Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Diagnostic Performance and Agreement of Deep Learning- Based Digital Pathology for Liver Fibrosis Assessment: A Systematic Review and Meta-Analysis

Qiuxia Qiu
Meizhou City People's Hospital
Qiuer Qiu
Meizhou City People's Hospital
Jiapeng Zhang
1025973954@qq.com

Jiaying University

Systematic Review

Keywords: liver fibrosis, deep learning, digital pathology, whole-slide imaging, diagnostic accuracy, histopathology, artificial intelligence

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10974067/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Deep-learning-based digital pathology is increasingly used to stage and quantify liver fibrosis, but the evidence spans heterogeneous diseases, stains, model architectures, and outcome definitions. Methods: We systematically searched PubMed/MEDLINE, Embase, Web of Science Core Collection, Scopus, IEEE Xplore, and CNKI without year or language restrictions. Human liver pathology studies were eligible when an explicit deep-learning or neural-network method was applied to fibrosis staging, classification, agreement, detection, segmentation, or quantitative assessment. Direct staging/classification/agreement evidence was designated A-layer evidence; segmentation and continuous quantitative evidence was designated B-layer evidence. Diagnostic accuracy was pooled when compatible exact 2x2 data were available. Risk of bias and applicability were assessed using QUADAS-3. Results: Of 825 records, 28 reports representing 25 study families were included (17 A-layer and 11 B-layer reports). Four independent study families contributed exact 2x2 data for advanced fibrosis (F > = 3 or an etiologically equivalent threshold). Pooled sensitivity was 0.87 (95% CI 0.58–0.97; I2 = 64.3%) and specificity was 0.86 (95% CI 0.74–0.93; I2 = 33.3%). The pooled LR + was 6.15 (95% CI 2.96–12.78), LR- was 0.15 (95% CI 0.04–0.60), and DOR was 53.46 (95% CI 7.08-403.72). Including Pan 2026 as reported proportions produced similar sensitivity and specificity. Agreement, AUC, and B-layer outcomes were too heterogeneous for a single pooled estimate. All nine QUADAS-3 estimates had high overall risk of bias and high overall applicability concern. Conclusions: Deep-learning digital pathology shows promising performance for advanced liver fibrosis assessment, but the evidence base remains small, heterogeneous, and methodologically fragile. Multicentre external validation, locked models, standardized reporting, and prospective clinical evaluation are required before routine deployment.

1. Introduction

Liver fibrosis is a shared consequence of many chronic liver diseases and is a major determinant of progression to cirrhosis, liver-related complications, and death. In metabolic dysfunction-associated steatotic liver disease (MASLD), contemporary clinical guidance emphasizes identifying advanced fibrosis because fibrosis stage is closely linked to liver-related risk and can alter surveillance and management pathways [1]. Histopathological assessment remains particularly important when tissue-level architecture, inflammatory activity, or disease-specific staging cannot be resolved adequately by non-invasive tests.

However, conventional biopsy-based staging has intrinsic limitations. Fibrosis is spatially heterogeneous, ordinal staging systems compress complex architecture into a small number of categories, and interpretation may vary between observers. In a paired-biopsy study of NAFLD, fibrosis stage differed by at least one stage in 41% of paired samples; among patients with bridging fibrosis in one specimen, 35% had only mild or no fibrosis in the paired specimen [2]. These limitations create a strong rationale for methods that improve reproducibility and extract more continuous and spatially explicit information from digitized tissue.

Digital pathology has enabled whole-slide imaging, computational measurement of collagen and architectural features, and increasingly the application of deep learning to histological data. Recent liver-fibrosis studies have used convolutional and residual networks, weakly supervised multiple-instance learning, graph neural networks, Transformers, generative adversarial networks, and segmentation models to reproduce fibrosis stages or derive quantitative fibrosis phenotypes [3]. The potential advantages include standardized scoring, scalable analysis, reproducible measurements, and localization of fibrotic patterns. At the same time, model performance can depend on disease spectrum, staining protocol, scanner, slide preparation, annotation strategy, and validation design, raising concerns about transportability across centres.

Previous syntheses have largely addressed either non-invasive deep-learning approaches to liver fibrosis [4] or artificial intelligence across digital pathology as a broad diagnostic field [5]. In the latter setting, pooled accuracy was high, but nearly all included studies had at least one high or unclear risk-of-bias or applicability concern [5]. A liver-specific synthesis that distinguishes direct fibrosis staging from segmentation and continuous quantification is therefore needed. This systematic review and meta-analysis evaluated the diagnostic performance and agreement of deep-learning-based digital pathology for liver fibrosis assessment and separately mapped complementary evidence on fibrosis detection, segmentation, and quantitative phenotyping.

2. Methods

2.1 Study design, reporting framework, and protocol status

This systematic review and meta-analysis evaluated the diagnostic performance, agreement, and complementary quantitative evidence of deep-learning-based digital pathology for liver fibrosis assessment. The review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 framework [6].

Because the eligible literature encompassed conventional fibrosis staging, binary diagnostic-accuracy analyses, model–pathologist agreement, collagen segmentation, fibrosis-pattern detection, and continuous fibrosis quantification, the evidence was organized into two analytical layers. The A-layer comprised studies directly evaluating fibrosis staging, classification, or agreement against a histopathological reference standard. The B-layer comprised deep-learning studies evaluating fibrosis detection, segmentation, collagen quantification, continuous fibrosis burden, spatial phenotyping, or related quantitative pathological outcomes that could not be directly combined with conventional diagnostic-accuracy estimates.

The review was retrospectively registered in PROSPERO (CRD420261500208).

2.2 Information sources and search strategy

Six bibliographic databases were searched: PubMed/MEDLINE, Embase, Web of Science Core Collection, Scopus, IEEE Xplore, and the China National Knowledge Infrastructure (CNKI). The formal searches were conducted on 5 June 2026 for PubMed, 6 June 2026 for Embase, 7 June 2026 for Web of

Science Core Collection, 9 June 2026 for Scopus, 7 June 2026 for IEEE Xplore, and 12 June 2026 for CNKI. No restrictions were imposed on publication year, language, or publication type.

The search strategy combined three core concepts: (1) liver fibrosis or cirrhosis; (2) digital pathology, histopathology, whole-slide imaging, liver biopsy microscopy, second harmonic generation, two-photon excitation fluorescence, or multiphoton microscopy; and (3) deep learning, artificial intelligence, machine learning, or neural-network methods. Deep-learning terms included, where supported by the database syntax, convolutional neural networks, ResNet, DenseNet, EfficientNet, U-Net, Transformers, multiple-instance learning, weakly supervised learning, self-supervised learning, and related architectures.

To preserve recall, performance-related terms such as accuracy, sensitivity, specificity, AUC, and validation were deliberately not required. Exclusionary NOT operators for CT, MRI, or ultrasound were also not used because otherwise eligible pathology studies could mention these modalities as comparators. Database-specific controlled vocabulary, field tags, and proximity operators were adapted to each platform. The complete search strings and search log are provided in Supplementary Appendix S1.

The formal searches yielded 825 records before deduplication: PubMed (n = 147), Embase (n = 429), Web of Science Core Collection (n = 87), Scopus (n = 143), IEEE Xplore (n = 3), and CNKI (n = 16).

2.3 Eligibility criteria

Original studies were eligible when they met all of the following criteria: (1) human liver tissue, liver biopsy specimens, histological sections, whole-slide images, or other digital microscopic pathology images were analyzed; (2) a deep-learning or neural-network method was directly applied to the fibrosis-related image-analysis task; and (3) the study evaluated liver fibrosis staging, grading, classification, detection, quantification, severity assessment, or agreement with pathologist-derived fibrosis scores.

Eligible deep-learning approaches included, but were not limited to, convolutional neural networks, ResNet-based architectures, U-Net-based models, Transformers, graph neural networks, multiple-instance learning, weakly supervised learning, and generative adversarial networks.

Studies were excluded when they used only non-pathological modalities such as CT, MRI, ultrasound, elastography, or FibroScan; relied exclusively on serum biomarkers, clinical variables, electronic health records, or omics data; evaluated non-fibrosis targets without a fibrosis-related endpoint; involved only non-human material; or were reviews, editorials, guidelines, or other non-original reports.

