

RESEARCH ARTICLE

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

Jenny Lee^{1*}, Yasaman Vali¹, Jérôme Boursier^{2,3}, Kevin Duffin⁴, Joanne Verheij⁵, M. Julia Brosnan⁶, Koos Zwinderman¹, Quentin M. Anstee^{7,8}, Patrick M. Bossuyt¹, Mohammad Hadi Zafarmand¹

1 Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands, **2** Hepato-Gastroenterology Department, Angers University Hospital, Angers, France, **3** HIFIH Laboratory, UPRES EA3859, Angers University, Angers, France, **4** Lilly Research Laboratories, Eli Lilly and Company Ltd (LLY), Indianapolis, IN, United States of America, **5** Department of Pathology, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands, **6** Internal Medicine Research Unit, Pfizer Inc, Cambridge, MA, United States of America, **7** The Newcastle Liver Research Group, Translational & Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom, **8** Newcastle NIHR Biomedical Research Centre, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom

* j.a.lee@amsterdamumc.nl



OPEN ACCESS

Citation: Lee J, Vali Y, Boursier J, Duffin K, Verheij J, Brosnan MJ, et al. (2020) Accuracy of cytokeratin 18 (M30 and M65) in detecting non-alcoholic steatohepatitis and fibrosis: A systematic review and meta-analysis. PLoS ONE 15(9): e0238717. <https://doi.org/10.1371/journal.pone.0238717>

Editor: Pavel Strnad, Medizinische Fakultät der RWTH Aachen, GERMANY

Received: June 30, 2020

Accepted: August 22, 2020

Published: September 11, 2020

Copyright: © 2020 Lee et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This systematic review has been conducted as part of the evidence synthesis efforts in the LITMUS (Liver Investigation: Testing Marker Utility in Steatohepatitis) study. The LITMUS project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 777377 to QMA. This Joint

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

Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. The funder provided support in the form of salaries for authors JL, YV and MHZ, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: QMA is coordinator of the IMI2 LITMUS consortium. He reports research grant funding from Abbvie, Allergan/Tobira, AstraZeneca, GlaxoSmithKline, Glympse Bio, Novartis Pharma AG, Pfizer Ltd., Vertex; consultancy on behalf of Newcastle University for Abbott Laboratories, Acuitas Medical, Allergan/Tobira, Blade, BNN Cardio, Cirus, CymaBay, EcoR1, E3Bio, Eli Lilly & Company Ltd., Galmed, Genfit SA, Gilead, Grunthal, HistoIndex, Indalo, Imperial Innovations, Intercept Pharma Europe Ltd., Inventiva, IQVIA, Janssen, Kenes, Madrigal, MedImmune, Metacrine, NewGene, NGMBio, North Sea Therapeutics, Novartis, Novo Nordisk A/S, Pfizer Ltd., Poxel, ProSciento, Raptor Pharma, Servier, Viking Therapeutics; and speaker fees from Abbott Laboratories, Allergan/Tobira, BMS, Clinical Care Options, Falk, Fishawack, Genfit SA, Gilead, Integritas Communications, MedScape. MJB is a share-holder and employee of Pfizer. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

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. 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 in [S1 File](#).

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 in [S1 File](#)). 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 $NAS \geq 4$ with at least one point for each criteria of steatohepatitis for the characterization of NASH. See S3 Table in [S1 File](#) 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 ($\geq F2$) and advanced ($\geq F3$) fibrosis were included. See S4 Table in [S1 File](#) for different scoring systems for liver fibrosis, and S5 Table in [S1 File](#) 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 summary 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).

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 (Fig 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 Figs in S1 File. 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.

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 Fig in S1 File). 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

