmed NAFLD. Journal of Hepato-Biliary-Pancreatic Sciences. 2021;28(7):593-603.

48. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 2019;110:12-22.

Supplementary Material

https://journals.lww.com/hep/Fulltext/2023/07000/Machine_learning_algorithm_improves_the_detection.21.aspx

08

Covariate-specific ROC curve analysis can accommodate differences between covariate subgroups in the evaluation of diagnostic accuracy

Jenny Lee
Nick van Es
Toshihiko Takada
Frederikus A. Klok
Geert-Jan Geersing
Jeffrey Blume
Patrick M. Bossuyt

Journal of Clinical Epidemiology, 2023.

Abstract

Objective: We present an illustrative application of methods that account for covariates in receiver operating characteristic (ROC) curve analysis, using individual patient data on D-dimer testing for excluding pulmonary embolism.

Study design and setting: Bayesian nonparametric covariate-specific ROC curves were constructed to examine the performance/positivity thresholds in covariate subgroups. Standard ROC curves were constructed. Three scenarios were outlined based on comparison between subgroups and standard ROC curve conclusion: (1) identical distribution/identical performance, (2) different distribution/identical performance, and (3) different distribution/different performance. Scenarios were illustrated using clinical covariates. Covariate-adjusted ROC curves were also constructed.

Results: Age groups had prominent differences in D-dimer concentration, paired with differences in performance (Scenario 3). Different positivity thresholds were required to achieve the same level of sensitivity. D-dimer had identical performance, but different distributions for YEARS algorithm items (Scenario 2), and similar distributions for sex (Scenario 1). For the later covariates, comparable positivity thresholds achieved the same sensitivity. All covariate-adjusted models had AUCs comparable to the standard approach.

Conclusion: Subgroup differences in performance and distribution of results can indicate that the conventional ROC curve is not a fair representation of test performance. Estimating conditional ROC curves can improve the ability to select thresholds with greater applicability.

Introduction

Biomarkers are regularly investigated for their ability to classify subjects as diseased or non-diseased. Receiver operating characteristic (ROC) curves are, unarguably, the most widely used tool for evaluating the discriminatory capacity, initially popular with the evaluation of imaging modalities. Their use has now spread to all tests that deliver results on an ordinal, interval or ratio scale (1). The overall diagnostic accuracy of a medical test is then expressed as the corresponding area under the ROC curve (AUC). The shape of a ROC curve illustrates the trade-off between the sensitivity and specificity of a test at various positivity thresholds, which converts a continuous classifier into a dichotomous one. Oftentimes, a desired level of classification is specified, to maximize the true positive or true negative results, and identify the corresponding threshold (2).

The result of a test can be associated with other factors than the presence or absence of the target condition. For instance, older patients tend to have higher D-dimer values than younger ones, while males have higher hemoglobin levels than females. In the presence of such associations, there may also be covariate-specific (such as age or sex) differences in test performance. Moreover, selecting thresholds from a standard ROC curve can be misleading, as compared to subgroup specific ROC curves, when strong associations between the marker and covariate are present, and result in differences in sensitivity and specificity between covariate subgroups. Thus, when covariate information is available, it should be considered, as neglecting such information may inflate our estimates of the relative proportion of false negative or false positive test results for certain subgroups (3).

In light of this, several methods that account for covariates in ROC curve analysis have been proposed (4, 5). They allow assessment of covariate-specific and covariate-adjusted ROC curves; the former models ROC curves for each stratum of a given covariate (e.g. men and women), while the other models a single ROC curve that can be interpreted as the weighted average of covariate-specific curves (6). Despite the widespread consideration of covariate effects in randomized trials of interventions, it is not yet standard practice in diagnostic accuracy studies (7, 8).

Associations between covariates and the positivity threshold are even less considered, when, in fact, it has direct implications for how the test will be implemented for practical use. Understanding the magnitude of potential covariate effects and applying appropriate techniques are therefore fundamental to produce robust and reliable results that can be translated into clinical practice.

We here present an illustrative application of the use of covariate-specific and covariate-adjusted ROC analyses, to encourage a more widespread application of such methods in evaluations of diagnostic accuracy. The following sections are structured as follows: the motivating example, an outline of conventional ROC curve analysis and possible scenarios when considering covariates, the application, and concluding remarks.

Motivating example

Pulmonary embolism (PE) is a common venous thromboembolic disease that can cause significant morbidity and mortality (9, 10). Patients with suspected venous thromboembolism (VTE), comprised of PE and deep vein thrombosis, usually undergo imaging testing, such as compression ultrasonography or computed tomography pulmonary angiography (CTPA), for a confirmation or exclusion of diagnosis. However, signs and symptoms indicating PE are non-specific, and therefore PE is not confirmed in many patients with the suspected disease. Considering the additional risks and costs of performing CTPA, scoring systems and tests have been proposed to indicate those at greater risk.

Diagnostic clinical scores comprised of clinical characteristics, such as the Wells score, have been developed to classify patients with suspected PE into pre-test probability, and ultimately minimize the number of patients subjected to CTPA testing (11, 12). More recently, the YEARS algorithm was proposed, consisting of only three components, offering a more simplified decision rule (11).

D-dimer is a sensitive plasma marker of endogenous fibrinolysis that appears following blood clot degradation (13). Measuring levels of this degradation product is commonly used as a diagnostic test in patients with signs and symptoms suggestive of venous thromboembolism. A threshold of 500 ng/mL was initially proposed for D-dimer to rule out VTE in patients with non-high pre-test probability. A more recent study factored the patient's pretest probability and proposed an additional upper threshold of 1000 ng/mL for those without any YEARS items to increase the proportion of patients in whom imaging can be withheld (14). Yet the optimal approach for adjusting d-dimer thresholds has still to be determined (15).

Other factors have shown to influence D-dimer concentration. Age, for example, is associated with D-dimer positivity (16). The D-dimer concentration naturally increases with age, leading to many older patients without PE presenting with D-dimer levels above the conventional threshold of 500 ng/mL (17). When D-dimer testing is performed among elderly, the proportion of false-positive results is higher leading to unnecessary imaging (18, 19). Age-dependent threshold values for D-dimer were proposed and its diagnostic performance has been compared to the conventional threshold (20, 21). The age-adjusted D-dimer threshold was defined as age(years)*10 ng/mL for patients aged over 50 years, based on evaluating optimal values for 10-year interval age groups. Studies have also shown the influence of other factors, such as setting (inpatient/outpatient) and cancer status (22, 23).

Individual patient data cohort

We consider data from a large individual patient data (IPD) meta-analysis of studies assessing the accuracy of clinical decision rules and D-dimer testing for detection of VTE among patients with suspected PE (24). In the IPD cohort, data from 21,621 patients, from 16 studies recruited between 1990 and 2020, were included in the analysis. In this cohort, 15% was diagnosed with PE. PE diagnosis was objectively confirmed with either CTPA or clinical follow-up of at least one month in those without initial anticoagulation treatment upon initial testing. The characteristics of the IPD cohort are described in Supplementary Table 2.

Receiver operating characteristic (ROC) curve analysis

Conventional ROC curve analysis

In a diagnostic accuracy study, ROC curves can be constructed where the results of one or more index tests are compared against the results of the clinical reference standard, the best available test to evaluate the presence or absence of the target condition (25).

Positivity threshold

If a positivity threshold is defined, the diagnostic accuracy of an index test can be expressed by estimates of its sensitivity and specificity. If higher index test results make the target condition more likely, sensitivity corresponds to the proportion of those with a target condition whose test result exceeds the positivity threshold. Analogously, the specificity refers to the proportion of those without the target condition whose test result does not exceed the positivity threshold.

If no positivity threshold can be defined, or none was defined a priori, one can consider the full ROC curve. The y-axis of the ROC curve displays all possible values of the sensitivity (or true positive fraction, TPF). The x-axis displays all possible values of the specificity, from right to left, or of the false positive fraction (FPF, one minus specificity), from left to right.

The ROC curve links the TPF and FPF; it is based on the survival function (one minus the cumulative distribution function) of the test results in the subgroup with the target condition, as indicated by the reference standard, and links this to the survival function of the test results in the subgroup without the target condition. The area under the ROC curve (AUC, also known as AUROC) takes values between zero and one, where one indicates perfect performance and 0.5 refers to performance no better than flipping a coin.

ROC curve analysis incorporating covariates

In most diagnostic accuracy studies, all available index test and reference standard results are used to construct ROC curves. No other patient or study characteristics are considered as covariates. By now it is well known that diagnostic accuracy is not a fixed property of a test and that it can vary between population subgroups, test types, settings, and depending on the position of the test in the clinical pathway (26, 27).

Covariate-specific ROC curve scenarios and implications

If the covariate can be indicated by one, dichotomous variable, the investigators can create two subgroups and correspondingly create two different ROC curves. In line with terminology in the statistical literature, we will refer to these as covariate-specific ROC curves.

Three scenarios can be drawn based on a comparison of these two ROC curves, as well as conclusions regarding the standard ROC curve. We illustrate the scenarios using sex as the covariate of interest.

Scenario 1. Identical distribution, identical performance

It is possible that the covariate-specific ROC curves are completely identical. That would be the case, for example, if the underlying distributions of test results in those with and those without the target condition are identical in men and women. Each positivity threshold would then yield the same sensitivity and specificity in women as in men. The AUC would be the same in men and in women.

Implication

If no difference exists and the covariate-specific ROC curves are identical, the standard AUC expresses performance well, since the covariate-specific AUC are one and the same.

Scenario 2. Different distribution, identical performance

In a different scenario, the distribution of test results differs between men and women. Again, as an example, men may have higher values, on average, than women, both in those with and in those without the target condition. In that case, a single positivity threshold would yield a different sensitivity in men compared to women, and a different specificity. Sensitivity will be higher in men but specificity lower.

It is still possible that the two covariate-specific ROC curves are identical. For example, if the distributions of those with and without the target condition have the same difference in means between men and women, without any differences in variance, then the two covariate-specific curves will be the same, as well as the AUC. Overall performance, as expressed by the AUC, will be the same in men and women, but different positivity thresholds must be selected to yield the same sensitivity and specificity.

Implication

As demonstrated by Janes and Pepe (2008), if a difference in distributions exists but the covariate-specific ROC curves are identical, the standard AUC can present a biased upward estimate of test performance (3). This will be the case if one subgroup, say men, is more likely to have the target condition. The standard ROC curve will also capture that additional difference between men and women and will lie above the covariate-specific ROC curve. The standard AUC, though correctly estimated, will then also show upward bias, since it does not only express the performance of the test but is also based on the pre-existing difference in prevalence between men and women.

If, in an alternative scenario, the prevalence between men and women is the same, the standard ROC curve will be attenuated: it will lie below the covariate-specific ROC curves. The standard AUC will not express performance well. The identical covariate-specific AUC, based on thresholds that differ between men and women, will be higher: it reflects the gain in performance that is possible from using such stratified positivity thresholds.

Scenario 3. Different distribution, different performance

In a third scenario, the distributions of the test results in those with and without the target condition differ in such a way that the covariate-specific ROC curves are no longer identical. That can happen in diverse ways. It is possible that men without the target condition have the same distribution as women without the target condition but, with the target condition, men have much higher values than women. If so, overall performance will also be different. Depending on the distributions, a single positivity threshold may also lead to different values for sensitivity and specificity in the subgroups.

Implication

If the covariate-specific ROC curves and AUC differ, the standard AUC is not a fair representation of performance, as it ignores the potentially meaningful differences between the subgroups. Presenting the significantly different covariate-specific ROC curves and the corresponding AUC may be more informative for clinical decision-making.