Terminology such as artificial intelligence or machine learning alone was not considered sufficient for eligibility. A report had to document an eligible deep-learning or neural-network component that was directly involved in the fibrosis-analysis task. Consequently, quantitative image-analysis platforms such as qFibrosis, qSepta, qNodule, or FibroNest were excluded when the report did not document an eligible DL/NN architecture for the fibrosis task.

Conference abstracts and posters were not excluded on publication type alone. Potentially relevant conference reports with sufficient methodological and outcome information were assessed against the same eligibility criteria as full articles.

2.4 Study selection

Records from the six databases were merged into a master bibliographic dataset. Duplicate records were identified using DOI, PMID, title, and other bibliographic fields. Potential DOI–title conflicts were reviewed rather than being merged automatically.

Title and abstract screening was independently performed by Qiuxia Qiu (Reviewer 1) and Qiuer Qiu (Reviewer 2) using prespecified high-sensitivity eligibility criteria. Each record was categorized as include, exclude, or uncertain. Potentially relevant records were retained when the available information was insufficient for confident exclusion. Disagreements between the two reviewers were resolved by adjudication by Jiapeng Zhang.

Full-text and otherwise evaluable reports were subsequently assessed against the predefined eligibility criteria. Exclusion reasons were recorded using prespecified codes, including wrong population/non-human material, wrong modality/non-pathology, absence of an eligible DL/NN method, wrong fibrosis target or outcome, insufficient or unclear reporting, and duplicate/preprint/companion reporting without unique eligible evidence. Final inclusion and study-family assignments were confirmed during adjudication.

2.5 Study-family identification and management of overlapping reports

Multiple publications or conference reports originating from the same underlying patient cohort, clinical trial, or digital-pathology dataset were grouped into a common study family. Reports were considered related when they shared a clearly identifiable cohort, trial registration, dataset size and source, recruitment setting, or other distinctive sample characteristics.

Companion abstracts, preprints, theses, and duplicate conference versions were not treated as independent studies when a more complete representative report was available. When multiple eligible reports within the same study family provided distinct outcomes, they could contribute separately to descriptive evidence synthesis but were not treated as statistically independent observations within the same pooled meta-analysis. For example, Pan 2026 and Heinemann 2022 used the same 467-biopsy dataset and were treated as one underlying study family for independence considerations.

2.6 Data extraction and evidence classification

Data were extracted using a standardized framework. Extracted variables included report and study-family identifiers, publication year, liver-disease population, sample size, histological stain or imaging modality, digital-pathology input, deep-learning architecture, fibrosis target, staging system, reference standard, validation design, and reported performance measures.

Diagnostic-performance data included sensitivity, specificity, area under the receiver operating characteristic curve (AUC), accuracy, positive and negative likelihood ratios, and diagnostic odds ratios when available. Model–reference agreement measures included Cohen κ, weighted κ, quadratic weighted κ, correlation coefficients, and other reported agreement statistics.

When sufficient information was available, raw true-positive (TP), false-positive (FP), false-negative (FN), and true-negative (TN) counts were extracted directly or reconstructed only when the reported data uniquely determined all four cells. Rounded sensitivity and specificity values were not converted into integer 2×2 counts when multiple integer solutions were possible or when the reported proportions were mathematically incompatible with the known denominators.

B-layer outcomes, including Dice coefficients, mean average precision, segmentation accuracy, continuous fibrosis measures, correlations, and fibrosis-pattern metrics, were extracted separately and were not pooled with A-layer diagnostic-accuracy outcomes.

2.7 Definition of quantitative synthesis groups

The primary quantitative synthesis evaluated advanced fibrosis, defined as F ≥ 3 versus F0–F2 or the closest etiologically equivalent threshold. Four independent study families provided compatible exact 2×2 tables and were included in this primary analysis: Preechathammawong 2024, Yang 2026, Panzeri 2025, and Naik 2023.

A sensitivity analysis additionally incorporated Pan 2026, which reported sensitivity and specificity for F3–F4 versus F0–F2 but did not provide reconstructable exact 2×2 counts.

A secondary exploratory analysis evaluated significant fibrosis (F ≥ 2) using the three study families with compatible exact 2×2 data: Shabanian 2025, Preechathammawong 2024, and Panzeri 2025. Other fibrosis-stage thresholds, including cirrhosis (F4/S4), multiclass staging, and AUC-only estimates, were synthesized descriptively.

2.8 Risk of bias and applicability assessment

Risk of bias and applicability were assessed using QUADAS-3 version 1.2 at the accuracy-estimate level [7]. QUADAS-3 was selected rather than QUADAS-2 because several reports contributed multiple thresholds or distinct diagnostic-accuracy estimates. Its estimate-level framework and explicit separation of the Participants, Index test, Target condition, and Analysis domains matched the unit of inference used in this review and allowed analytical bias to be considered separately from participant- and index-test-related concerns. Three synthesis questions were prespecified: Q1, detection of advanced fibrosis (F ≥ 3 or an etiologically equivalent threshold); Q2, detection of significant fibrosis (F ≥ 2); and Q3, other direct fibrosis-staging accuracy estimates, including cirrhosis (F4/S4), retained for narrative synthesis but not compatible with the primary pooled analyses.

Risk of bias was assessed across the Participants, Index test, Target condition, and Analysis domains. Applicability was assessed for Participants, Index test, and Target condition, followed by an overall applicability judgment. Nine accuracy estimates were selected for QUADAS-3 assessment. The Klimov 2019 cirrhosis classifier was included because it provided a direct staging-accuracy estimate. The PATHAI PSC and Frankfurt NASH conference reports were not selected for estimate-level QUADAS-3 assessment because the available reports did not provide a direct fibrosis-stage accuracy estimate suitable for the prespecified diagnostic-accuracy synthesis.

2.9 Statistical analysis

The primary meta-analysis was based on exact 2×2 diagnostic tables for advanced fibrosis. Study-specific sensitivity and specificity were calculated from the original uncorrected cell counts. Exact binomial 95% confidence intervals were used for the study-level sensitivity and specificity displayed in forest plots.

Sensitivity and specificity were pooled separately on the logit scale using random-effects models. Between-study variance was estimated using restricted maximum likelihood (REML), and confidence intervals for pooled effects used the Knapp–Hartung adjustment.

Positive likelihood ratios (LR+), negative likelihood ratios (LR-), and diagnostic odds ratios (DORs) were analyzed on the logarithmic scale. For transformed pooling, when any cell in a study-level 2x2 table was zero, 0.5 was added to all four cells of that study to permit finite logit/log-scale transformations. Study-level sensitivity, specificity, and exact binomial 95% confidence intervals shown in forest plots were calculated from the original uncorrected counts.

Statistical heterogeneity was summarized using the I^2 statistic, interpreted cautiously because of the small number of studies. For Pan 2026, the reported sensitivity and specificity were included only in a sensitivity analysis. Variances were approximated from the reported proportions and the known numbers of advanced-fibrosis and non-advanced-fibrosis cases; exact TP/FP/FN/TN values were not reconstructed.

A bivariate/HSROC model was explored for the F ≥ 3 dataset. Because only four independent study families were available and the fitted random-effects correlation approached the parameter boundary, the model was considered unstable and was not retained for primary inference.

AUCs were not formally pooled because fibrosis thresholds, validation designs, internal versus external validation, and measures of dispersion differed substantially. Cross-validation standard deviations were not treated as sampling standard errors. Agreement measures were synthesized descriptively rather than formally pooled because studies reported non-equivalent statistics, including binary Cohen κ, weighted κ, quadratic weighted κ, and correlation coefficients, and variance estimates were frequently unavailable.

No formal assessment of small-study effects or publication bias was undertaken because the number of studies in each quantitative synthesis was too small for reliable funnel-plot or regression-based evaluation. Quantitative analyses were performed in R, primarily using the metafor package; exploratory bivariate diagnostic analyses used mada.

3. Results

3.1 Study selection

The database search identified 825 records: 147 from PubMed, 429 from Embase, 87 from Web of Science Core Collection, 143 from Scopus, 3 from IEEE Xplore, and 16 from CNKI. After removal of 291 duplicate records, 534 unique records underwent title and abstract screening, of which 435 were excluded. Ninety-nine reports were sought for retrieval. One report could not be located despite full-text searching, leaving 98 reports for eligibility assessment.