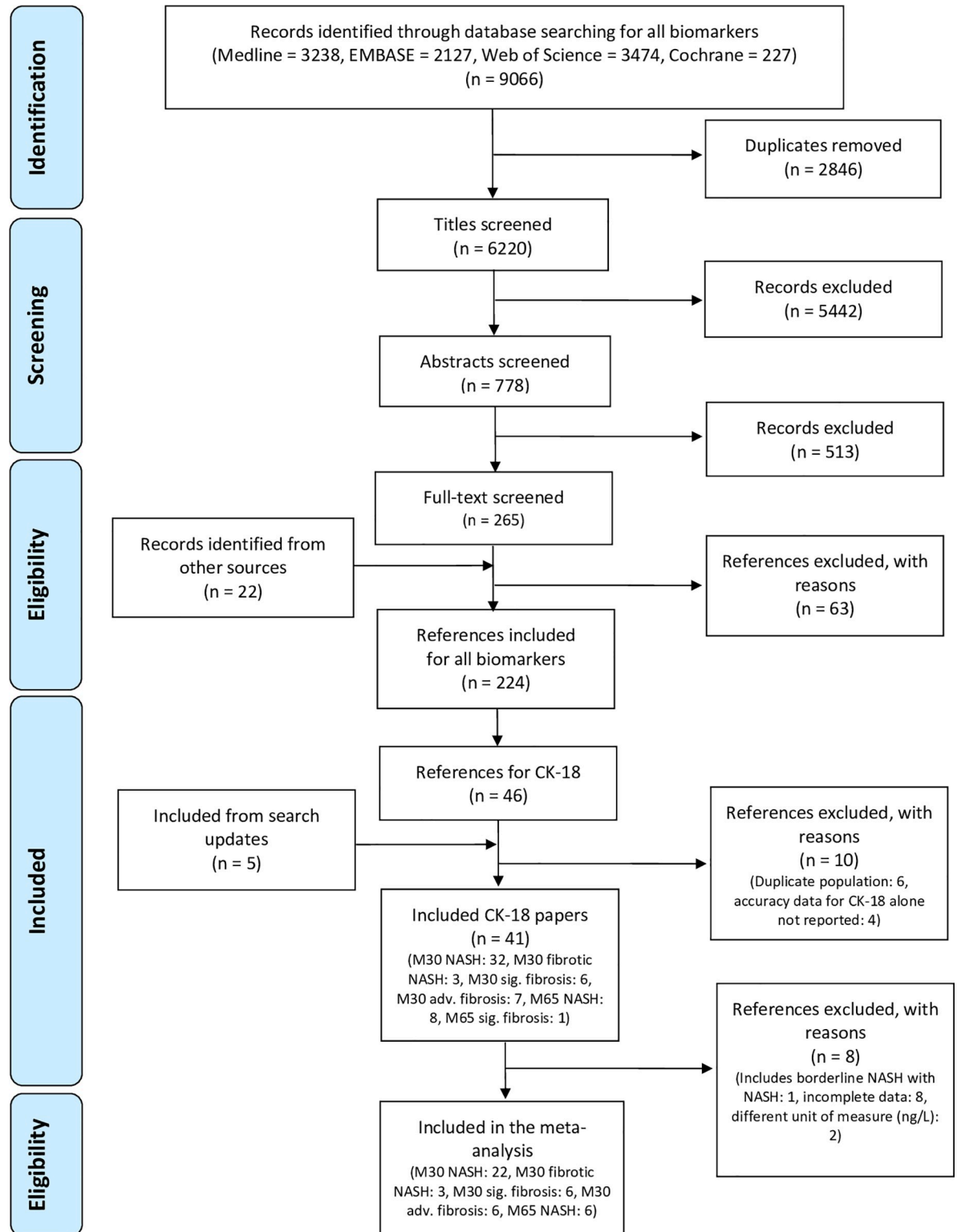


Fig 1. PRISMA flow diagram of included primary studies.

<https://doi.org/10.1371/journal.pone.0238717.g001>

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	39	Biopsy proven NAFLD	30.6 (5.4)	77.8 (56.1) [‡]	49.6 (30.5) [‡]	NR
			Adv. fibrosis	22					
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	18	Biopsy proven NAFLD	28.7 (4.4)	117 (87.9) [‡]	63.1 (33.3) [‡]	DM: 54 HTN: 65
			Adv. fibrosis	9					
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

(Continued)

Table 1. (Continued)

Study ID	Country	N (female)	Target condition	Prev (%)	Population	BMI, mean (SD)	ALT, median (IQR)	AST, median (IQR)	Comorbidities (%)
26. Rosso 2016 [43]	Italy	105 (29)	Sig. fibrosis	59	Biopsy proven NAFLD	28.1 (3.9)	65 (57–79)	36 (33–41)	NR
			Adv. fibrosis	36					
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	54 23	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
			NASH						
			Sig. fibrosis	51					
			Adv. Fibrosis	17					
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.

<https://doi.org/10.1371/journal.pone.0238717.t001>

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 (Fig 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