Bayesian non-parametric model

All of the performed analyses were based on the Bayesian nonparametric approach proposed by In´acio de Carvalho et al (28). This approach incorporates covariate information by using a single-weights dependent Dirichlet process mixture of normal distributions. Specifically, the model includes a mixture of normal distributions with means that follow a regression model, which may be linear or nonlinear, dependent on the covariate(s) (29). This allows for the construction of covariate-specific ROC curves, specified for the conditional CDF which changes as a function of the covariate, as opposed to just considering the mean or variance of the distribution, as in other semiparametric approaches (30).

Statistical analysis

Cumulative distribution function (CDF) plots and histograms were created for covariate subgroups to explore the distribution and density of index test results among the diseased and non-diseased.

The standard empirical ROC curve was constructed without incorporating covariate information (7). We utilized the Bayesian nonparametric approach to construct covariate-specific ROC curves (28), including ordinal and continuous covariates which were dichotomized, where necessary, into clinically relevant categories. For each Bayesian nonparametric model, we estimated the densities and distribution by disease status. In addition, we also constructed covariate-adjusted ROC curves, initially developed by Janes and Pepe (31), but adapted to the Bayesian non-parametric approach. Here we included covariates without categorization. Diagnostic accuracy was expressed as the AUC with its 95% confidence interval (95% CI). For each individual ROC curve, positivity thresholds corresponding to a sensitivity of 0.98, 0.95, 0.90 were identified.

All statistical analyses were performed using R software version 4.0.3, using the ROCnReg package (32). For detailed introduction and illustration of various frameworks for covariate consideration in ROC curve analysis, we refer to the ROCnReg guidance document (32).

Application

Subgroup differences

We conducted a series of exploratory analyses to landscape the distribution of index test values across covariate subgroups and in the diseased and non-diseased subgroups. There were differences in D-dimer concentration between age groups, more prominent in the non-diseased group, with much wider dispersion among the diseased. Differences were less pronounced for other covariates (Supplementary Figure 1).

Overall, there was unanimous right-skewed distribution of test results. We further visually confirmed largely overlapping distributions of test results between sex subgroups, with differences in frequency of lower test results among the non-diseased for some covariates (Supplementary Figure 1). The PE prevalence varied between some of the covariate subgroups, for example those based on age and on the presence of YEARS items (Table 1).

Performance estimates with conventional ROC approach

We first constructed the standard ROC curve to evaluate performance without incorporating any covariate information. Using an empirical estimator, we found D-dimer had an AUC of 0.87 (95% CI: 0.86, 0.88) in detecting VTE. This was considered as the benchmark performance indicator.

Performance estimates with covariate-specific ROC curve analysis

When constructing covariate-specific ROC curves, our interest was in evaluating whether the discriminatory capacity of the index test varies between covariate subgroups. In our case, we were interested in the possible effect of age, sex, and pretest probability by use of the YEARS algorithm on the performance of the index test, D-dimer.

Performance of the index test varied significantly between age groups (Table 1). In older patients the AUC was lower (0.84 [0.84, 0.85]) than in the younger group (AUC of 0.88 [0.87,0.89]). We saw a noticeable gap in the CDF between the age groups (Figure 1).

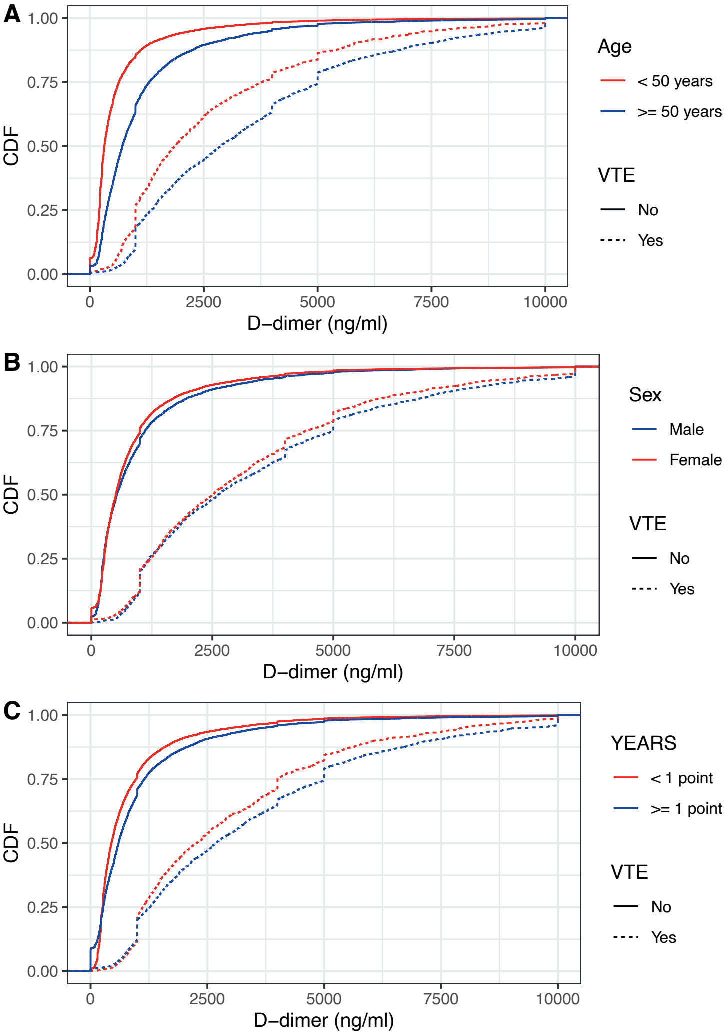


Figure 1. Cumulative distribution function (CDF) plots by covariate and venous thromboembolism (VTE) status

Table 1. Covariate-specific performance of D-dimer and prevalence in corresponding subgroups

		Proportion with VTE	AUC (95% CI)	Sensitivity = 0.98		Sensitivity = 0.95		Sensitivity = 0.90	
				Threshold* (95% CI)	Specificity (95% CI)	Threshold* (95% CI)	Specificity (95% CI)	Threshold (95% CI)	Specificity (95% CI)
Standard		15%	0.87 (0.86, 0.88)	470	0.47	686	0.61	910	0.70
Age	< 50 years	9%	0.88 (0.87, 0.89)	114 (0, 205)	0.13 (0.03, 0.26)	402 (325, 473)	0.57 (0.47, 0.65)	662 (596, 725)	0.75 (0.72, 0.78)
	≥ 50 years	18%	0.84 (0.84, 0.85)	401 (346, 452)	0.28 (0.23, 0.33)	655 (610, 698)	0.47 (0.44, 0.50)	893 (852, 934)	0.60 (0.58, 0.62)
Sex	Male	18%	0.86 (0.85, 0.87)	351 (286, 413)	0.35 (0.27, 0.41)	609 (555, 661)	0.54 (0.51, 0.56)	850 (799, 901)	0.65 (0.62, 0.67)
	Female	13%	0.87 (0.86, 0.87)	297 (235, 357)	0.29 (0.21, 0.37)	560 (507, 610)	0.55 (0.51, 0.58)	803 (755, 850)	0.67 (0.65, 0.69)
YEARS	YEARS = 0	8%	0.87 (0.86, 0.88)	311 (228, 392)	0.31 (0.18, 0.43)	575 (510, 638)	0.58 (0.54, 0.62)	818 (757, 877)	0.69 (0.66, 0.72)
	YEARS ≥ 1	21%	0.85 (0.85, 0.86)	327 (269, 385)	0.32 (0.27, 0.36)	582 (536, 627)	0.50 (0.47, 0.53)	825 (782, 868)	0.64 (0.61, 0.66)

Venous thromboembolism (VTE), area under the receiver operating characteristic curve (AUC) and 95% confidence interval (95% CI)

YEARS algorithm components: clinical signs or symptoms of deep vein thrombosis, haemoptysis, pulmonary embolism likely diagnosis.

*D-dimer thresholds expressed in ng/mL

Younger patients tend to have lower index test results (Supplementary Figure 1). There are also differences in the proportion with and without VTE in the two subgroups. This is an example of Scenario 3 (different distribution, different performance). Providing only the standard ROC curve analysis is not a fair representation of performance, as it ignores meaningful differences between age subgroups, in this case compromised performance in older patients.

In contrast, the covariate-specific ROC curves and AUCs were nearly identical between men and women, consistent with the similar distribution of test results in subgroups by sex (Supplementary Figure 1), and nearly overlapping CDF (Figure 1). We can assume that there are no meaningful differences based on sex, an example of Scenario 1 (identical distribution, identical performance). In this case, we can conclude the standard AUC fairly expresses the performance.

Patients with and without items of the YEARS had similar performance (Table 1), with some differences in the distribution of index test results (Figure 1). However, differences in the distribution are less noticeable in the lower ranges of the index test values (Supplementary Figure 1, F). This resembles Scenario 2 (different distribution, identical performance). The prevalence of the target condition differs drastically between the subgroups (8% vs 21%). The standard ROC curve analysis in this case may therefore be biased, as it does not consider underlying differences.

The former analyses considered the effect of a single covariate. It is also possible that there is an interaction between a pair of covariates, such as age and sex. By specifying the parameters of the regression model to include both covariates, we can model their effect on performance. In Supplementary Figure 2 we can see that discriminatory capacity of the index test is slightly lower in younger men and younger women, with no pronounced differences by sex. In cases where the effect is unclear, further significance testing should be performed (33).

Selection of covariate-specific positivity thresholds

We also compared standard vs. covariate-specific threshold values for desired performance levels (Table 1). We found different thresholds were necessary to achieve

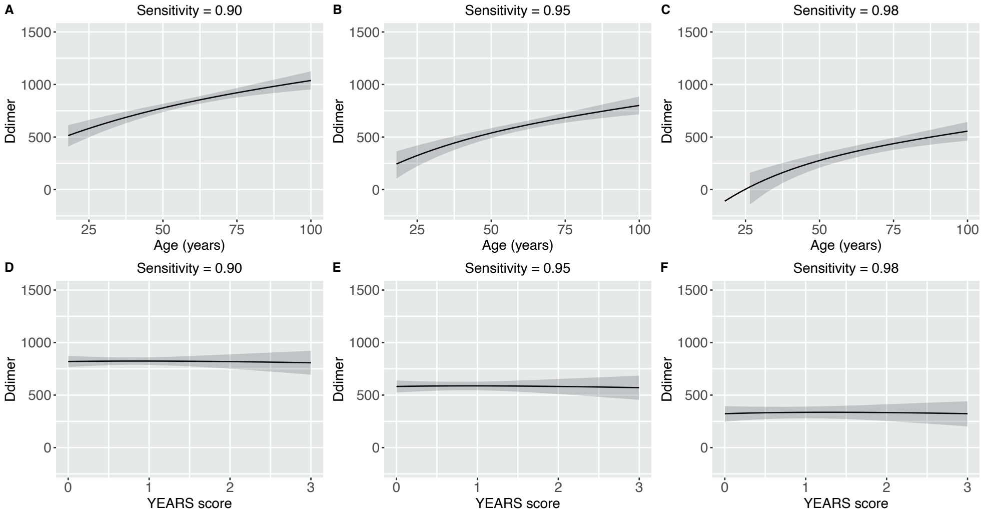


Figure 2. Threshold values for D-dimer, along age (A, B, C) and the YEARS score (D, E, F), modeling using the Bayesian nonparametric approach. Posterior mean (solid black line) and 95% pointwise credible band for D-dimer thresholds, corresponding to sensitivity of 0.98, 0.95 and 0.90.

the same level of performance. Taking age as an example, there is a difference of nearly 250 ng/mL in thresholds for younger versus older patients to achieve a sensitivity of 0.95.