Seventy reports were excluded after eligibility assessment. The principal reason was the absence of an eligible deep-learning or neural-network method for the fibrosis task (n = 45). Other reasons included an ineligible fibrosis target or outcome (n = 9), duplicate, preprint, or companion reports without unique eligible data (n = 10), non-human or otherwise ineligible populations (n = 3), non-pathology modalities (n = 1), and insufficient or unclear reporting (n = 2). Ultimately, 28 reports representing 25 independent study families were included in the systematic review (Fig. 1). Seventeen reports contributed direct fibrosis staging, classification, or agreement evidence (A-layer), whereas 11 reports addressed fibrosis detection, segmentation, continuous quantification, or related digital-pathology phenotyping (B-layer).

3.2 Characteristics of the included evidence

The included studies covered a heterogeneous range of liver diseases and histological settings, including MASLD/NAFLD, MASH/NASH, biliary atresia, autoimmune hepatitis, primary sclerosing cholangitis, and mixed chronic liver disease populations. Sample sizes ranged from small proof-of-concept cohorts to multicentre datasets comprising several hundred biopsies or several thousand histopathology images.

Deep-learning architectures included conventional and residual convolutional neural networks, weakly supervised multiple-instance learning, graph convolutional networks, Transformer–ResNet architectures, U-Net-based segmentation models, CycleGAN-based virtual staining, YOLO object-detection models, and other end-to-end digital-pathology systems. Histological inputs included H&E, Masson trichrome, Sirius red, and picrosirius red whole-slide images, as well as multiphoton microscopy-derived images.

The 17 A-layer reports primarily evaluated ordinal fibrosis staging, dichotomized fibrosis thresholds, or model–pathologist agreement. The 11 B-layer reports addressed collagen segmentation, fibrosis-pattern detection, continuous fibrosis quantification, treatment-related fibrosis change, or spatial fibrosis phenotyping. Several reports originated from overlapping study programs or datasets and were grouped into study families to avoid treating correlated reports as independent studies (Table 1).

Table 1
Characteristics of studies included in the systematic review

Representative report	Layer	Population / disease	Sample size	Histological material / stain	Deep-learning approach	Reference standard / staging system	Valid
Levy 2021 [14]	A	Human liver biopsies	574 WSI	H&E WSI → virtual trichrome; paired real trichrome	CycleGAN-based virtual staining	Real trichrome with pathologist fibrosis staging	Paire real s comp
Pan 2026 [12]	A	MASLD-related fibrosis; human liver biopsies	467 total (327 training; 140 validation)	Masson Goldner / Masson trichrome digital biopsy images	Transformer + ResNet integrated model	Pathologist-assigned fibrosis grade (F0–F4)	Inde valid = 14 467-l
Zhan 2023 [18]	A	NAFLD liver biopsies	203 development + 42 + 33 external validation	Multiphoton microscopy / label-free fibrosis imaging	VGG16 / ResNet34 / MobileNetV3 + MLP (AutoFibroNet)	NASH-CRN fibrosis staging	Deve coho exter coho
Ratziu 2024 [16]	A	NASH/MASH clinical-trial participants	251 digitized patients	H&E + Masson digital biopsy slides	PathAI end-to-end CNN/GNN-based digital pathology	Consensus assessment by 2 expert pathologists	Post trial a
Panzeri 2025 [10]	A	MASH	59 biopsies/patients	Sirius-red digital biopsy images	ChatGPT-4 Vision image assessment	Senior pathologist fibrosis stage	Direc agair stagi
Heinemann 2022 [15]	A	NAFLD/NASH	296 development + 171 independent test	Human liver biopsy histology images	CNN-based deep learning	Pathologist histologic scores	Deve with test s
Jana 2020 [19]	A	NAFLD/NASH; human patients	30 patients	H&E pathology images ± CT	Deep CNN with pathology-only and multimodal fusion models	Pathologist fibrosis stage / NAS	Thre valid
Shabanian 2025 [13]	A	Children and young adults	217 trichrome WSI	Trichrome WSI	CLAM weakly supervised multiple-instance learning	Average of two pathologists using METAVIR and Ishak	Five- valid
Naik 2023 [11]	A	NAFLD; human liver biopsies	152 WSI	Sirius-red WSI	Weakly supervised multiple-instance deep learning	Expert pathologist fibrosis stage	Cros
Preechathammawong 2024 [8]	A	NAFLD	146 patients	Digital liver histology	SMART AI-PATHO; AI/DL architecture described in source/supplement	Pathologist METAVIR staging	Singl comp agair exac subg repo
Qu 2021 [20]	A	NAFLD/NASH	87 paired H&E/trichrome slide pairs	H&E + trichrome histopathology slides	Deep convolutional feature-extraction/classification network	Pathologist fibrosis stage / NAS	Repe valid stain magi setti
Wojciechowska 2021 [36]	A	NAFLD	271 biopsies	Picrosirius-red WSI	ResNet18 + graph convolutional / graph neural network	Pathologist fibrosis stage	Inter evalu

A-layer: direct fibrosis staging/classification/agreement evidence. B-layer: detection, segmentation, continuous quantification, spatial phenotyping, or related qua represented 25 study families; overlapping reports within the same family were not treated as statistically independent in the same pooled analysis. Abbreviation BA, biliary atresia; CLAM, clustering-constrained attention multiple-instance learning; CPA, collagen proportionate area; DL, deep learning; MASH, metabolic dysfu dysfunction-associated steatotic liver disease; NR, not reported; PSR, picrosirius red; WSI, whole-slide image.

Representative report	Layer	Population / disease	Sample size	Histological material / stain	Deep-learning approach	Reference standard / staging system	Valid
XFibrosis (Yin et al.) 2024 [21]	A	Human liver pathology; NAFLD and mixed validation datasets	132 Sirius-red WSI + 47 Masson/Triise WSI	Sirius-red and Masson/Triise liver biopsy WSI	Primal-dual graph convolutional vessel–fiber model (XFibrosis)	Expert pathologist fibrosis stage (FIB0–4)	Three valid data
Yang 2026 [9]	A	Children with biliary atresia	530 patients/WSI; test cohort n = 53	H&E whole-slide images	DSL-Net lightweight encoder–decoder model	BA-specific fibrosis staging system / pathologist reference	80% valid
Naglah 2022 [24]	B	Human liver transplant tissue	NR	H&E → virtual Masson WSI	Conditional GAN + fibrosis detector	Registered special-stain / fibrosis reference	Two
Wojciechowska 2026 (Decoding) [25]	B	MASLD + healthy volunteers	187 MASLD + 15 healthy	PSR-stained liver biopsies	Deep-learning collagen segmentation + collagen-deposition phenotype framework	Fibrosis stage / collagen proportionate area / omics associations	Coho phen asso analy
Wojciechowska 2026 (Enhancing) [26]	B	Four multicentre cohorts	686 digitized biopsies across > 20 sites	PSR-stained liver biopsy WSI	U-Net ensemble with uncertainty estimation	Dense collagen annotations / quantitative collagen reference	Train coho unse valid
Ercan 2024 [27]	B	Autoimmune hepatitis	123 paired biopsy slides/patients	Sirius-red WSI	CNN models on Aiforia platform	Pathologist manual annotations / Ishak context	Indep set e
Mizuochi 2025 [28]	B	Steatotic/steatohepatitis-associated liver disease	233 validation cases	Digital liver histology	CNN / DenseNet-based quantification	Pathologist fibrosis stage	Valid
Zou 2025 [29]	B	Human liver histopathology images	3976 images	Histopathology images	YOLOv4/v5/v7/v8/v9 comparison	Manual bounding-box labels	Imag 70/2 train/ split
PATHAI PSC conference report [17]	A	Primary sclerosing cholangitis; human liver biopsies	Training n = 110; validation n = 31	Trichrome and picrosirius-red digital histology	Annotation platform + convolutional neural networks	Central-pathologist Ishak fibrosis stage	Conf with traini sets
PathAI NASH cirrhosis heterogeneity conference report [30]	B	NASH cirrhosis; STELLAR/ATLAS clinical-trial biopsies	NR	Trichrome digital liver histology	Convolutional neural-network fibrosis models	Pathologist / trial histology review	Conf overl STEL prog
PathAI resmetirom conference report [31]	B	MASH/MASLD; resmetirom phase 2 biopsies	NR	Digital pathology / WSI	CNN cell/tissue models + fibrosis segmentation	Pathologist review / digital pathology comparator	Conf

A-layer: direct fibrosis staging/classification/agreement evidence. B-layer: detection, segmentation, continuous quantification, spatial phenotyping, or related qua represented 25 study families; overlapping reports within the same family were not treated as statistically independent in the same pooled analysis. Abbreviation BA, biliary atresia; CLAM, clustering-constrained attention multiple-instance learning; CPA, collagen proportionate area; DL, deep learning; MASH, metabolic dysfu dysfunction-associated steatotic liver disease; NR, not reported; PSR, picrosirius red; WSI, whole-slide image.