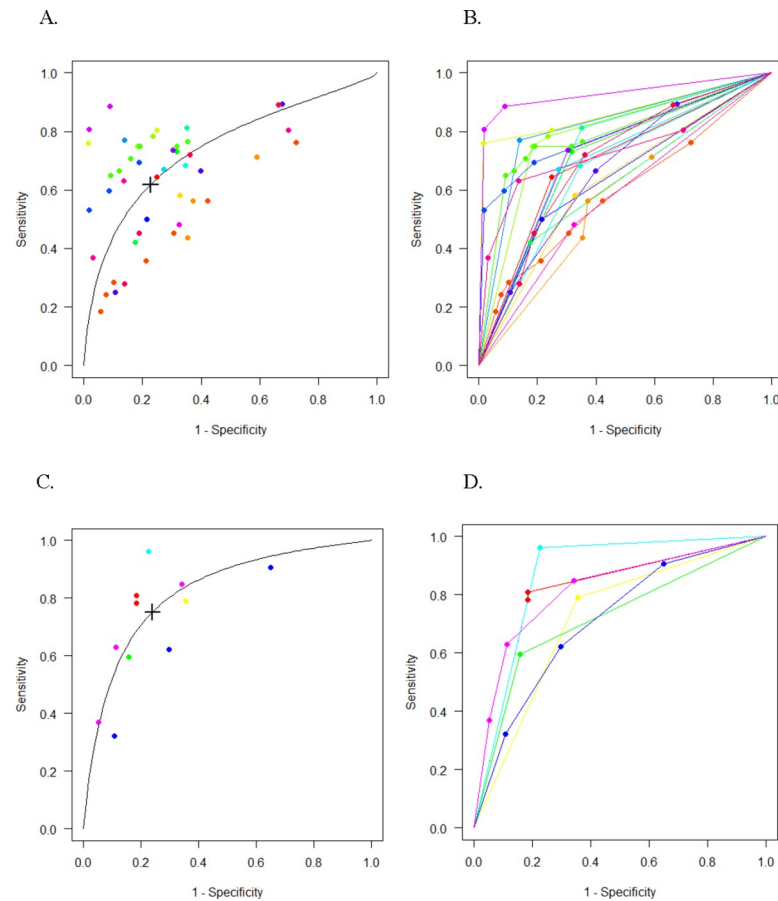


Fig 2. Multiple threshold SROC and ROC curves for detecting NASH. Multiple threshold SROC and ROC curves for CK-18 M30 (A-B) and M65 (C-D) in detecting NASH. Each point represents a reported threshold value, points of the same color represent thresholds reported within the same study. The x-axis indicates 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.

<https://doi.org/10.1371/journal.pone.0238717.g002>

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 Fig 3A and 3B.

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 Fig in S1 File). 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 (Fig 2C).

We again investigated the PPV and NPV in various clinical settings (Table 3, Fig 3C and 3D). Fixing sensitivity from 0.80 to 0.90, the specificity ranged from 0.70 to 0.51 at threshold

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

<https://doi.org/10.1371/journal.pone.0238717.t002>

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 Fig in S1 File). 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 studies [27, 43, 47, 55, 59] with a total of 1,155 participants (554 had significant fibrosis) (S6 Fig in S1 File). The resulting AUC was 0.68. See S7A Fig in S1 File 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 Fig in S1 File). We calculated an AUC of 0.75, with included threshold values ranging from 216 to 396 U/L (see S7B Fig in S1 File). One