Thus, higher positivity thresholds for D-dimer have to be selected for elderly patients, visually illustrated in Figure 2. This is consistent with the understanding that D-dimer levels increase with age. In such settings, recall Scenario 3, the covariate-specific ROC curves are different, and covariate-specific positivity thresholds should, ideally, be used. Differences were less prominent for other covariates. Looking at each point of the YEARS algorithm (scale of 0 to 3), the positivity thresholds are nearly identical, meaning the standard threshold would achieve the same sensitivity in both groups (Table 1). As mentioned previously, this may be explained by the similarities in distribution in the lower ranges, which corresponds to high sensitivity. In Figure 2 we can see that the same threshold would apply to any point on the YEARS algorithm to meet the desired sensitivity level. As no meaningful differences are observed, we can conclude that the ROC curves are identical.

Performance estimates with covariate-adjusted ROC curve analysis

In some cases, it may be informative to also present a covariate-adjusted ROC curve, one that takes covariate information into account: a weighted average of the covariate-specific ROC curves, with weights corresponding to the proportion of those with the target condition in the two subgroups. We constructed ROC curves that were adjusted for age, sex, the YEARS algorithm, and the combination of two covariates. The covariate-adjusted AUCs were almost identical to each other and reflected the standard pooled AUC of 0.87 (see Figure 3).

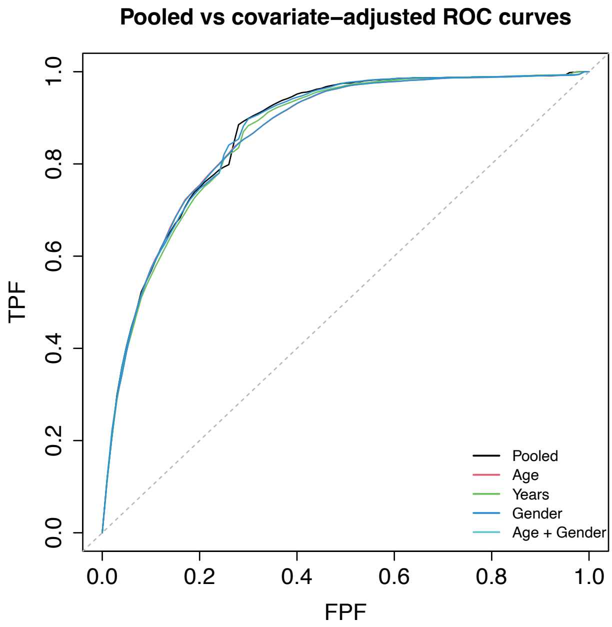


Figure 3. Standard (pooled) vs covariate-adjusted ROC curves.

Discussion

Most diagnostic accuracy studies that present ROC curves to summarize test performance ignore covariates. Yet constructing covariate-specific ROC curves can be informative for understanding the relationship between covariates and a test's performance and positivity threshold. There may be stratum-specific differences in performance that can influence further clinical decision making or, in the absence of any meaningful differences, we may conclude that the standard ROC curve produces fair estimates of performance.

Adjusting for covariate effects would be most necessary when comparing the performance of different tests, to alleviate any bias that may arise from unfair representation of patient characteristics where the performance may vary, as illustrated by Pepe et al. (3). We can also assume multicenter studies, with intrinsically different test settings, may benefit from adjusting for covariates. Yet comparisons between other techniques such as multi-level analysis or other proposed mixed methods for handling issues related to multicenter data are less established.

Importantly, this is an area that deserves more attention, particularly in the presence of clustered data. Covariate adjustment may be preferred with smaller sample sizes, which may be problematic for covariate-specific analysis. Covariate-adjusted ROC analyses can also consider continuous variables, in addition to categorical and binary ones.

Our study presents some limitations. In our IPD cohort, verification of the outcome was not the same for all patients in most studies and we relied on multiple reference standards. Imaging was performed for those with high clinical suspicion and/or high D-dimer, and clinical follow-up for those with low D-dimer levels.

In meta-analyses of diagnostic accuracy studies, heterogeneity in test performance is common across the primary studies. This may be due to patient or test characteristics and thus present a genuine difference in performance based on biology, but it may also be artefactual and due to study design flaws. Such artefactual factors can be identified with the QUADAS-2 tool (34). In the motivating example, we selected covariates based on a biologic basis, however, we can also utilize study design characteristics as covariates. Various approaches for including such covariates in meta-analysis of ROC curves have been proposed (35). Inclusion of artefactual covariates in conditional ROC curve analysis can also be incorporated to reflect, for example, center differences in multicenter diagnostic accuracy studies. This application warrants further exploration, as adjustment for center differences is another element of diagnostic accuracy research that remains less established.

We here presented results using a Bayesian nonparametric approach. Other methods, such as a semiparametric approach and a nonparametric kernel-based regression model, have been proposed (33). The Bayesian nonparametric approach is flexible to various distribution features, as it can adapt to skewness, nonlinearities, or data with higher variability. This makes it a practical choice for use with many different diseases and populations. The computational demand is however greater with this approach compared to some other models. We further note the limitations of methods such as the kernel-based regression models, which have long computation times and more limitations regarding number and type of covariates that can be included in the model. For an in-depth review, including an overview of various proposed statistical concepts and their application can be found in the work by In´acio and Rodríguez-Alvarez et al. and related publications (4, 5).

Incorporating covariate information into ROC curve analysis is not yet common practice, despite methods that have been proposed decades ago. We hope that the analysis presented here will lead to a more widespread application of such conditional ROC curves, which can provide more robust information on test performance and may improve our ability to select thresholds catered for specific subgroups.

References

1. Obuchowski NA, Bullen JA. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. Physics in Medicine & Biology. 2018;63(7):07TR1.

2. Pepe MS, Janes H, Li CI, Bossuyt PM, Feng Z, Hilden J. Early-phase studies of biomarkers: what target sensitivity and specificity values might confer clinical utility? Clinical chemistry. 2016;62(5):737-42.

3. Janes H, Pepe MS. Adjusting for Covariates in Studies of Diagnostic, Screening, or Prognostic Markers: An Old Concept in a New Setting. American Journal of Epidemiology. 2008;168(1):89-97.

4. Inácio V, Rodríguez-Álvarez MX. The covariate-adjusted ROC curve: the concept and its importance, review of inferential methods, and a new Bayesian estimator. Statistical Science. 2022;37(4):541-61.

5. Inácio V, Rodríguez-Álvarez MX, Gayoso-Diz P. Statistical evaluation of medical tests. Annual Review of Statistics and Its Application. 2021;8:41-67.

6. Rodríguez-Álvarez MX, Inacio V. ROCnReg: An R Package for Receiver Operating Characteristic Curve Inference with and without Covariate Information. arXiv preprint arXiv:200313111. 2020.

7. Metz CE, editor Basic principles of ROC analysis. Seminars in nuclear medicine; 1978: Elsevier.

8. Zweig M. ROC Plots: A Fundamental Evaluation Tool in Clinical Medicine/Zweig MH, Campbell G. Clinical Chemistry. 1993;39(4).

9. Martin KA, Molsberry R, Cuttica MJ, Desai KR, Schimmel DR, Khan SS. Time trends in pulmonary embolism mortality rates in the United States, 1999 to 2018. Journal of the American Heart Association. 2020;9(17):e016784.

10. Barco S, Mahmoudpour SH, Valerio L, Klok FA, Münzel T, Middeldorp S, et al. Trends in mortality related to pulmonary embolism in the European Region, 2000–15: analysis of vital registration data from the WHO Mortality Database. The Lancet Respiratory Medicine. 2020;8(3):277-87.

11. van der Hulle T, Cheung WY, Kooij S, Beenen LFM, van Bemmel T, van Es J, et al. Simplified diagnostic management of suspected pulmonary embolism (the YEARS study): a prospective, multicentre, cohort study. The Lancet. 2017;390(10091):289-97.

12. Wells PS, Ihaddadene R, Reilly A, Forgie MA. Diagnosis of venous thromboembolism: 20 years of progress. Annals of internal medicine. 2018;168(2):131-40.

13. Weitz JI, Fredenburgh JC, Eikelboom JW. A test in context: D-dimer. Journal of the American College of Cardiology. 2017;70(19):2411-20.

14. Van Es J, Beenen L, Douma R, Den Exter P, Mos I, Kaasjager H, et al. A simple decision rule including D-dimer to reduce the need for computed tomography scanning in patients with suspected pulmonary embolism. Journal of Thrombosis and Haemostasis. 2015;13(8):1428-35.

15. Konstantinides SV, Meyer G, Becattini C, Bueno H, Geersing G-J, Harjola V-P, et al. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS) The Task Force for the diagnosis and management of acute pulmonary embolism of the European Society of Cardiology (ESC). European heart journal. 2020;41(4):543-603.

16. Kabrhel C, Mark Courtney D, Camargo CA, Jr., Plewa MC, Nordenholz KE, Moore CL, et al. Factors associated with positive D-dimer results in patients evaluated for pulmonary embolism. Acad Emerg Med. 2010;17(6):589-97.

17. Harper P, Theakston E, Ahmed J, Ockelford P. D-dimer concentration increases with age reducing the clinical value of the D-dimer assay in the elderly. Internal medicine journal. 2007;37(9):607-13.

18. Righini M, Le Gal G, Perrier A, Bounameaux H. The challenge of diagnosing pulmonary embolism in elderly patients: influence of age on commonly used diagnostic tests and strategies. Journal of the American Geriatrics Society. 2005;53(6):1039-45.

19. Schutgens RE, Haas FJ, Biesma DH. Reduced efficacy of clinical probability score and D-dimer assay in elderly subjects suspected of having deep vein thrombosis. British journal of haematology. 2005;129(5):653-7.

20. Schouten HJ, Geersing GJ, Koek HL, Zuithoff NPA, Janssen KJM, Douma RA, et al. Diagnostic accuracy of conventional or age adjusted D-dimer cut-off values in older patients with suspected venous thromboembolism: systematic review and meta-analysis. BMJ : British Medical Journal. 2013;346:f2492.

21. Douma RA, Le Gal G, Söhne M, Righini M, Kamphuisen PW, Perrier A, et al. Potential of an age adjusted D-dimer cut-off value to improve the exclusion of pulmonary embolism in older patients: a retrospective analysis of three large cohorts. Bmj. 2010;340.

22. Den Ouden M, Ubachs JH, Stoot J, Van Wersch J. Thrombin-antithrombin III and D-dimer plasma levels in patients with benign or malignant ovarian tumours. Scandinavian journal of clinical and laboratory investigation. 1998;58(7):555-60.

23. Douma RA, van Sluis GL, Kamphuisen PW, Söhne M, Leebeek FW, Bossuyt PM, et al. Clinical decision rule and D-dimer have lower clinical utility to exclude pulmonary embolism in cancer patients. Explanations and potential ameliorations. Thromb Haemost. 2010;104(4):831-6.

24. Geersing GJ, Kraaijpoel N, Büller HR, van Doorn S, van Es N, Le Gal G, et al. Ruling out pulmonary embolism across different subgroups of patients and healthcare settings: protocol for a systematic review and individual patient data meta-analysis (IPDMA). Diagn Progn Res. 2018;2:10.

25. Bossuyt PM, Olsen M, Hyde C, Cohen JF. An analysis reveals differences between pragmatic and explanatory diagnostic accuracy studies. Journal of Clinical Epidemiology. 2020;117:29-35.

26. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ. 2002;324(7338):669-71.

27. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6(11):e012799.

28. de Carvalho VI, Jara A, Hanson TE, de Carvalho M. Bayesian nonparametric ROC regression modeling. Bayesian Analysis. 2013;8(3):623-46.

29. De Iorio M, Johnson WO, Müller P, Rosner GL. Bayesian Nonparametric Nonproportional Hazards Survival Modeling. Biometrics. 2009;65(3):762-71.

30. Janes H, Pepe MS. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. Biometrika. 2009;96(2):371-82.

31. Pepe MS. Three Approaches to Regression Analysis of Receiver Operating Characteristic Curves for Continuous Test Results. Biometrics. 1998;54(1):124-35.