Representative report	Layer	Population / disease	Sample size	Histological material / stain	Deep-learning approach	Reference standard / staging system	Valid
Klimov 2019 AASLD conference report [22]	A	Human liver biopsies; mixed fibrosis stages	244 matched scans; 205 retained after exclusions	H&E WSI with matched Masson trichrome for training/reference	LinkNet semantic segmentation + whole-slide features + ML classifier	Expert fibrosis stage labels; S4 vs S0–S3	Cros inde set
PathAI disease progression conference report [32]	B	F3/F4 NASH; STELLAR/ATLAS biopsies	NR	Digital liver histology	End-to-end fibrosis model + deep CNN feature extraction	Pathologist histology + clinical outcome linkage	Conf overl STEL prog
Frankfurt NASH DL conference report [23]	A	NASH liver biopsies	80	Masson-stained liver biopsy slides	Deep-learning automated histology scoring model	Pathologist NAS / fibrosis scoring	Conf no re exac
DELTA Liver Fibrosis Score conference report [33]	B	Advanced NASH fibrosis; STELLAR-3 / ATLAS biopsies	NR	Digital pathology / fibrotic regions of interest	End-to-end image model (DELTA Liver Fibrosis Score)	Pathologist / trial histology response	Long clinic analy overl STEL prog
Panzeri collagen segmentation conference report [34]	B	Human liver biopsies	NR	Sirius-red digital biopsy images	U-Net3 + collagen segmentation + unsupervised feature extraction	Pathologist-assessed fibrosis progression	Conf abstr

A-layer: direct fibrosis staging/classification/agreement evidence. B-layer: detection, segmentation, continuous quantification, spatial phenotyping, or related qua represented 25 study families; overlapping reports within the same family were not treated as statistically independent in the same pooled analysis. Abbreviation BA, biliary atresia; CLAM, clustering-constrained attention multiple-instance learning; CPA, collagen proportionate area; DL, deep learning; MASH, metabolic dysfu dysfunction-associated steatotic liver disease; NR, not reported; PSR, picrosirius red; WSI, whole-slide image.

Table 2
Pooled diagnostic performance of deep learning-based digital pathology for liver fibrosis staging

Analysis	k	Metric	Pooled estimate	95% CI	I2 (%)	Interpretation
Primary F > = 3 exact 2x2	4	Sensitivity	0.870	0.577–0.971	64.3	Primary
Primary F > = 3 exact 2x2	4	Specificity	0.864	0.739–0.934	33.3	Primary
Primary F > = 3 exact 2x2	4	LR+	6.15	2.96–12.78	-	Primary
Primary F > = 3 exact 2x2	4	LR-	0.149	0.037–0.597	-	Primary
Primary F > = 3 exact 2x2	4	DOR	53.46	7.08-403.72	-	Primary
Sensitivity analysis: F > = 3 + Pan 2026	5	Sensitivity	0.857	0.716–0.935	57.2	Pan 2026 reported proportions only
Sensitivity analysis: F > = 3 + Pan 2026	5	Specificity	0.856	0.783–0.908	12.4	Pan 2026 reported proportions only
Exploratory F > = 2	3	Sensitivity	0.862	0.204–0.993	66.4	Exploratory
Exploratory F > = 2	3	Specificity	0.759	0.301–0.958	66.4	Exploratory
Exploratory F > = 2	3	LR+	3.35	0.99–11.34	-	Exploratory
Exploratory F > = 2	3	LR-	0.265	0.136–0.518	-	Exploratory
Exploratory F > = 2	3	DOR	16.76	5.39–52.09	-	Exploratory

Primary F > = 3 pooling used logit-transformed sensitivity/specificity with random-effects REML and Knapp-Hartung confidence intervals. For transformed pooling only, a 0.5 continuity correction was applied to all four cells of a study when any cell was zero; raw study-level sensitivity/specificity used uncorrected counts. Pan 2026 contributed reported sensitivity/specificity only in the sensitivity analysis. Abbreviations: DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

3.3 Risk of bias and applicability

Nine accuracy estimates were evaluated using QUADAS-3 at the estimate level (Fig. 2). Overall risk of bias was judged high for all nine estimates. In the Participants domain, five estimates (55.6%) were judged high risk and four (44.4%) had insufficient information. Index-test risk of bias was high for six

estimates (66.7%), low for one (11.1%), and based on insufficient information for two (22.2%). Target-condition risk of bias was low for four (44.4%) and based on insufficient information for five (55.6%). The Analysis domain was low risk for seven estimates (77.8%) and high risk for two (22.2%).

Overall applicability concerns were high for all nine estimates. Participant applicability was high concern for all nine estimates. Index-test applicability was low concern for five (55.6%), high concern for two (22.2%), and insufficiently reported for two (22.2%). Target-condition applicability was low concern for all nine estimates. The Klimov 2019 conference report was judged high risk in the Participants, Index-test, and Analysis domains because of selective exclusions and cross-validation-based feature and model selection without an independent test set. The PATHAI PSC and Frankfurt NASH conference reports were not selected for estimate-level QUADAS-3 assessment because they did not provide a directly extractable fibrosis-stage diagnostic-accuracy estimate suitable for the prespecified synthesis.

3.4 Primary diagnostic-accuracy synthesis for advanced fibrosis

Four independent study families provided reconstructable 2×2 data for the primary advanced-fibrosis analysis. Preechathammawong 2024 and Yang 2026 contributed two of these datasets [8, 9]. Panzeri 2025 and Naik 2023 contributed the other two [10, 11]. Although the underlying disease populations and histological staging systems differed, all four analyses represented a binary advanced-fibrosis threshold corresponding to F ≥ 3 or its etiological equivalent.

The pooled sensitivity for advanced fibrosis was 0.87 (95% CI 0.58–0.97), with moderate-to-substantial between-study heterogeneity (I^2=64.3%). The pooled specificity was 0.86 (95% CI 0.74–0.93), with lower heterogeneity (I^2=33.3%) (Fig. 3A–B).

The pooled positive likelihood ratio was 6.15 (95% CI 2.96–12.78), whereas the pooled negative likelihood ratio was 0.15 (95% CI 0.04–0.60). The corresponding diagnostic odds ratio was 53.46 (95% CI 7.08–403.72). Because only four independent study families were available and the full bivariate model yielded a boundary random-effects correlation, a formal HSROC model was not retained as the primary inference.

3.5 Sensitivity analysis incorporating Pan 2026

Pan 2026 [12] reported advanced-fibrosis sensitivity and specificity but did not provide an exact reconstructable 2×2 table. The reported sensitivity of 0.86 was incompatible with a unique integer numerator among the 52 advanced-fibrosis cases under conventional two-decimal rounding. Accordingly, Pan 2026 was excluded from the exact 2×2 primary synthesis but was incorporated into a sensitivity analysis using the reported sensitivity and specificity and the known positive and negative denominators.

After inclusion of Pan 2026, pooled sensitivity was 0.857 (95% CI 0.716–0.935; I^2=57.2%), and pooled specificity was 0.856 (95% CI 0.783–0.908; I^2=12.4%). These estimates were closely aligned with the four-study primary analysis, indicating that inclusion of the additional reported proportions did not materially alter the overall diagnostic-performance estimates.

3.6 Exploratory analysis of significant fibrosis

Three study families provided compatible 2×2 data for significant fibrosis defined as F ≥ 2 or an equivalent threshold. Shabanian 2025 contributed one dataset [13]. The other two were Preechathammawong 2024 and Panzeri 2025 [8, 10].

The pooled sensitivity was 0.862 (95% CI 0.204–0.993; I^2=66.4%), and pooled specificity was 0.759 (95% CI 0.301–0.958; I^2=66.4%). The pooled LR + was 3.35 (95% CI 0.99–11.34), LR − was 0.265 (95% CI 0.136–0.518), and DOR was 16.76 (95% CI 5.39–52.09).