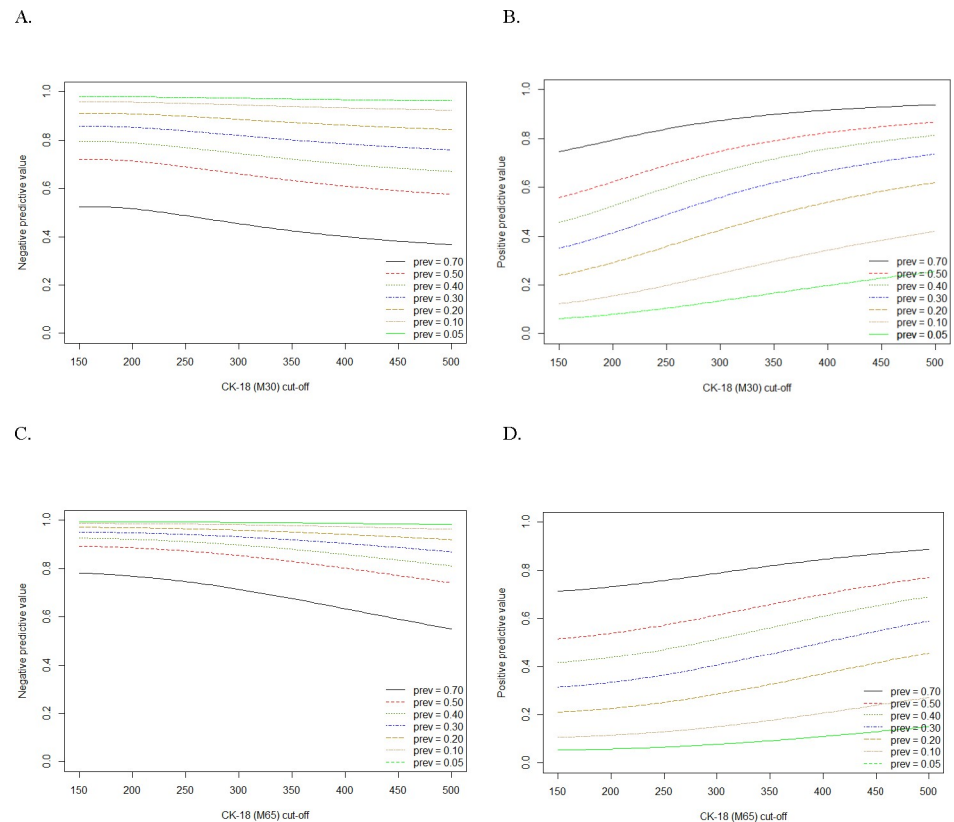


Fig 3. Positive and negative predictive values and thresholds for detecting NASH. Plots illustrating the negative and positive predictive values of M30 (A-B) and M65 (C-D) in detecting NASH at corresponding threshold values, projected by the multiple thresholds model. Each colored line represents a different 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.

<https://doi.org/10.1371/journal.pone.0238717.g003>

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].

Table 3. Performance of M65 in detecting NASH: Positive and negative predictive values for different settings of 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			0.92	0.45	37

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

<https://doi.org/10.1371/journal.pone.0238717.t003>

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

Author	Target condition and population	Scoring system	Ingredients	AUC
1. Anty 2010	NAFLD grading among morbidly obese	The Nice Model	Metabolic syndrome, ALT, CK-18	Training: 0.88 Validation: 0.83
2. Boursier 2018	Fibrotic NASH among NAFLD	MACK-3	HOMA, AST, CK-18	Validation: 0.85
3. Cao 2013	NASH among NAFLD		ALT, platelets, M30, and TG	0.92
4. Chuah 2019	Fibrotic NASH among NAFLD	MACK-3	HOMA, AST, CK-18	0.80
5. Ergelen 2015	Fibrosis among NAFLD		TE, M30	F ≥ 2: 0.89 F ≥ 3: 0.93
6. Grigorescu 2012	NASH among NAFLD		M65, IL-6 and adiponectin	0.90
7. Pirvulescu 2012	NASH (including borderline NASH) among morbidly obese patients	NASH-score	BMI, ALT, AST, ALP, HOMA-IR, M65, and adiponectin	0.96
8. Rosso 2016	Fibrosis among NAFLD		TE, M30	F ≥ 2: 0.84 F ≥ 3: 0.87
9. Tada 2018	NASH among NAFLD	FIC-22	FIB-4 and CK-18	0.82
10. Tamimi 2011	NASH among NAFLD		Soluble fas and CK-18	Training: 0.93 Validation 0.79
11. 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
12. Younossi 2008	NASH among NAFLD		M30 and M65, adiponectin, resistin	Training: 0.91 Validation: 0.73
13. 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: triglyceride, HOMA-IR: homeostatic model assessment for insulin resistance, TE: transient elastography, IL-6: interleukin 6, BMI: body mass 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.

<https://doi.org/10.1371/journal.pone.0238717.t004>

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) 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 (Tables 2 and 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 PEVIVA [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 in [S1 File](#)), 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 [15] (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 used 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.

Supporting information

S1 File.
(DOCX)

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.

Author Contributions

Conceptualization: Jenny Lee, Yasaman Vali, Jérôme Boursier, Kevin Duffin, M. Julia Brosnan, Quentin M. Anstee, Patrick M. Bossuyt, Mohammad Hadi Zafarmand.

Data curation: Jenny Lee, Jérôme Boursier.

Formal analysis: Jenny Lee, Koos Zwinderman, Patrick M. Bossuyt, Mohammad Hadi Zafarmand.

Funding acquisition: M. Julia Brosnan, Quentin M. Anstee.

Investigation: Jenny Lee, Yasaman Vali, Jérôme Boursier, Kevin Duffin, Joanne Verheij, Quentin M. Anstee, Patrick M. Bossuyt, Mohammad Hadi Zafarmand.

Methodology: Jenny Lee, Patrick M. Bossuyt, Mohammad Hadi Zafarmand.

Supervision: Patrick M. Bossuyt, Mohammad Hadi Zafarmand.

Writing – original draft: Jenny Lee.

Writing – review & editing: Jenny Lee, Yasaman Vali, Jérôme Boursier, Kevin Duffin, Joanne Verheij, M. Julia Brosnan, Koos Zwinderman, Quentin M. Anstee, Patrick M. Bossuyt, Mohammad Hadi Zafarmand.

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, Charatcharoenwithaya 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; 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, Randhawa 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 non-alcoholic 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.