32. Rodríguez-Álvarez MX, Vanda I. ROCnReg: an R package for receiver operating characteristic curve inference with and without covariates. 2021.

33. Rodríguez-Álvarez MX, Roca-Pardiñas J, Cadarso-Suárez C. ROC curve and covariates: extending induced methodology to the non-parametric framework. Statistics and Computing. 2011;21(4):483-99.

34. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36.

35. Doebler P, Holling H. Meta-analysis of diagnostic accuracy and ROC curves with covariate adjusted semiparametric mixtures. Psychometrika. 2015;80(4):1084-104.

Supplementary Material

Supplementary Table 1. List of study team members

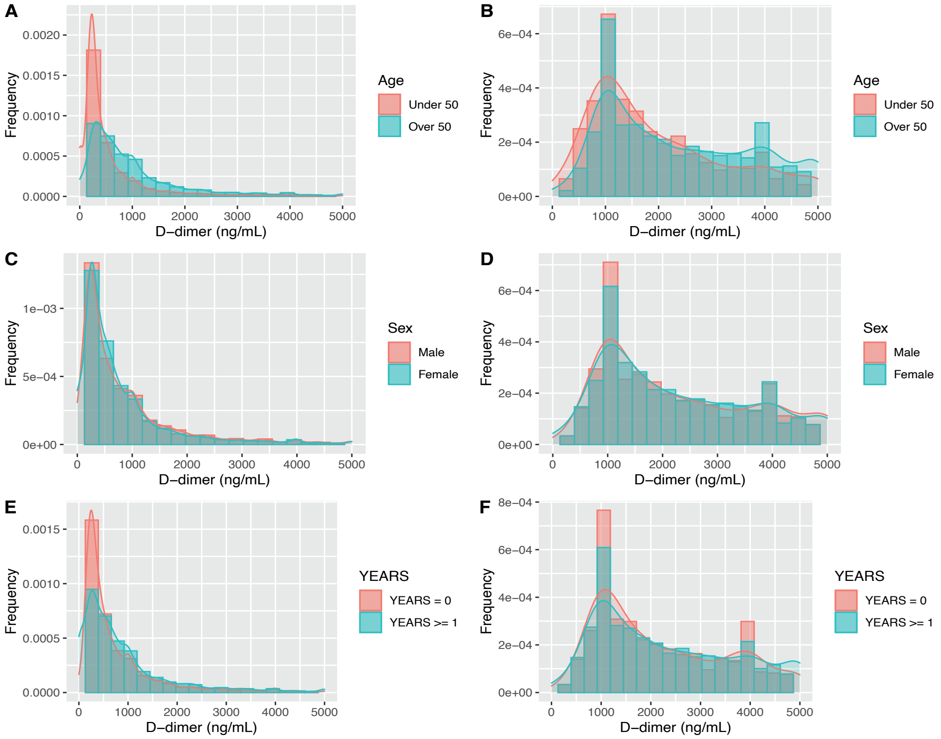
Study team	Geert-Jan Geersing 1 Toshihiko Takada 1 2 Frederikus A Klok 3 Harry R Büller 4 D Mark Courtney 5 Yonathan Freund 6 Javier Galipienzo 7 Gregoire Le Gal 8 Waleed Ghanima 9 Jeffrey A Kline 10 Menno V Huisman 3 Karel G M Moons 1 11 Arnaud Perrier 12 Sameer Parpia 13 14 Helia Robert-Ebadi 12 Marc Righini 12 Pierre-Marie Roy 15 Maarten van Smeden 1 Milou A M Stals 3 Philip S Wells 8 Kerstin de Wit 14 16 Noémie Kraaijpoel 4 Nick van Es 4
Affiliations	1 Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands. 2 Department of General Medicine, Shirakawa Satellite for Teaching And Research (STAR), Fukushima Medical University, Fukushima, Japan. 3 Department of Medicine, Thrombosis and Haemostasis, Dutch Thrombosis Network, Leiden University Medical Center, Leiden, the Netherlands. 4 Department of Medicine, Amsterdam University Medical Center, Amsterdam Cardiovascular Sciences, Amsterdam, the Netherlands. 5 Department of Emergency Medicine, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America. 6 Sorbonne University, Emergency Department, Hôpital Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, Paris, France. 7 Service of Anesthesiology, MD Anderson Cancer Center Madrid, Madrid, Spain. 8 Department of Medicine, University of Ottawa, Ottawa Hospital Research Institute, Ottawa, Canada. 9 Department of Medicine, Østfold Hospital Trust, Norway and Institute of Clinical Medicine, University of Oslo, Oslo, Norway. 10 Department of Emergency Medicine, Wayne State School of Medicine, Detroit, Michigan, United States of America. 11 Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands. 12 Division of Angiology and Hemostasis, Geneva University Hospitals and Faculty of Medicine, Geneva, Switzerland. 13 Department of Oncology, McMaster University, Hamilton, Canada.

	14 Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada. 15 UNIV Angers, UMR (CNRS 6015-INSERM 1083) and CHU Angers, Department of Emergency Medicine, F-CRIN InnoVTE, Angers, France. 16 Department of Emergency Medicine, Queen's University, Kingston, Canada.

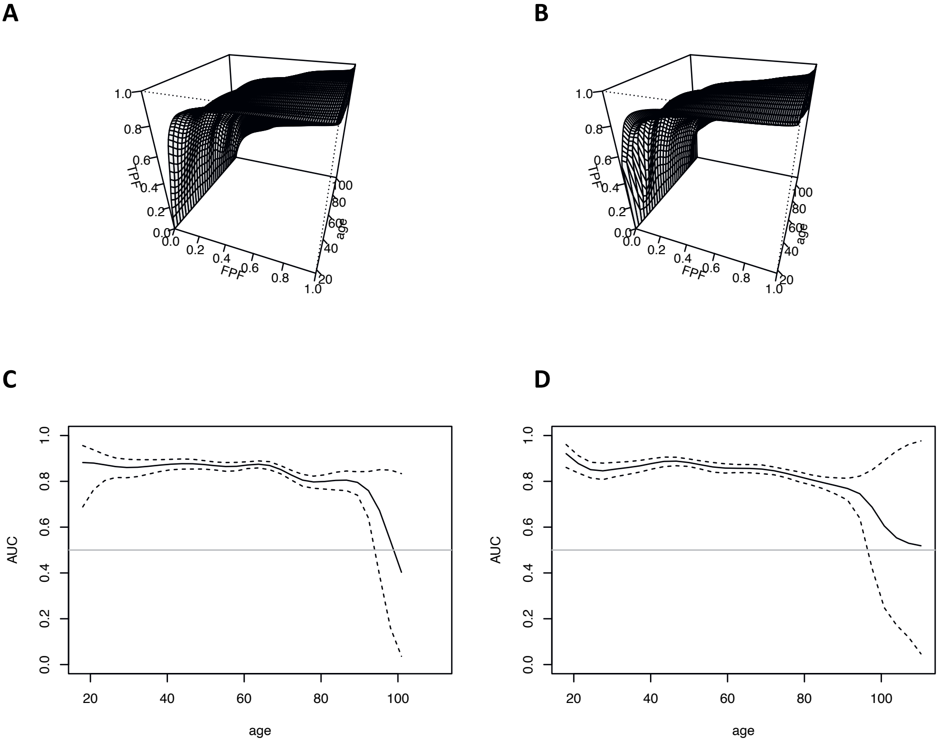
Supplementary Table 2. Characteristics of the study group

	Individual patient data
Total	21621
Females	13334 (61.7%)
Mean age, years (SD)	54.58 (18.29)
Median D-dimer, ng/mL (IQR)	630 (285, 1402)
VTE	3150 (14.6%)
Clinical signs of DVT	1454 (6.7%)
Hemoptysis	875 (4.0%)
Inpatient setting	1344 (6.2%)
Active cancer	1939 (9.0%)
YEARS score	
0	11114 (51.4%)
1	9430 (43.6%)
2	1049 (4.9%)
3	28 (0.1%)

All values expressed numbers, unless otherwise noted
Standard deviation (SD), interquartile range (IQR), venous thromboembolism (VTE), deep vein thrombosis (DVT)



Supplementary Figure 1. Histograms for D-dimer frequency by covariate in those without VTE (left) and with VTE (right)



Supplementary Figure 2. Graphical results for covariate-specific ROC, including interaction between age and sex. Top row: Posterior mean of the covariate-specific ROC curve along age, separately for men (A) and women (B). Bottom row: Posterior mean and 95% pointwise credible band for the covariate-specific AUC along age, separately for men (C) and women (D). True positive fraction (TPF), false positive fraction (FPF), area under the receiver operating characteristic curve (AUC).

09

Future prospects and opportunities

General Discussion

This thesis focuses on handling challenges related to bias and variability in test accuracy research. We emphasized the importance of careful methodological consideration when designing and executing test accuracy studies, based on observations that important elements associated to test performance are frequently overlooked.

The journey from biomarker discovery to implementation is a long and arduous process, with many markers failing to meet the standards for regulatory approval and clinical use (1). Oftentimes the initial enthusiasm of promising biomarker performance is met with attenuated accuracy with external validation. Our systematic reviews and meta-analyses on the performance of non-alcoholic fatty liver disease (NAFLD) biomarkers confirmed the observation that, over repeated validation studies, many were found to have more modest performance (Chapter 2) (2, 3) . This was further confirmed in our large external validation of seventeen different biomarkers; most did not meet our performance prerequisite - we selected a priori at an area under the receiver operating characteristic curve (AUC) of 0.80 - despite their selection based on pathophysiologic understanding (Chapter 6).

We also consistently observed that many test accuracy studies are conducted with a simplistic, one-size-fits-all mindset. If we consider a continuous biomarker, we can imagine that even the decision of selecting a positivity threshold will generally require different considerations. On whom will the test be used? In what setting? For what purpose? Such factors are valuable for balancing the trade-off between the sensitivity and specificity of a test. Yet, in many published test accuracy studies, meaningful and sometimes cross-level interactions between patient characteristics, center differences, index test properties, and disease prevalence remain ignored despite documented implications of such issues in the literature (4, 5).

These issues are not unique to test accuracy research, as interventional studies come with their own set of biases and variation, but methodology to address these concerns is less developed or applied in test evaluation. Take for example adjusting for covariate or center differences: these factors are regularly controlled for in interventional studies, but sparsely regarded when evaluating test performance. That is not to say there haven't been advancements in methodology. Novel approaches have been proposed but are slow to adaptation. This leaves the question: why? What are the barriers to their implementation?

Outside of clinical practice, biomarkers are increasingly studied to serve a variety of purposes in drug development, for example to enrich the clinical trial population with those more likely to benefit from treatment. In therapeutic trials for non-alcoholic steatohepatitis (NASH), liver biopsies are required to enroll eligible patients, but this comes at the cost of high screen failure rates, partially due to inherent limitations of liver biopsy itself (6) but also because of flaws in the process of selecting patients who will undergo this procedure. In the LITMUS (Liver Investigation: Testing Marker Utility in Steatohepatitis) consortium (7), we examined the ability of non-invasive tests to preselect those who would undergo biopsy, thereby improving patient selection efficiency, by devising a novel strategy for selecting a positivity threshold which corresponds to a desired screen failure rate. Providing the evidentiary standards to demonstrate utility requires careful consideration of appropriate study design methodology, and, where there is need, development of innovative strategies to realize the unmet needs.

While it is an impossible task to eliminate all sources of bias or variability in test accuracy studies, more contentious consideration of appropriate methods can improve the reliability of their results. In the following section we outline points that may be considered in future research to improve the evaluation of medical tests.