These results were considered exploratory because only three study families were available and the negative-reference spectrum in Panzeri 2025 was restricted, with no F0 cases. The F ≥ 2 analysis was therefore not treated as a co-primary diagnostic endpoint.

3.7 Agreement between digital-pathology models and histological reference assessments

Agreement outcomes were heterogeneous in both definition and statistical formulation and therefore were not pooled into a single κ meta-analysis (Table 3).

Table 3
Agreement between deep learning-based digital pathology and histopathological fibrosis assessment

Study / report	Study family	Comparison / endpoint	Agreement metric	Value	Uncertainty / supporting result	Metric source	Role / note
Levy 2021 [14]	LEVY-VIRTUAL-TRICHROME	Virtual vs real trichrome; advanced-fibrosis/stage agreement	Cohen κ	0.76	95% CI 0.65–0.87; stage correlation ρ = 0.86 (95% CI 0.84–0.88)	Reported	Narrative agreement
Ratziu 2024 [16]	SEMAGLUTIDE-PATHAI	Categorical fibrosis stage (F0–F4), baseline	Weighted κ	0.3405	Overall categorical accuracy ≈ 0.498; baseline→week72 change κ ≈ 0.2190	Reported	Narrative agreement
Panzeri 2025 [10]	HUMANITAS-MASH-01	Full fibrosis staging vs senior pathologist	Weighted κ	0.78	Overall staging accuracy = 0.81	Reported	Narrative agreement
Heinemann 2022 [15]	BI-MASLD-467	Stage-level fibrosis agreement in independent test set	Quadratic weighted κ	0.62	Fibrosis MAE = 0.76 on independent test set	Reported	Narrative agreement
Preechathammawong 2024 [8]	PREECHATHAMMAWONG-DL	Overall METAVIR F0–F4 agreement	Cohen κ	0.5948	Overall agreement = 68.49%; p < 0.001	Reported	Narrative agreement
Preechathammawong 2024 [8]	PREECHATHAMMAWONG-DL	Significant fibrosis F ≥ 2 vs F0–F1	Binary Cohen κ	0.6153	Agreement = 80.82%; p < 0.001	Reported	Exploratory F ≥ 2
Preechathammawong 2024 [8]	PREECHATHAMMAWONG-DL	Advanced fibrosis F ≥ 3 vs F0–F2	Binary Cohen κ	0.7401	Agreement = 89.73%; p < 0.001	Reported	Primary F ≥ 3
Yang 2026 [9]	YANG-BA-DSLNET	Advanced fibrosis-equivalent ≥Stage 3	Binary Cohen κ	0.8271	Derived from exact 2×2 table (TP = 34, FP = 1, FN = 3, TN = 15)	Derived from exact 2×2	Primary F ≥ 3-equivalent
Panzeri 2025 [10]	HUMANITAS-MASH-01	Advanced fibrosis F ≥ 3 vs lower stages	Binary Cohen κ	0.8245	Derived from exact 2×2 table (TP = 33, FP = 5, FN = 0, TN = 21)	Derived from exact 2×2	Primary F ≥ 3
Naik 2023 [11]	NAIK-LIFIB	Advanced fibrosis F ≥ 3 vs F ≤ 2	Binary Cohen κ	0.5789	Derived from exact 2×2 table (TP = 57, FP = 13, FN = 19, TN = 63)	Derived from exact 2×2	Primary F ≥ 3
Shabanian 2025 [13]	SHABANIAN-PEDS	Significant fibrosis F ≥ 2 (METAVIR)	Cohen κ	0.57 ± 0.15	Five-fold cross-validation; aggregated exact 2×2 gives κ = 0.5778	Reported and independently derived	Exploratory F ≥ 2
PATHAI PSC conference report [17]	PATHAI-PSC	ML Ishak score vs central-pathologist Ishak stage	Spearman correlation ρ	0.67	ρ = 0.67 for both trichrome and picrosirius-red images	Reported	Narrative agreement

Reported agreement metrics and binary kappa values derived from exact 2x2 tables are intentionally distinguished and were not pooled. For Heinemann 2022, the detailed Results and Table 1 *of the source article report fibrosis test-set MAE = 0.76, although the article abstract reports 0.78.*

Levy 2021 [14] reported a Cohen κ of 0.76 (95% CI 0.65–0.87) for advanced-fibrosis agreement between virtual and real trichrome assessment, accompanied by a bias-adjusted stage correlation of ρ = 0.86 (95% CI 0.84–0.88). Heinemann 2022 [15] reported a quadratic weighted κ of 0.62 for fibrosis staging and a fibrosis mean absolute error of 0.76 in the independent test set. Ratziu 2024 [16] reported lower categorical agreement, with baseline agreement of approximately 49.8% and weighted κ of approximately 0.34. Panzeri 2025 [10] reported a weighted κ of 0.78 and overall staging accuracy of 0.81.

For the four studies with exact F ≥ 3 binary tables, descriptive Cohen κ values derived from the same 2×2 data were 0.7401 for Preechathammawong 2024, 0.8271 for Yang 2026, 0.8245 for Panzeri 2025, and 0.5789 for Naik 2023. These derived binary κ estimates were not formally pooled because a sufficiently justified and harmonized study-level variance framework was not available and because they represented a different agreement construct from reported weighted or quadratic weighted κ statistics.

The PATHAI PSC conference report [17] provided additional agreement-related evidence: the machine-learning Ishak fibrosis score correlated with central-pathologist Ishak stage (ρ = 0.67 for both trichrome and picrosirius-red images). This evidence was synthesized narratively rather than treated as a binary diagnostic-accuracy estimate.

3.8 AUC-based and other non-poolable staging-performance evidence

Several studies reported AUCs or other staging-performance measures that could not be incorporated into the exact 2×2 synthesis (Table 4). Formal AUC pooling was not performed because target definitions, validation designs, and measures of dispersion differed substantially, and several reported “±” values represented cross-validation variability rather than sampling standard errors.

Table 4
Reported AUCs/AUROCs and other non-poolable staging-performance evidence

Study / report	Study family	Endpoint	Validation / dataset	N	AUC / AUROC	Dispersion / precision	Diagnostic vs prognostic	Reason not pooled as AUC
Pan 2026 [12]	BI-MASLD-467	Advanced fibrosis F3–F4 vs F0–F2	Validation cohort	140	0.90	NR	Diagnostic	Exact 2×2 was not uniquely reconstructable from rounded sensitivity/specificity; used only in sensitivity analysis
Pan 2026 [12]	BI-MASLD-467	Cirrhosis F4 vs F0–F3	Validation cohort	140	0.88	NR	Diagnostic	Different target condition from primary F ≥ 3 synthesis
Zhan 2023 [18]	AUTOFIBRONET-WENZHOU	F3–4 one-vs-rest	External validation cohort 1	42	0.90	Reported directly	Diagnostic	AUC-only external validation
Zhan 2023 [18]	AUTOFIBRONET-WENZHOU	F3–4 one-vs-rest	External validation cohort 2	33	0.94	Reported directly	Diagnostic	AUC-only external validation
Jana 2020 [19]	JANA-MULTIMODAL	Fibrosis-stage prediction	H&E pathology-only; 3-fold CV	30	0.8385	± 0.0968 across folds	Diagnostic	Multiclass stage-prediction AUC; cross-validation dispersion
Jana 2020 [19]	JANA-MULTIMODAL	Fibrosis-stage prediction	Mid-fusion single-loss (CT + H&E); 3-fold CV	30	0.8984	± 0.0792 across folds	Diagnostic	Multimodal model; not directly comparable with pathology-only DTA threshold
Qu 2021 [20]	QU-NAFLD-DL	Fibrosis-stage prediction	Repeated experiments across H&E/trichrome and magnifications	87 paired slide sets	0.8170–0.8385	Setting-specific variability reported	Diagnostic	Different stain/magnification settings and multiclass staging
XFibrosis (Yin et al.) 2024 [21]	XFIBROSIS	FIB ≥ 3	Liver-Fibrosis-SR (Sirius red); 3-fold CV	132	0.8611	± 0.0736 across folds	Diagnostic	Cross-validation SD is not a sampling SE
XFibrosis (Yin et al.) 2024 [21]	XFIBROSIS	FIB ≥ 3	FibrosisMT-v1.0 (Masson/Triise); 3-fold CV	47	0.9572	± 0.0605 across folds	Diagnostic	Cross-validation SD is not a sampling SE
Shabanian 2025 [13]	SHABANIAN-PEDS	Significant fibrosis F ≥ 2 (METAVIR)	5-fold cross-validation	217	0.88	± 0.07 across folds	Diagnostic	Exploratory threshold; cross-validation SD
Naik 2023 [11]	NAIK-LIFIB	Advanced fibrosis F ≥ 3	Cross-validation	152	0.87	± 0.06 across folds	Diagnostic	Primary synthesis uses exact 2×2; AUC dispersion is cross-validation SD
Yang 2026 [9]	YANG-BA-DSLNET	BA-specific ≥Stage 3	Independent test cohort	53	0.906	NR	Diagnostic	Etiology-specific threshold; primary synthesis uses exact 2×2
Klimov 2019 AASLD conference report [22]	KLIMOV-LINKNET-205	Cirrhosis S4 vs S0–S3	Cross-validated conference report	205 retained cases	0.70–0.93 (individual features)	Classifier AUC not reported	Diagnostic	Feature-level AUC range only; final optimized classifier reported sensitivity/specificity rather than AUC
PATHAI PSC conference report [17]	PATHAI-PSC	Progression to cirrhosis	Trichrome / picrosirius red; conference validation	31 validation cases	0.86 / 0.76	NR	Prognostic	Prognostic AUROC, not direct staging-accuracy AUC
Frankfurt NASH DL conference report [23]	FRANKFURT-NASH-DL80	Automated fibrosis scoring	Conference report	80	NR	NR	Diagnostic / agreement narrative	No extractable harmonized AUC or exact 2×2 in available report