Future prospects

Expanding the scope of QUAPAS

In Chapter 4, we reported a systematic review on the prognostic accuracy of commonly used liver fibrosis scores in prognosticating liver related outcomes. Risk of bias and applicability assessment is an essential step in the evidence synthesis process as weaknesses in study design, conduct, and statistical analysis can produce misleading results and, moreover, waste valuable resources. Various risk of bias tools have been developed but no tool to critically evaluate prognostic accuracy existed. Most publications relied on existing tools, such as QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2), developed for diagnostic test accuracy studies (8). While this may be an intuitive choice, it is problematic as there are different sources of bias that arise when evaluating the prognostic ability of a test compared to the cross-sectional examination of diagnostic performance.

To improve bias assessment in prognostic accuracy studies, we modified the QUADAS-2 tool to develop QUAPAS (Quality Assessment of Prognostic Accuracy Studies) (Chapter 5) (9). The tool as it stands is intended to examine bias and applicability in prognostic accuracy studies, where the performance of a single index test is evaluated in reference to the occurrence of a future event. Analogous to tests used for a diagnostic purpose, oftentimes we are interested in comparing the prognostic performance between several tests. When there are many competing tests, knowledge of their relative performance provides the evidence based required to make informed decisions and recommendations.

For diagnostic accuracy studies, the understanding that different sources of bias can arise when designing studies addressing comparative accuracy questions was the precursor for developing QUADAS-C (Quality Assessment of Diagnostic Accuracy Studies–Comparative) (10). Much like the diagnostic counterpart, QUAPAS would benefit from an extension that allows more systematic assessment of biases that arise in a comparative setting. Empirical studies that examine study design elements that can introduce bias, and the magnitude of their effect on performance, can supplement the development of such an extension.

The field could further benefit from systematically evaluating sources of bias that occur in prognostic accuracy studies in general, as one of the limitations in developing QUAPAS was the reliance on theoretical knowledge from experts in the field in the absence of existing empirical data from meta-epidemiologic studies. Resources like the work by Whiting et al. could support further refinement of QUAPAS (11).

Piloting QUAPAS, we found that risk of bias assessment, particularly in the analysis domain, was hampered by incomplete reporting (12). Reporting guidelines and risk of bias assessment tools go hand-in-hand. Reviewers are unable to detect study design flaws or applicability concerns if they are not clearly reported, this was the precursor for developing the STARD (Standards for Reporting of Diagnostic Accuracy Studies) guideline for diagnostic accuracy studies (13). In this regard, we highlight the importance of primary study investigators' adherence to appropriate reporting guidelines, to promote overall transparency and reproducibility, as well as to assist reviewers in conducting quality assessment. Further modification of STARD for use in prognostic accuracy studies may alleviate some of the challenges we faced when piloting QUAPAS.

Barriers to implementing appropriate methodology

Scenario 1. The knowledge gap

There may be different barriers that interfere with implementation of appropriate methods in test accuracy research. The problem may simply be ignorance. Clinical and laboratory researchers may lack specialized training in test accuracy methodology, contributing to inadequate emphasis on the importance of accounting for factors that modify test performance. Studies that rely on the conventional framework for evaluating test accuracy, as one would find in an introductory clinical epidemiology textbook, do not allow us to fully appreciate the differences in performance across unique patient characteristics or settings.

A practical solution to closing the knowledge gap is to involve experienced statisticians and methodologists as part of the wider study team. This, paired with the expertise of clinicians and stakeholders in related industries, can allow a more rigorous consideration and implementation of appropriate methods.

Medical programs may consider implementing a more structured research training program, for example in the form of a structured research rotation (14). Clinical researchers may individually elect more specialized training, one which emphasizes the evaluation of medical tests. This can better enforce core principles of evidence-based practices and support the development of necessary skills for both critically conducting and interpreting test accuracy studies.

Scenario 2. Fragmented knowledge

In some cases, novel solutions to address analytical concerns have been developed but these do not reach the clinical audience. This is often a result of the gap in disciplines. For example, methods that account for the time dependency in evaluating prognostic accuracy are available (15), but a quick search of the literature will demonstrate that these proposals are not applied in practice. In some cases, there may be many different methods proposed, for example to conduct covariate-adjusted ROC curve analysis, but no clear guidance on appropriate scenarios in which they should be applied. This, paired with far fewer publications that compare related methods to demonstrate the relative merits of each approach, as well as key distinguishing factors, can make decision making arduous for a beginning and perhaps even experienced researchers.

To bridge the gap between statistical and clinical disciplines, methodologists can develop resources to support the design of test accuracy studies. This was the intention behind Chapter 8. We aimed to explain the added value of conducting conditional ROC curve analysis, supplemented by a practical application of novel statistical methods in a digestible manner. The didactic nature of this paper is catered to a clinical audience so that they may better consider and apply appropriate methods that have been proposed and make valid interpretations. We also point to more detailed resources, such as the work by Kamarudin et al., that landscape current methods for time-dependent ROC curve analysis and provide guidance on their application (16).

The development of methodological guidelines for clinical research should be emphasized as much as clinical guidelines. Oftentimes we reward and incentivize new and novel ideas, however, quality resources such as the rigorous comparison of methods are also warranted.

Scenario 3. Lack of resources

Another barrier behind inadequate study methodology may be that appropriate solutions have not yet been developed to address outstanding issues in test accuracy research. The element of center effects, for example, is frequently ignored by clinical researchers, and literature describing challenges behind multicenter test accuracy studies that, when ignored, can lead to misleading conclusions is largely absent. Now there is even greater interest in the combination of markers to detect diseases, yet the implications for developing and validating such models using data from multiple centers does not appear to be widely appreciated. Future studies can direct attention towards evaluating the role of center effects in evaluating test accuracy and develop ways in which existing methods, for example those used in therapeutic trials, can be applied in test accuracy studies.

Methodology for evaluating test accuracy beyond the diagnostic context is far less developed. Most attention in test accuracy research is on the diagnostic context of use, leaving scarcity of methods in other areas, such as monitoring. There is a level of complexity introduced when expressing accuracy in the monitoring context of use. This will reflect not only the repeated measurements over time, but differences in the characteristics of the condition or outcome being monitored. A test may be used to monitor a patient for an imminent event, for early detection of a condition, or for rejection of a graft, to list a few examples. With the different scenarios in which monitoring is necessary, methods for expressing performance in such scenarios deserves more specialized attention.

Where there is room for improvement, statisticians and methodologists can develop and propose new approaches that have undergone rigorous testing. The implementation of novel approaches can then be facilitated by following the suggestions discussed above, emphasizing the role of methodologists in bridging the gap to between statisticians and clinicians.

Leveraging machine learning

In the NAFLD space, there is a big pool of biomarkers and multimarker scores that have been proposed but, for the majority of the evaluated tests, we found suboptimal performance (Chapter 6). Even multimarker scores, which generally perform better than single markers, had disappointing accuracy in detecting 'at-risk NASH', indicating that there is still room for improvement in terms of how we develop diagnostic models. That was the focus of Chapter 7. We employed supervised machine learning algorithms to develop a series of models in detecting stages of NAFLD. In a direct comparison of the multimarker scores developed using simple regression approaches and those derived using machine learning algorithms, the latter models had superior performance.

We focus here on the application of machine learning techniques to address two fundamental shortcomings in test accuracy research: development of diagnostic models and imperfect reference standards. While this thesis evaluated the application of machine learning in the context of developing and evaluating tests used to detect NAFLD conditions, we may presume these challenges may be synonymous to test development and evaluation in related fields.

Development of diagnostic models

Many of the challenges we face when developing diagnostic models are related to data. Prior to designing a study, investigators may ponder whether the available data is of sufficient quality and scope. While there are ways to address incomplete data, like multiple imputation, the quality of the data may still be compromised, resulting in a lot of noise. One of the fundamental capabilities of machine learning, and its distinguishing factor from statistics alone, is the emphasis on making predictions by finding patterns within the data. This makes the application of machine learning for diagnostics particularly interesting, as the intricate algorithms may better detect signals amidst noise in an imperfect dataset. They may also be better suited to detect diseases with complex etiologies, such as NAFLD.

Ideally, diagnostic models are built using a prospectively recruited study group from a well-defined patient population, where predictor data are preselected based on underlying biologic plausibility. But constructing such cohorts can be extremely resource intensive and, in some cases, impractical. A sensible alternative is to utilize existing data sources, such as registry or electronic health record (EHR) data. Machine learning algorithms may still be able to detect and combine signals from different patterns within the scope of the available data. We saw a glimpse of this in our own analysis. We developed models with two different sets of predictors, one which only utilized routinely collected clinical data, and a second set that included novel blood-based biomarkers with hypothesized involvement in disease progression. Much to our surprise, the models developed with only the clinical parameters performed just as well, and even marginally better, compared to the models that included the more specialized, and more resource intensive, biomarkers.

This area of research still requires more refinement. With the availability of many different machine learning algorithms, more work is needed to understand which algorithms and accompanying study design elements will produce the most optimal and parsimonious models. This should be paired with a clear emphasis on the context in which the test will be applied in practice.

We have established the importance of evaluating medical tests in appropriate settings and patient profiles, and this applies to the cohort used for the development and validation of such diagnostic models. In the example above, we discuss developing models using EHR or similar data. It is important to highlight here that models developed using such data sources are then only generalizable to the setting it represents, in this case a more general primary care setting. This can however be regarded as an advantage of applying machine learning to such data sources, as primary care settings are likely those with less diagnostic resources, compared to specialized tertiary care setting's access to specialists as well as more advanced diagnostic technologies.

Training clinically useful machine learning models will depend on datasets of sufficient size and representativeness. In the LITMUS consortium, data from 39 centers in 13 different countries were pooled together to form a single, harmonized dataset. This was essential for allocating sufficient patient data, but came at the high cost of legal and logistical obstacles that were extremely resource intensive, and dependent on the willingness of respective stakeholders to collaborate and share data. In the future, we can look towards innovative ways of training and validating models such as the use of centralized repositories and blockchain-based technology. In a recent publication, investigators locally trained machine learning models that were centrally combined by use of swarm learning, applied to histopathology images for the detection of solid tumors (17). They found that the swarm learning trained models outperformed the local models and were comparable to the models trained in the learning dataset. Utilizing such technologies can alleviate the many road-blocks related to data privacy and sharing, as well as maintenance of centralized patient cohorts, without compromising accuracy.

The imperfect reference standard

The application of machine learning can extend beyond developing diagnostic models. Their use be can leveraged to address limitations we face when attempting to evaluate the performance of a medical test against an imperfect reference standard. This is a prevalent issue in test accuracy studies. In the NAFLD space, there are several studies that have demonstrated high inter/intra observer variability when it comes to interpreting liver histology data (6). The strength of the test accuracy results is thus relative to the accuracy of the histological diagnosis, and limited by variability in the interpretation of histological features. In the absence of a perfect reference standard, this element will always remain a barrier in obtaining valid results.

Machine learning techniques have already gained popularity to address this caveat, with several solutions approved by the FDA for use in medical imaging (18). Their use has proliferated to support radiologists and histopathologists in not only automating manual tasks in imaging analysis, but also in detecting patterns and information that may be missed by the human eye (19). Studies have also been conducted to measure specific features of NAFLD conditions using convolutional neural network algorithms on digitalized histopathology images (20). Such machine learning based predictions can offer a potential solution for improving the reliability and reproducibility of the reference standard results, which remains a limitation in majority of studies evaluating accuracy of NAFLD biomarkers.

Concluding remarks

Understanding the full capacity of test performance is not possible without, at least in part, systematically considering factors that can modify performance or produce biased results. These efforts are beneficial for expanding ones understanding of the operational characteristics of medical tests and how tests can be optimized for implementation for specific purposes, settings or patient profiles. Methodologists play a key role in the development of both new methodology and resources for guiding the implementation of such innovations in future test accuracy studies. This can promote generation of more reliable data on biomarker performance, which are required both for use in clinical practice and for successful vetting by regulatory agencies. We hope the work in this thesis raises greater awareness of the intricacies in test accuracy evaluation, and promotes more conscientious consideration of novel methods and technologies in future works.

References

1. Ioannidis JP, Bossuyt PM. Waste, leaks, and failures in the biomarker pipeline. Clinical chemistry. 2017;63(5):963-72.

2. Lee J, Vali Y, Boursier J, Duffin K, Verheij J, Brosnan MJ, et al. Accuracy of cytokeratin 18 (M30 and M65) in detecting non-alcoholic steatohepatitis and fibrosis: A systematic review and meta-analysis. PLoS One. 2020;15(9):e0238717.

3. Vali Y, Lee J, Boursier J, Spijker R, Löffler J, Verheij J, et al. Enhanced liver fibrosis test for the non-invasive diagnosis of fibrosis in patients with NAFLD: a systematic review and meta-analysis. Journal of hepatology. 2020;73(2):252-62.

4. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. Journal of clinical epidemiology. 2009;62(1):5-12.

5. Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J. Sources of Variation and Bias in Studies of Diagnostic Accuracy. Annals of Internal Medicine. 2004;140(3):189-202.

6. 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.

7. Hardy T, Wonders K, Younes R, Aithal GP, Aller R, Allison M, et al. The European NAFLD Registry: A real-world longitudinal cohort study of nonalcoholic fatty liver disease. Contemp Clin Trials. 2020;98:106175.

8. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36.

9. Lee J, Mulder F, Leeflang M, Wolff R, Whiting P, Bossuyt PM. QUAPAS: An Adaptation of the QUADAS-2 Tool to Assess Prognostic Accuracy Studies. Annals of Internal Medicine. 2022;175(7):1010-8.

10. Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF, et al. QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. Annals of internal medicine. 2021;174(11):1592-9.

11. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Group Q-S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. Journal of clinical epidemiology. 2013;66(10):1093-104.

12. Lee J, Vali Y, Boursier J, Spijker R, Anstee QM, Bossuyt PM, et al. Prognostic accuracy of FIB-4, NAFLD fibrosis score and APRI for NAFLD-related events: a systematic review. Liver International. 2021;41(2):261-70.

13. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clinical chemistry. 2015;61(12):1446-52.

14. Kanna B, Deng C, Erickson SN, Valerio JA, Dimitrov V, Soni A. The research rotation: competency-based structured and novel approach to research training of internal medicine residents. BMC medical education. 2006;6(1):1-8.

15. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000;56(2):337-44.

16. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC medical research methodology. 2017;17(1):1-19.

17. Saldanha OL, Quirke P, West NP, James JA, Loughrey MB, Grabsch HI, et al. Swarm learning for decentralized artificial intelligence in cancer histopathology. Nature Medicine. 2022;28(6):1232-9.

18. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ digital medicine. 2020;3(1):1-8.

19. Kather JN, Calderaro J. Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. Nature Reviews Gastroenterology & Hepatology. 2020;17(10):591-2.

20. Taylor-Weiner A, Pokkalla H, Han L, Jia C, Huss R, Chung C, et al. A machine learning approach enables quantitative measurement of liver histology and disease monitoring in NASH. Hepatology. 2021;74(1):133-47.

APPENDIX

Summary / Samenvatting
List of publications
Portfolio
Letter of thanks
About the author

Summary

Researchers face a variety of challenges in biomarker and test evaluation research. Robust study design and analytical decisions can provide potential solutions or alleviate some of the problems associated to bias and variability in test performance. In this thesis we focused on addressing some of the challenges in the evidence synthesis and generation of test accuracy data.

In **Chapter 2** we conducted a systematic review and meta-analysis on the accuracy of a non-invasive biomarker, circulating cytokeratin-18 (CK-18). We aimed to provide summary estimates with increased precision for the accuracy of CK-18's two antigens (M30 and M65) in detecting non-alcoholic steatohepatitis (NASH) and fibrosis. To produce summary measures with sufficient granularity, meta-analyses were performed for five groups based on the CK-18 antigen and target condition, using one of two methods: linear mixed-effects multiple thresholds model or bivariate logit-normal random-effects model. The mixed effects multiple thresholds model was selected to account for heterogeneous reporting of positivity threshold values, a common occurrence for biomarkers on a continuous scale. This approach further allowed modeling of predictive values across a range of disease prevalence. Among the 41 included primary studies, we found modest performance for both CK-18 antigens as a stand-alone test. However, they have the potential to reach high negative predictive values in low prevalence settings that likely mirror a primary care center, indicating potential for excluding those without the disease. Primary studies further demonstrated that the value of CK-18 can be maximized when used in combination with other synergistic markers, as multi-marker scores that included CK-18 had higher accuracy among non-alcoholic fatty liver disease (NAFLD) patients.

Various imaging modalities can also be used to stage liver fibrosis. In **Chapter 3** we conducted an individual patient data (IPD) analysis on the ability of liver stiffness measurement by vibration controlled transient elastography (LSM-VCTE), and common multimarker scores, Fibrosis-4 index (FIB-4) and NAFLD Fibrosis Score (NFS), to propose diagnostic strategies that could reduce the need for liver biopsies, which remain the reference standard for staging fibrosis. Considering the suboptimal performance of stand-alone non-invasive tests, here the biomarkers were assessed both individually and as part of a sequential testing strategy. We further aimed to better understand factors that interact with diagnostic performance. Our analysis showed that while accuracy was good

for LSM-VCTE alone, the sequential combination of markers increased sensitivity and specificity, with the lower threshold ruling out cases of advanced fibrosis and a higher threshold ruling in cirrhosis. The sequential strategy also reduced the number of intermediary cases that will require further testing, likely a liver biopsy. Subgroup analysis further demonstrated that performance was variable across different subgroups based as body mass index (BMI) and age. This study can serve as a benchmark for future testing strategies that consider newer multi-marker scores for the staging of fibrosis, although greater emphasis on upper and lower thresholds is warranted and our findings could benefit from further validation.

A test's ability to prognosticate future events can support disease management and risk stratification. FIB-4, NFS and APRI are multimarker-scores commonly used for detecting fibrosis among NAFLD patients. **Chapter 4** focuses on synthesizing the available data on the accuracy of these models in prognosticating NAFLD-related events. This systematic review demonstrated that all three markers performed well in predicting liver-related events. However, the markers had highly inconsistent performance in prognosticating changes in fibrosis stage, which may, in part, be explained by the differences in the time horizon, definition of the target event, and baseline disease prevalence, as well as patient and center differences. Building on the understanding that fibrosis is strongest predictor for long-term clinical outcomes in NAFLD patients, future studies can focus on evaluation the comparative accuracy between liver biopsy and non-invasive tests. Comparable performance between histology-confirmed fibrosis and non-invasive tests may support the implementation of such biomarkers for risk stratification. The vast level of heterogeneity did not allow calculation of summary measures, however, rather highlighted several shortcomings in that can be better addressed in future studies.

One of the challenges in Chapter 4 was the absence of an appropriate risk of bias and applicability tool to critically assess the included prognostic accuracy studies. This inspired the work in **Chapter 5**, where we utilized an existing tool called QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) to develop QUAPAS (Quality Assessment of Prognostic Accuracy Studies), an adaptation of QUADAS-2 for prognostic accuracy studies. Studies that evaluate prognostic accuracy have key distinguishing study design elements compared to diagnostic accuracy studies, as different sources of biases accompany such longitudinal study designs, while aspects of evaluating test accuracy, in some domains, may overlap with evaluation of its diagnostic performance. The tool was developed using the framework of QUADAS-2, combining questions likely to identify bias evaluated and collated from QUIPS (Quality in Prognosis Studies) and PROBAST (Prediction Model Risk of Bias Assessment Tool). QUAPAS follows the same steps as QUADAS-2. Risk of bias is judged in 5 domains: participants, index test, outcome, flow and timing, and analysis. Signaling questions assist the final judgment for each domain. Applicability concerns are assessed for the first 4 domains. Compared to QUADAS-2, QUAPAS was able to identify studies at risk of bias that were not captured before. The reliability of risk of bias tools is dependent on proper reporting, we found that this hampered the bias judgement for the analysis domain and is an area for improvement. Future meta-epidemiologic studies that systematically study biases in prognostic accuracy studies can supplement future refinement of the tool.

Following the evidence synthesis phase, in **Chapter 6**, we evaluated the diagnostic accuracy of seventeen biomarkers, multimarker scores, and vibration-controlled transient elastography (VCTE) for the detection of at-risk NASH and advanced fibrosis, using data from the LITMUS (Liver Investigation: Testing Marker Utility in Steatohepatitis) Metacohort, an international cohort of biopsy-confirmed NAFLD patients. We secondarily aimed to increase the efficiency of future therapeutic trial enrolment by identifying thresholds for each marker that meet the acceptable screen failure rate. Performance was expressed as the area under the ROC curve (AUC), using biopsy as the reference standard, and compared against the Fibrosis-4 Index for Liver Fibrosis (FIB-4) in the same subgroup. Among 966 patients included in the analysis, no single biomarker or multi-marker significantly exceeded the predefined AUC 0.80 acceptability threshold, with performance comparable to that of FIB-4 for detection of at-risk NASH. SomaSignal, ADAPT score and liver stiffness measurement were able to detect advanced fibrosis with better accuracy. We further found that biopsy screen failure rates could be minimized to one third for trial recruitment, if only marker-positives undergo biopsy. The performance of biomarker performance and pre-screening strategy for identifying at-risk NASH patients for clinical trials will be further evaluated.

In **Chapter 7**, to address the challenges behind accurate detection of NASH and at-risk NASH, we employed a supervised machine learning algorithm called gradient boosting machine (GBM) to develop a diagnostic model. Utilizing the Metacohort as the learning data (training set 75%, validation set 25%), 35 predictors representing both clinical and biomarker data were included to train models for detecting NASH, at-risk NASH, and fibrosis stages. Missing data were handled by multiple imputation. Two models were trained for each condition: clinical versus extended (clinical data and biomarkers). Two variants of the NASH and at-risk NASH models were constructed: direct and composite models. We found the clinical and extended models were comparable, indicating no improvement with the addition of biomarkers in the model training for NASH, although small improvements were seen for fibrosis. The composite approach, which aggregates independent GBM models trained for each component of NASH/at-risk NASH, significantly improved detection compared to the direct approach. The composite GBM models outperformed existing single and multi-marker scores, both established and newly proposed marker combinations for each respective target condition. The models will be validated in the prospective LITMUS cohort.

Lastly, in **Chapter 8** we provide an illustrative application of a Bayesian nonparametric approach in evaluating biomarker performance incorporating covariate information as diagnostic accuracy is not a fixed property but may be associated to several factors related study or patient characteristics. Three scenarios were outlined based on a comparison between covariate subgroups and conclusions regarding the standard ROC curve. They are Scenario (1) identical distribution, identical performance, Scenario (2) different distribution, identical performance, and Scenario (3) different distribution, different performance. Differences in performance and distribution of results between covariate subgroups can indicate that the conventional ROC curve is not a fair representation of test performance. We then analyzed individual patient data (IPD) on D-dimer testing for excluding pulmonary embolism. Covariate-specific and covariate-adjusted ROC curve analyses were performed to examine performance and positivity thresholds of D-dimer in each covariate subgroup. We observed different scenarios when considering age, sex and pre-test probability, dependent on index test concentration and performance between covariate subgroups. Similarly, between some covariate subgroups, different positivity thresholds were necessary to achieve identical sensitivity. Application of conditional ROC curves can improve our ability understand variability in test performance, and improve threshold selection with improved applicability.

Chapter 9 presents a general discussion of the work reported in this thesis and discusses prospects for future studies and development.