Cross-validation +/- values represent fold-to-fold variability and were not treated as sampling standard errors. For Klimov 2019, 0.70–0.93 refers to individual whole-slide feature AUCs; a final optimized classifier AUC was not reported.

Pan 2026 [12] reported an AUC of 0.90 for F3–F4 advanced fibrosis and 0.88 for cirrhosis. Zhan 2023 [18] reported external-validation AUCs of 0.90 and 0.94 for F3–4 classification in two external cohorts. Jana 2020 [19] reported an AUC of approximately 0.839 for the pathology-only model and approximately

0.898 for the multimodal mid-fusion model. Qu 2021 [20] reported fibrosis-stage AUCs ranging approximately from 0.817 to 0.839 across staining and magnification settings.

XFibrosis [21] reported AUCs of approximately 0.861 in the Sirius-red dataset and 0.957 in the Masson-trichrome dataset for FIB ≥ 3. Shabanian 2025 [13] and Naik 2023 [11] reported cross-validation AUCs of approximately 0.88 and 0.87, respectively, although the associated dispersion estimates reflected fold-to-fold variation rather than sampling uncertainty. Yang 2026 [9] reported an AUC of 0.906 for the biliary-atresia-specific ≥Stage 3 threshold.

The Klimov 2019 conference study [22] used a LinkNet-based deep-learning segmentation stage followed by whole-slide feature extraction and machine-learning classification of S4 versus S0–S3. In the report, individual whole-slide features yielded AUCs ranging from 0.70 to 0.93, while the optimized classifier achieved cross-validated sensitivity above 95% and specificity of 95%. Because the final classifier did not provide an exact reconstructable 2×2 table and the target was cirrhosis rather than the primary F ≥ 3 threshold, the study was retained as structured staging-performance evidence rather than added to the primary meta-analysis.

The PATHAI PSC conference report [17] further showed that the machine-learning Ishak fibrosis score predicted subsequent progression to cirrhosis, with reported AUROCs of 0.86 for trichrome and 0.76 for picrosirius-red images. These were prognostic rather than direct staging-accuracy AUCs and were therefore not combined with diagnostic AUC evidence. The Frankfurt NASH conference report [23] similarly provided direct automated fibrosis-scoring evidence but did not report a harmonizable exact diagnostic-accuracy estimate; it was therefore retained as structured A-layer staging-performance evidence (Table 4) rather than pooled with the primary synthesis.

3.9 Secondary fibrosis detection, segmentation, and quantitative digital-pathology evidence

Eleven B-layer reports evaluated fibrosis-related digital-pathology tasks that were not directly comparable with conventional binary staging accuracy (Table 5).

Table 5

Deep learning-based fibrosis detection, segmentation, and quantitative assessment not included in the primary diagnostic-accuracy meta

Study / report	Study family	Task	Architecture	Input modality	Validation / sample	Main quantitative results
Naglah 2022 [24]	NAGLAH-CGAN	Fibrosis detection / semantic segmentation	cGAN + fibrosis detector	H&E → virtual Masson WSI	Two validation sets	Set 1: accuracy 0.87 (95% CI 0.85–0.88), Dice 0.24 (0.22–0.26); Set 2: accuracy 0.78 (0.77–0.79), Dice 0.45 (0.42–0.47)
Wojciechowska 2026 (Decoding) [25]	WOJ-DECODING	Collagen-deposition phenotyping / quantification	DL collagen segmentation + CDP framework	PSR-stained liver biopsies	187 MASLD + 15 healthy	Continuous spatial collagen-deposition phenotypes and associations with fibrosis/clinical/transcriptomic features
Wojciechowska 2026 (Enhancing) [26]	WOJ-ENHANCING	Collagen segmentation / fibrosis measurement	U-Net ensemble + uncertainty estimation	PSR liver biopsy WSI	686 biopsies across 4 cohorts; unseen HepaT1ca validation	Training/internal cohorts Dice ≈ 0.83–0.90; unseen HepaT1ca U-Net Mini Dice = 0.555 ± 0.104; uncertainty maps used for QC
Ercan 2024 [27]	AIH-AIFORIA	Fibrosis-pattern detection in autoimmune hepatitis	CNN models on Aiforia platform	Sirius-red WSI	Independent test-set slides	Fibrosis detection: accuracy 87.6%, precision 73.3%, recall 68.3%
Mizuochi 2025 [28]	MIZUOCHI-DL-Q	Continuous fibrosis quantification	CNN / DenseNet	Digital liver histology	233 validation cases; SLD subgroup n = 47	DL-derived %fibrosis correlated with histological fibrosis stage: R = 0.72, p < 0.001 in SLD subgroup
Zou 2025 [29]	ZOU-YOLO	Fibrosis and liver-lesion detection / localization	YOLOv4/v5/v7/v8/v9	Histopathology images	3976 images; 70/20/10 image-level split	YOLOv9 precision = 0.88, recall = 0.91, F1 = 0.89, mAP@0.5 = 0.86; fibrosis-specific mAP = 0.82
PathAI NASH cirrhosis heterogeneity conference report [30]	PATHAI-STELLAR-ATLAS	Intra-cirrhotic fibrosis heterogeneity quantification	Convolutional fibrosis models	Trichrome digital histology	STELLAR/ATLAS clinical-trial biopsies	Stage-area distributions and spatial fibrosis heterogeneity described; clinical-event associations explored
PathAI resmetirom conference report [31]	PATHAI-RESMETIROM-PH2	Continuous fibrosis/collagen quantification	CNN cell/tissue models + fibrosis segmentation	Digital pathology / WSI	Resmetirom phase 2 biopsies	Quantitative fibrosis/collagen changes evaluated in treatment-response analyses
PathAI progression conference report	PATHAI-STELLAR-ATLAS	Histologic predictors of disease progression	End-to-end fibrosis model + deep CNN feature extraction	Digital liver histology	F3/F4 NASH; STELLAR/ATLAS program	Fibrosis-feature–based progression associations reported
DELTA Liver Fibrosis Score conference report [33]	PATHAI-STELLAR-ATLAS	Longitudinal fibrosis-response quantification	End-to-end image model (DELTA)	Digital fibrotic ROIs	STELLAR-3 / ATLAS biopsies	Longitudinal response metrics and fibrosis-stage distribution changes reported
Panzeri collagen segmentation conference report [34]	PANZERI-UNET3P-COLLAGEN	Collagen segmentation / unsupervised fibrosis-feature extraction	U-Net3 + + latent-feature extraction	Sirius-red digital biopsy images	Conference report	Segmentation accuracy > 95%; latent collagen features associated with pathologist-assessed fibrosis progression

These B-layer reports were synthesized narratively because their endpoints were segmentation, continuous quantification, spatial phenotyping, treatment respon harmonized patient-level binary fibrosis staging.