Samenvatting

Onderzoekers staan voor meerdere uitdagingen als ze biomarkers en tests willen evalueren. Met een robuuste onderzoeksopzet en gepaste statistische analyses kunnen ze de kans op vertekening verkleinen en rekening houden met variabiliteit in testprestaties. Het onderzoek dat in dit proefschrift staat beschreven kwam uitdagingen tegen bij de synthese van eerder gerapporteerd onderzoek en bij primair, nieuw onderzoek naar de prestaties van biomarkers als medische tests. Dat deze we vooral binnen onderzoek naar bestaande en nieuwe tests voor patiënten met niet-alcoholische leververvetting (NAFLD).

In **Hoofdstuk 2** rapporteren we een systematisch literatuuronderzoek met meta-analyse naar de nauwkeurigheid van een niet-invasieve biomarker: circulerend cytokeratine-18 (CK-18). We wilden samenvattende en precieze schattingen berekenen van de accuratesse van de twee antigenen van CK-18 (M30 en M65) bij het detecteren van niet-alcoholische steatohepatitis (NASH) en fibrose. We voerden meta-analyses uit met twee verschillende statistische methoden: een lineair mixed-effects multiple thresholds model of een bivariaat logit-normal random-effects model. Het mixed effects multiple thresholds model werd gekozen om rekening te houden met de heterogene rapportage van drempelwaarden, wat vaak voorkomt bij studies van continu te meten biomarkers. Binnen de 41 geselecteerde primaire studies zagen we een matige accuratesse. Er kan waarschijnlijk wel een hoge negatief voorspellende waarde worden bereikt bij een lage prevalentie, zoals in eerstelijnszorgcenta. Uit het gerapporteerde onderzoek concludeerden we verder dat de prestaties van CK-18 kunnen worden verbeterd in combinatie met andere markers; multimarkerscores met inbegrip van CK-18 presteerden doorgaans beter.

Ook verschillende vormen van beeldvorming kunnen worden gebruikt om het stadium van leverfibrose vast te stellen. **Hoofdstuk 3** beschrijft een analyse op basis van individuele patiëntgegevens (IPD) van het meten van de stijfheid van de lever met “vibration controlled transient elastography” (LSM-VCTE), in vergelijking met twee multimarkerscores: de Fibrosis-4 index (FIB-4) en de NAFLD Fibrosis Score (NFS). Deze tests kunnen alle worden ingezet om de inzet van het leverbiopt, nog steeds de referentiestandaard voor stadiëring van fibrose, te verminderen. De niet-invasieve tests presteerden suboptimaal, daarom werden de ze ook als onderdeel van een sequentiële teststrategie beoordeeld. Enkel voor LSM-VCTE was de accuratesse redelijk, terwijl sequentiële gebruik van markers leidde tot een hogere sensitiviteit en specificiteit; met een lage drempel kon gevorderde fibrose worden uitgesloten en met hoge drempel cirrose aangetoond. De sequentiële strategie verminderde ook het aantal patiënten met een resultaat tussen een hoge en een lage drempel. Een subgroepanalyse toonde verder aan dat de prestaties varieerden tussen subgroepen die waren ingedeeld op basis van body mass index (BMI) of leeftijd. Deze IPD meta-analyse kan dienen als referentiepunt bij het ontwikkelen van alternatieve teststrategieën met multimarkerscores voor de stadiëring van fibrose. Verdere validatie van de teststrategieën blijft nodig.

Het vermogen van een test om het beloop te voorspellen kan de patiëntenzorg voor patiënten ondersteunen, bij voorbeeld door het toepassen van risicostratificatie. FIB-4, NFS en APRI zijn multimarkerscores die vaak worden gebruikt voor het evalueren van de mate van fibrose bij patiënten met NAFLD. **Hoofdstuk 4** richt zich op een synthese van de beschikbare gegevens over de accuratesse van deze multimarkers bij het voorspellen van NAFLD-gerelateerde gebeurtenissen. Ons systematisch literatuuronderzoek toonde aan dat deze multimarkers goed presteerden bij het voorspellen van levergerelateerde gebeurtenissen. De markers presteerden echter zeer inconsistent bij het voorspellen van veranderingen in de mate van fibrose, wat gedeeltelijk kan worden verklaard door verschillen in tijdshorizon, definities en omvang en ernst van de ziekte bij aanvang van de studie, naast verschillen tussen studiegroepen en centra. De mate van leverfibrose staat bekend als de sterkste voorspeller van het beloop bij patiënten met NAFLD. Daarom kunnen toekomstige studies zich richten op de evaluatie van de prestaties van niet-invasieve tests ten opzichte van het leverbiopt. Vanwege de enorme heterogeniteit tussen de studies konden we geen samenvattende schattingen berekenen. De tekortkomingen in de studies die we zagen kunnen toekomstig onderzoek worden vermeden.

Een van de uitdagingen in het onderzoek in Hoofdstuk 4 was het ontbreken van een geschikt instrument om het risico op vertekening en beperkingen in de toepasbaarheid van de studies te beoordelen. Dit vormde de inspiratie voor het werk in **Hoofdstuk 5**, waar we een bestaand instrument, QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2), gebruikten om QUAPAS (Quality Assessment of Prognostic Accuracy Studies) te ontwikkelen, een aanpassing van QUADAS-2 voor studies naar prognostische accuratesse. Dergelijke longitudinale acuratessestudies verschillen qua opzet van crosssectionele studies naar de diagnostische accuratesse, met andere bronnen van mogelijke vertekening. QUAPAS werd ontwikkeld met behulp van het raamwerk van QUADAS-2, aangevuld met vragen die op vertekening kunnen wijzen uit andere tools: QUIPS (Quality in Prognosis Studies) en PROBAST (Prediction Model Risk of Bias Assessment Tool).

Het gebruik van QUAPAS verloopt identiek aan dat van QUADAS-2. Het risico op vertekening wordt beoordeeld voor vijf domeinen: de deelnemers, de indextest, de uitkomst, flow en timing, en de analyse. Signaleringsvragen moeten helpen tot een oordeel over vertekening te komen, voor elk van deze vijf domeinen. Zorgen over de toepasbaarheid worden beoordeeld binnen de eerste vier domeinen. QUAPAS stelde ons, beter dan QUADAS-2, in staat om studies te identificeren met een gevaar voor vertekende resultaten. De geloofwaardigheid van dit soort instrumenten staat of valt met een correcte rapportage van de studie. We moesten vaststellen dat de gebrekkige verslaglegging van de primaire studies ons oordeel over vertekening sterk belemmerde, zeker voor het analysedomein. Rapportage is dus voor verbetering vatbaar. Toekomstige meta-epidemiologische studies kunnen onze kennis over vertekening in dit type studies verder versterken, wat tot een verdere verbetering en verfijning van QUAPAS zou kunnen leiden.

Waar de eerste hoofdstukken zich vooral richtten op een synthese van bestaande kennis, biedt **Hoofdstuk 6** een andere aanpak. Daarin brengen we verslag uit van een vergelijkende studie naar de diagnostische accuratesse van zeventien biomarkers, multimarkerscores en “vibration controlled transient elastography” (VCTE) voor de detectie van risicovolle NASH en gevorderde fibrose. We maakten daarbij gebruik van gegevens verzameld binnen het LITMUS (Liver Investigation: Testing Marker Utility in Steatohepatitis) Metacohort, een combinatie van internationale cohortstudies in patiënten met door een biopt bevestigde NAFLD. We gingen ook na hoe goed deze markers en multimarkers waren in het mogelijk preselecteren van patiënten die in aanmerking komen voor geneesmiddelenonderzoek, waarbij voor deelname een combinatie van actieve NASH en relevante fibrose (”at-risk NASH”) vereist is. De prestatie van een marker werd uitgedrukt als het oppervlak onder de ROC-curve (AUC), waarbij biopsie als referentiestandaard werd gebruikt, en vergeleken met die van de Fibrosis-4-index voor leverfibrose (FIB-4) in dezelfde subgroep. Van 966 patiënten konden resultaten worden gebruikt. Geen enkele biomarker of multi-marker overschreed op statistisch significante wijze de vooraf gedefinieerde AUC drempel (0,80). Meestal waren de prestaties vergelijkbaar zijn met die van FIB-4. SomaSignal, de ADAPT-score en de meting van de leverstijfheid waren wel in staat gevorderde fibrose te detecteren met een aanvaardbare accuratesse. We zagen verder dat de screening op “at-risk NASH” van potentiële deelnemers aan geneesmiddelenonderzoek kon worden verbeterd als alleen zij die een resultaat boven een bepaalde drempel halen een biopsie ondergaan. In de prospectieve LITMUS cohortstudie zullen de prestaties van deze markers en multimarkers verder worden onderzocht.

Om de detectie van NASH en “at-risk NASH” verder te verbeteren deden we ook een beroep op een vorm van “machine learning”: gradient boosting machine (GBM) genaamd (**Hoofdstuk 7)**. Met data uit het LITMUS Metacohort (75% in een trainingsset, 25% in een validatieset) over 35 voorspellers, zowel klinische gegevens als biomarkers, bouwden we modellen voor het detecteren van NASH, “at-risk NASH” en fibrosestadia. Ontbrekende gegevens werden ingevuld door meervoudige imputatie. Voor elke doelconditie werden twee modellen getraind: een klinisch model en een uitgebreid model, met zowel klinische gegevens als biomarkers. Voor NASH en “at-risk NASH” bouwden we telkens twee modellen: een direct model en een samengesteld model. We ontdekten dat de prestaties van de klinische en van de uitgebreide modellen redelijk vergelijkbaar waren. De toevoeging van biomarkers leidde dus niet to een betere detectie van NASH, hoewel er voor fibrose kleine verbeteringen werden waargenomen. De samengestelde aanpak, met afzonderlijke GBM-modellen voor elke component van NASH/at-risk NASH, verbeterde de detectie aanzienlijk in vergelijking met de directe aanpak. De samengestelde GBM-modellen presteerden wel beduidend beter dan de enkelvoudige markers en de bestaande multimarkerscores. Deze nieuwe modellen zullen worden gevalideerd in het lopende LITMUS-cohort.

In **Hoofdstuk 8** bieden we een illustratieve toepassing van een Bayesiaanse, nonparametrische methode voor het evalueren van de prestaties van biomarkers waarbij rekening wordt gehouden met informatie over covariaten. De diagnostische accuratesse van een test is immers meestal geen vaststaande eigenschap; prestaties kunnen variëren op basis van studie- of patiëntkenmerken. Er werden drie scenario’s geschetst, op basis van een vergelijking van de verdelingen van testresultaten tussen subgroepen. Elk scenario leidt tot andere conclusies over de validiteit van een niet-gecorrigeerde, standaard ROC-curve. Dit zijn, achtereenvolgens, scenario (1) identieke distributie, identieke prestaties, scenario (2) verschillende distributie, identieke prestaties, en scenario (3) verschillende distributie, verschillende prestaties. Door dit soort verschillen is het goed mogelijk dat de conventionele ROC-curve niet goed genoeg de feitelijke prestaties van de test weergeeft. We gebruikten vervolgens individuele patiëntgegevens (IPD) uit een reeks evaluaties van D-dimeer-tests voor het uitsluiten van een longembolie. Covariaat-specifieke en covariaat-gecorrigeerde ROC-curves werden berekend, dit om de prestaties van D-dimeer te onderzoeken rekening houdend met covariaten zoals leeftijd of geslacht. We konden de genoemde scenario’s ook voor D-dimeer waarnemen. Daarnaast waren binnen de respectievelijke subgroepen ook andere positiviteitsdrempels nodig om een vooraf gedefinieerde sensitiviteit te bereiken. Door in plaats van de conventionele ROC curve ook conditionele ROC curves te berekenen kunnen we de accuratesse van een test beter begrijpen, en ook meer geldige afkapwaarden selecteren.