Naglah 2022 [24] evaluated fibrosis segmentation using a conditional GAN-based framework. In one validation set, accuracy was 0.87 (95% CI 0.85–0.88) with a Dice coefficient of 0.24 (95% CI 0.22–0.26); in the second validation set, accuracy was 0.78 (95% CI 0.77–0.79) with a Dice coefficient of 0.45 (95% CI 0.42–0.47).

Wojciechowska and colleagues reported two complementary quantitative fibrosis frameworks. The collagen-deposition phenotype study [25] linked AI-derived spatial collagen patterns to fibrosis severity and biological features, whereas the multicentre U-Net study [26] evaluated collagen segmentation and fibrosis measurement across four cohorts comprising 686 biopsies. In the latter study, performance varied across cohorts, including a Dice value of approximately 0.555 ± 0.104 in the previously unseen HepaT1ca cohort, highlighting the challenge of generalization across centres.

Ercan 2024 [27] reported approximately 87.6% accuracy, 73.3% macro precision, and 68.3% macro recall for fibrosis-pattern classification in autoimmune hepatitis. Mizuochi 2025 [28] demonstrated a correlation between deep-learning-derived percentage fibrosis and histological fibrosis stage (R = 0.72, p < 0.001) in the steatotic liver disease subgroup.

Zou 2025 [29] evaluated multiple YOLO architectures for histopathological object detection. YOLOv9 achieved overall precision of 0.88, recall of 0.91, F1 score of 0.89, and mAP@0.5 of 0.86, with fibrosis-specific mAP of approximately 0.82.

Additional conference reports extended the evidence beyond conventional staging. PathAI-associated analyses characterized intra-stage fibrosis heterogeneity [30] and quantified treatment-related histologic changes after resmetirom [31]. Separate reports identified fibrosis features associated with disease progression [32] and evaluated longitudinal treatment response with the DELTA Liver Fibrosis Score [33]. A U-Net3+-based collagen-segmentation study [34] reported segmentation accuracy exceeding 95% and identified latent collagen features associated with pathologist-assessed fibrosis progression. Because these studies used fundamentally different endpoints—including segmentation accuracy, continuous fibrosis burden, spatial phenotyping, treatment response, and prognostic associations—they were synthesized narratively rather than pooled with the A-layer diagnostic-accuracy estimates.

4. Discussion

4.1 Principal findings

This review identified 28 eligible reports representing 25 study families across two analytically distinct evidence layers. Seventeen A-layer reports evaluated fibrosis staging, classification, or agreement, whereas 11 B-layer reports assessed segmentation, detection, continuous quantification, or spatial phenotyping. Four independent study families contributed exact 2×2 data to the primary F ≥ 3-equivalent meta-analysis, yielding pooled sensitivity of 0.87 (95% CI 0.58–0.97), specificity of 0.86 (95% CI 0.74–0.93), LR + of 6.15, LR − of 0.15, and DOR of 53.46. Incorporating Pan 2026 as reported proportions produced similar sensitivity and specificity, while agreement, AUC, and B-layer outcomes remained too heterogeneous for a single pooled estimate. All nine diagnostic-accuracy estimates assessed with QUADAS-3 had high overall risk of bias and high overall applicability concern. Taken together, the review identifies a promising but still methodologically immature evidence base.

4.2 Clinical interpretation of the diagnostic-accuracy findings

The pooled sensitivity and specificity are clinically encouraging, but they do not establish that an autonomous deep-learning system can replace histopathological assessment. An LR + of approximately 6 can materially increase the post-test probability of advanced fibrosis, and an LR− near 0.15 can reduce it; however, the associated confidence intervals were wide and arose from only four independent study families. These estimates are therefore better viewed as provisional measures of discrimination than as universal rule-in or rule-out thresholds. Their clinical meaning depends on disease spectrum, pretest probability, slide quality, reference-standard procedures, and the intended workflow.

Importantly, digital pathology does not remove the invasiveness or sampling error of liver biopsy. The paired-biopsy discordance described in the Introduction [2] arises before an algorithm sees the slide: two cores from the same liver may contain different amounts or patterns of fibrosis. A deep-learning model can standardize interpretation and quantify features within the tissue that was actually sampled, but it cannot recover fibrotic architecture that is absent from that core. Its most defensible value boundary is therefore improved reproducibility of slide interpretation, quantitative augmentation, multicentre harmonization, and potentially more sensitive clinical-trial endpoints—not correction of biopsy sampling representativeness itself.

This distinction also clarifies where digital pathology sits in the clinical pathway. In many care pathways, non-invasive tests such as FIB-4 and elastography are used upstream for risk stratification and to inform whether further specialist evaluation or biopsy is needed. Deep-learning pathology acts downstream, after tissue has been obtained, and its direct comparator or complement is manual histopathological interpretation rather than non-invasive triage. The pooled evidence also spans biologically different populations: Yang 2026 [9] evaluated a pediatric biliary-atresia-specific ≥Stage 3 threshold, whereas the other primary studies mainly addressed adult metabolic liver disease. This spectrum and etiologic difference is a plausible contributor to the moderate-to-substantial heterogeneity in sensitivity (I^2=64.3%), but with k = 4 its contribution cannot be isolated from differences in stains, thresholds, reference standards, or model architectures by subgroup analysis or meta-regression.

The sensitivity analysis incorporating Pan 2026 [12] changed the pooled estimates little, which supports the numerical stability of the primary summary to that additional report. Equally important, exact TP, FP, FN, and TN counts were not forced from rounded proportions. Avoiding such pseudo-precision preserves the distinction between directly observed diagnostic tables and secondary performance summaries. This emphasis on what the data can and cannot support provides the basis for interpreting the agreement measures and risk-of-bias findings that follow.

4.3 Agreement, heterogeneity, and domain shift

Agreement evidence addressed several different constructs. Weighted or quadratic weighted κ statistics describe concordance across ordinal fibrosis stages, whereas binary κ values derived from exact 2×2 tables quantify agreement at one dichotomized threshold; correlation coefficients assess association rather than categorical agreement. These measures are not interchangeable and were therefore not pooled. The observed range—from modest categorical concordance in a clinical-trial analysis to stronger agreement in virtual-staining assessment—supports a near-term role centered on decision support, second reading, and quantitative augmentation rather than unsupervised replacement of pathologists [14, 16].

The QUADAS-3 profile identifies why apparently strong headline accuracy should be interpreted cautiously. In the Participants domain, five of nine estimates were high risk and four had insufficient information, reflecting selected disease spectra, restricted cohorts, or exclusions that could affect case mix; participant applicability was high concern for all nine estimates. Index-test risk was high for six estimates and insufficiently reported for two, commonly because model or threshold development, cross-validation, feature selection, or implementation details limited independence from model-development decisions. Target-condition risk was low in four estimates but insufficiently reported in five, indicating that the definition of fibrosis itself was often clinically recognizable while blinding, reference-standard implementation, or independence details were incompletely documented. The Analysis domain was more reassuring overall—seven estimates were low risk—but Shabanian and Klimov were high risk because analysis/model-selection procedures did not provide a

fully independent evaluation. Thus, the pooled sensitivity of 0.87 and specificity of 0.86 should be regarded as provisional estimates that may be optimistic, not as transportable performance benchmarks for routine practice.

Technical domain shift adds a separate layer of uncertainty. The included studies used H&E, trichrome, Sirius red, picrosirius red, and multiphoton-derived images; architectures ranged from CNNs and residual networks to graph models, multiple-instance learning, Transformers, U-Net variants, GANs, and YOLO detectors. The multicentre U-Net study provides a concrete example: Dice performance of approximately 0.83–0.90 in training/internal cohorts fell to 0.555 ± 0.104 in the unseen HepaT1ca cohort [26]. Such a drop demonstrates why an apparently accurate internal model cannot be assumed to generalize across laboratories, stains, scanners, or patient populations. The methodological pattern is therefore at least as important as the pooled point estimates when relating this review to the wider literature.

4.4 Relation to previous evidence

The present findings are directionally consistent with the broader digital-pathology literature but are more cautious in magnitude and interpretation. McGenity et al. reported pooled sensitivity of 96.3% and specificity of 93.3% across AI applications to whole-slide images from many disease areas, while also finding that 99% of included studies had at least one high or unclear risk-of-bias or applicability concern [5]. Those estimates are not directly comparable with this liver-fibrosis synthesis because the targets, validation designs, and meta-analytic models differ, but the methodological warning is highly relevant. Similarly, a systematic review of non-invasive deep learning for liver fibrosis found that 87.5% of studies had high risk of bias and applicability concerns [4]. Across settings, high apparent discrimination has repeatedly coexisted with limitations in participant selection, reference standards, model development, and external validation.