Tot slot bespreken we in **hoofdstuk 9** in meer algemene zin de relevantie van de bevindingen uit ons onderzoek en wijzen we op aanknopingspunten voor toekomstig onderzoek.

Publications

Included in thesis

1. **Lee J**, Mulder F, Leeflang M, Wolff R, Whiting P, Bossuyt PM. QUAPAS: An Adaptation of the QUADAS-2 Tool to Assess Prognostic Accuracy Studies. Annals of Internal Medicine. 2022;175(7):1010-8.
2. **Lee J**, Vali Y, Boursier J, Duffin K, Verheij J, Brosnan MJ, et al. Accuracy of cytokeratin 18 (M30 and M65) in detecting non-alcoholic steatohepatitis and fibrosis: A systematic review and meta-analysis. PLoS One. 2020;15(9):e0238717.
3. **Lee J**, Vali Y, Boursier J, Spijker R, Anstee QM, Bossuyt PM, et al. Prognostic accuracy of FIB-4, NAFLD fibrosis score and APRI for NAFLD-related events: a systematic review. Liver International. 2021;41(2):261-70.
4. **Lee J**, Westphal M, Vali Y, Boursier J, Ostroff R, Alexander L, et al. Machine learning algorithm improves detection of NASH (NAS-based) and at-risk NASH, a development and validation study. Hepatology. 2023.
5. Mózes FE, **Lee JA**, Selvaraj EA, Jayaswal ANA, Trauner M, Boursier J, et al. Diagnostic accuracy of non-invasive tests for advanced fibrosis in patients with NAFLD: an individual patient data meta-analysis. Gut. 2022;71(5):1006-19.
6. **Lee J**, Vali Y, Boursier J, Petta S, Wonders K, Tiniakos D, et al. Biomarkers for staging fibrosis and non-alcoholic steatohepatitis in non-alcoholic fatty liver disease (the LITMUS project): a comparative diagnostic accuracy study. The Lancet Gastroenterology & Hepatology. 2023.
7. **Lee J**, van Es N, Takada T, Klok FA, Geersing G-J, Blume J, et al. Covariate-specific ROC curve analysis can accommodate differences between covariate subgroups in the evaluation of diagnostic accuracy. Journal of Clinical Epidemiology. 2023.

Outside of thesis

8. Mak AL, **Lee J**, van Dijk A-M, Vali Y, Aithal GP, Schattenberg JM, et al. Systematic review with meta-analysis: diagnostic accuracy of pro-C3 for hepatic fibrosis in patients with non-alcoholic fatty liver disease. Biomedicines. 2021;9(12):1920.
9. Selvaraj EA, Mózes FE, Jayaswal ANA, Zafarmand MH, Vali Y, **Lee JA**, et al. Diagnostic accuracy of elastography and magnetic resonance imaging in patients with NAFLD: a systematic review and meta-analysis. Journal of hepatology. 2021;75(4):770-85.
10. Vali Y, **Lee J**, Boursier J, Spijker R, Löffler J, Verheij J, et al. Enhanced liver fibrosis test for the non-invasive diagnosis of fibrosis in patients with NAFLD: a systematic review and meta-analysis. Journal of hepatology. 2020;73(2):252-62.
11. Vali Y, **Lee J**, Boursier J, Spijker R, Verheij J, Brosnan MJ, et al. FibroTest for evaluating fibrosis in non-alcoholic fatty liver disease patients: a systematic review and meta-analysis. Journal of Clinical Medicine. 2021;10(11):2415.
12. Van Dijk A-M, Vali Y, Mak AL, **Lee J**, Tushuizen ME, Zafarmand MH, et al. Systematic review with meta-analyses: diagnostic accuracy of FibroMeter tests in

patients with non-alcoholic fatty liver disease. Journal of Clinical Medicine. 2021;10(13):2910.

13. Mózes FE, **Lee JA**, Vali Y, Alzoubi O, Staufer K, Trauner M, et al. Performance of non-invasive tests and histology for the prediction of clinical outcomes in patients with non-alcoholic fatty liver disease: an individual participant data meta-analysis. The Lancet Gastroenterology & Hepatology. 2023.

Portfolio

Courses	Year
• AMC World of Science, AMC Graduate School	June 2019
• Advanced Analysis of Prognosis Studies, AMC Graduate School	March 2019
• Research Data Management, AMC Graduate School	May 2019
• Markers and Prediction Research, Erasmus Summer Program (NIHES)	August 2019
• Fundamentals of Medical Decision Making, Erasmus Summer Program (NIHES)	August 2019
• Causal Inference, Erasmus Summer Program (NIHES)	August 2020
• Advanced Machine learning, Statistical Horizons	October 2021
(Inter)national conferences	
With presentations	
• Using Real World Data and Designs to Optimize Decisions, the 35th International Conference on Pharmacoepidemiology and Therapeautic Risk Management, the International Society for Pharmacoepidemiology (ICPE), Philadelphia, PA, USA	August 2019
• Embracing Diversity, the 26th Cochrane Colloquium, The Cochrane Collaboration, Santiago, Chile	October 2019
• The Digital International Liver Congress 2021, The European Association for the Study of Liver Disease (EASL), virtual conference	June 2021
• The Liver Meeting 2021, The American Association for the Study of Liver Disease (AASLD), virtual conference	November 2021
• The International Liver Congress 2022, The European Association for the Study of Liver Disease (EASL), London, UK	June 2022
• Peer review and scientific publication congress, Chicago, USA	September 2022

Without presentations	
• Amsterdam Public Health Annual Meeting, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands	Nov 2018
• The International Liver Congress 2019, The European Association for the Study of Liver Disease (EASL), Vienna, Austria	April 2019
• Precision Medicine, Amsterdam Public Health Spring Meeting, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands	July 2019
• The Digital International Liver Congress 2020, The European Association for the Study of Liver Disease (EASL), virtual conference	September 2020
• MEMTAB 2020 symposium, virtual conference	December 2020
• WEON 2021, Science in an Online Society, virtual conference	June 2021
Teaching	
• Applied Epidemiology: Evaluation of Medical Tests European Educational Program in Epidemiology, (EEPE), Florence, Italy	July 2022
Seminars	
• Weekly BiTE seminars	2018-2022
• Bi-monthly KEBB department seminars	2018-2022

Letter of thanks

This thesis would not have been possible without the contribution and support from the community I had the privilege of surrounding myself with over the last several years. This letter is to you all.

First and foremost, I would like to express my highest gratitude to my supervisors, Hadi Zafarmand and Patrick Bossuyt. Hadi, thank you for investing your time into shaping the first year of my PhD research. I learned a lot from our 1-on-1s, many of which are skills that have made me a more meticulous and conscientious researcher. Patrick, I truly believe I won the lottery with your mentorship over the last 4.5 years. You have inspired, encouraged, and allowed me to flourish as an independent thinker. We spent much time discussing not only research methods and how to solve data puzzles, but also common interests in various forms of art and literature, and more personal matters. For all these conversations I am truly grateful. I hope to carry forward the lessons I have learned from you over the years, my favorite being to lead with humility.

I would also like to extend my thanks to the many LITMUS investigators, without whom the work within this thesis would truly be impossible. Dear Quentin Anstee, thank you for the many great opportunities to collaborate and your generous guidance on NAFLD/NASH. It was a pleasure to work under your leadership and alongside the many great LITMUS investigators.

Dear LITMUS WP2 colleagues, thank you for your unwavering support and participation in the output of our work package. Your expertise and guidance were indispensable resources and I am grateful for your trust and support over the years. I truly believe we were the most fun work package, as I carry with me many good memories from our in-person meetings from the General Assembly to many international conferences.

Dear Michael Pavlides and Ferenc Moses, it was a great pleasure to work with you so closely on both the evidence synthesis and WP5 imaging studies. Your expertise and ambitions for each project were very inspiring. Thank you for the great collaboration over the years.

Thank you to the QUAPAS steering committee members, Mariska Leeflang, Robert Wolf, Penny Whiting, Jill Hayden and Patrick Bossuyt. Developing QUAPAS would not have been possible without the collective expertise from each of you. Thank you for the guidance, for asking the difficult questions, and your trust in me. Dear Frits Mulder, thank you for the many hours that you dedicated to this project on top of your busy schedule, I thoroughly enjoyed our many conversations.

At the Epidemiology and Data Sciences department, I would like to thanks the BiTE group and my fellow PhD colleagues. Dear Mariska Leeflang, Miranda Langedaam, Yasaman Vali, Bada Yang, Mona Ghannad, Maria Olsen, Amber Boots, Marileen Wiegersma, Linda Pluyman, and Mariska Tuut, thank you for your comradery and many corridor conversations over the years. Dear Yasaman, we experienced many elements of the LITMUS project and office life together, and I am proud of all our combined achievements. I will remember our tea time ritual and many trips together with a smile. Dear Bada, I am thankful for the friendship that came from humble beginnings in the office. Thank you for always setting the bar so high and for accepting the role as my paranymph. Dear Koos Zwinderman, thank you for your support and interest in tackling the many statistical questions and challenges I faced. You never failed to provide a solution and I am thankful for your guidance.

Dear Onno Holleboom, Anne-Linde, Anne-Marieke, thank you for your collaboration and support in the evidence synthesis projects and much more in LITMUS. It was a great pleasure to work with you and your team.

To my family, dear mom, you have done your absolute most to encourage, love, and prepare me for a bigger and brighter future. Thank you for giving me the freedom to chase my dreams, wherever they take me. Dear James, my day one number one. Thank you for always believing and seeing the best in me, even when I can’t seem to. There are really no words to express how grateful I am to have you as my brother.

To my Dutch family, dear Ger, Bea, Lesley, Martijn, Philip and Boris, thank you for being my second family when mine felt so far away. You welcomed me with arms wide open and gave me another place to call home. For that I will always be thankful.

I would also like to thank my friends, new and old, near and far, who kept me sane. I am in awe of your ability to navigate my mind to the many beautiful aspects of life outside of work and research. Thanks to all the inspiring and open-minded friends for the long conversations, the laughs, the encouragements, for the shared meals, the long walks, and the many late nights and early mornings. You kept me afloat.

And lastly, I would like to thank my partner. Dear Steven, you have been with me for so many of the highs and lows. Thank you for all your patience, support and unconditional love through it all. It only goes up from here.

About the author

Jenny Lee was born on July 8th, 1992 in Tokyo, Japan. Shortly after, she moved with her family to the America, living in Washington State and California as well as Seoul, South Korea in her early years. She received her high school degree in 2010 and began her Bachelor of Science as a pre-medical student at University of Washington in Seattle, receiving a degree in Biology in 2014. During this period, she realized her interest in public health research and spent six months volunteering for HIV/AIDS prevention in KwaZulu Natal, South Africa. She took a gap year and travelled solo to over 10 countries in Southeast Asia and Oceania before settling in Munich, Germany in 2016 to pursue a Master of Science in Epidemiology and Biostatistics at the Ludwig Maximilian University, with a specialization in clinical and pharmaco-epidemiology. Jenny worked as an epidemiologist with the Global Epidemiology department at Bayer Pharmaceuticals AG in Berlin, Germany, where she supported stages of drug discovery and clinical development for Cardiovascular and Woman’s Health therapeutic areas. After completing her Master degree in 2018, she accepted a PhD position with the Department of Epidemiology and Data Science at the Amsterdam University Medical Center, Amsterdam, the Netherlands. Her PhD research focused on the discovery and validation of biomarkers in detecting conditions on the spectrum of non-alcoholic fatty liver disease (NAFLD) and their ability to expedite patient recruitment for therapeutic trials (Phase III). Jenny also focused on methodologic topics such as assessment of bias, addressing challenges in the test accuracy space, and utilizing machine learning techniques for prediction models. Following her PhD, she works as epidemiologist for a US based start-up, focusing on providing large scale data-driven solutions to support stages of drug development.