Evidence-source independence also deserves attention. Several included reports arose from commercial or industry-sponsored development programs, including PathAI-associated analyses in PSC and STELLAR/ATLAS settings [17, 30]. Related developer-associated reports addressed resmetirom and disease progression [31, 32]. The DELTA Liver Fibrosis Score report provided further industry-linked evidence [33]. Other studies involved commercial platforms or industry collaborators [16, 27]. Such provenance does not invalidate the results, and industry partnerships are common in translational digital pathology. However, when a developer or sponsor is closely involved in model creation and evaluation, independent replication, transparent protocol reporting, disclosure of model changes, and access to complete performance data become especially important because selective reporting or optimism cannot be excluded. Evidence-source independence is not itself a formal QUADAS-3 domain in this review, so this concern should be considered complementary to—not a substitute for—the structured risk-of-bias assessment.

This review also differs from prior work by separating direct stage classification from quantitative fibrosis phenotyping. Digital pathology can measure collagen burden, spatial architecture, and within-stage heterogeneity that are not reducible to an ordinal F0–F4 label [3, 25]. The B-layer evidence should therefore not be interpreted as failed diagnostic-accuracy evidence; it represents a different translational pathway. Continuous or spatial fibrosis measures may ultimately improve detection of progression or regression, trial-endpoint sensitivity, and characterization of heterogeneity, but those potential advantages require validation against reproducible biological changes and patient-level outcomes. These distinctions also frame the principal limitations and strengths of the present review.

4.5 Strengths and limitations

The limitations are substantial and should temper interpretation of all pooled results. First, the primary meta-analysis contained only four independent study families. This limited precision, produced a wide sensitivity confidence interval, and precluded credible meta-regression or subgroup analyses by etiology, stain, scanner, staging system, or model architecture; consequently, the sources of I^2=64.3% for sensitivity can only be discussed descriptively rather than attributed quantitatively. Second, all nine QUADAS-3 estimates had high overall risk of bias and high overall applicability concern, with particular weaknesses in participant selection and index-test development or validation. Third, several studies relied on cross-validation rather than a locked external test cohort, and disease-specific thresholds were not fully interchangeable. Formal small-study-effect or publication-bias analyses were not performed because the quantitative syntheses were too small for reliable funnel-plot or regression-based assessment.

The composition of the evidence base introduces additional uncertainty. Five of the 11 B-layer reports were conference-level reports rather than full peer-reviewed articles, and three A-layer reports were also conference reports. Their inclusion reduced the risk of discarding potentially relevant gray literature but also introduced incomplete methodological and reporting detail. This should not be equated automatically with publication bias, which was not formally estimable here. In addition, the review did not perform a separate structured assessment of sponsor or developer independence beyond the domains captured by QUADAS-3. Finally, PROSPERO registration was retrospective (CRD420261500208), so readers should distinguish clearly between analyses that were predefined in the completed review framework and decisions that could have been influenced by the observed evidence.

Against these limitations, several design features strengthen the review. Six databases were searched without year or language restrictions, and conference abstracts and posters were not excluded solely because of publication type. Reports arising from the same cohort or clinical-trial program were explicitly linked into study families to reduce double counting. Quantitative synthesis was restricted to exact 2×2 data that were directly available or uniquely reconstructable, and rounded proportions were not converted into arbitrary integer tables. Risk of bias and applicability were assessed at the accuracy-estimate level using QUADAS-3 [7], and direct staging evidence was separated from segmentation and continuous quantitative outcomes so that fundamentally different tasks were not combined in one pooled estimate. These safeguards make the remaining uncertainty visible rather than removing it, and they point directly to priorities for future studies.

4.6 Implications for future research

Future diagnostic-accuracy studies should prioritize multicentre external validation with patient-level separation between development and test data, locked models, prespecified clinically meaningful fibrosis thresholds, and complete reporting of confusion matrices. Stain and scanner robustness should be tested

explicitly in sites and disease populations not represented in model development. Calibration, decision-curve analysis, and prospective human-in-the-loop studies are needed to determine whether image-level discrimination improves pathologist consistency, trial-endpoint reliability, or patient-level decisions. Studies should also report non-evaluable slides, failure modes, uncertainty, and the performance consequences of distribution shift. For B-layer tools, continuous digital features should be linked to reproducible histological change, treatment response, and clinical outcomes rather than validated solely by segmentation metrics.

Reproducibility and study planning also require greater attention. Model and code availability was inconsistent across the evidence base: some investigators made methods or tools openly available [15, 26], whereas several commercial platforms remained proprietary. Future work should therefore provide open code and representative test data when feasible, or at minimum publish a locked evaluation protocol, model version, preprocessing pipeline, and sufficient test-set outputs to permit independent assessment and regulatory review. A priori sample-size justification, prospective registration or protocol publication where appropriate, and adherence to diagnostic-accuracy reporting standards such as STARD 2015 [35] would reduce analytic flexibility and improve interpretability. These steps are necessary for the field to move from promising retrospective performance toward clinically credible, independently reproducible, and auditable pathology decision support.

5. Conclusion

Deep-learning-based digital pathology shows promising diagnostic performance for advanced liver fibrosis and provides additional capabilities for agreement assessment, collagen segmentation, and continuous fibrosis quantification. In the primary F > = 3 synthesis, pooled sensitivity and specificity were approximately 0.87 and 0.86, respectively, and results were similar after incorporation of an additional study reporting rounded sensitivity and specificity. However, the small number of independently evaluable studies, methodological heterogeneity, high risk of bias, applicability concerns, and limited external validation preclude claims of routine clinical readiness. The most defensible current role is as pathology decision support and quantitative augmentation, with future progress dependent on multicentre validation, standardized reporting, and prospective clinical evaluation.

Declarations

Ethics approval and consent to participate:

Not applicable. This study is a systematic review and meta-analysis of published data.

Consent for publication:

Not applicable.

Competing interests:

The authors declare no conflicts of interest.

Generative AI disclosure

Generative AI tools were used only for language polishing, formatting, and improving the clarity of the manuscript. They were not used to make final decisions regarding study eligibility, data extraction, risk-of-bias judgments, statistical analysis, or scientific interpretation. All scientific content, analyses, and conclusions were reviewed and verified by the authors.

Funding:

This research received no external funding.

Author Contribution

Qiuxia Qiu: Conceptualization, methodology, literature search, study selection, data extraction, statistical analysis, visualization, and writing—original draft. Qiuer Qiu: Study selection, data extraction, data verification, and writing—review and editing. Jiapeng Zhang: Conceptualization, methodology, supervision, adjudication, statistical verification, interpretation of results, project administration, and writing—review and editing. All authors read and approved the final manuscript.

Data Availability

The data analyzed in this systematic review and meta-analysis were extracted from published studies. All data supporting the findings of this study are available within the article and its supplementary materials. No new individual-level patient data were generated.

References

1. European Association for the Study of the Liver (EASL), European Association for the Study of Diabetes (EASD). European Association for the Study of Obesity (EASO). EASL-EASD-EASO Clinical Practice Guidelines on the management of metabolic dysfunction-associated steatotic liver disease (MASLD). J Hepatol. 2024;81(3):492–542. https://doi.org/10.1016/j.jhep.2024.04.031.
2. Ratziu V, Charlotte F, Heurtier A, et al. Sampling variability of liver biopsy in nonalcoholic fatty liver disease. Gastroenterology. 2005;128(7):1898–906. https://doi.org/10.1053/j.gastro.2005.03.084.
3. Miyaaki H, Miuma S, Fukusima M, et al. Liver fibrosis analysis using digital pathology. Med Mol Morphol. 2024;57(3):161–6. https://doi.org/10.1007/s00795-024-00395-y.
4. Anteby R, Klang E, Horesh N, et al. Deep learning for noninvasive liver fibrosis classification: a systematic review. Liver Int. 2021;41(10):2269–78. https://doi.org/10.1111/liv.14966.
5. McGenity C, Clarke EL, Jennings C, et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. NPJ Digit Med. 2024;7:114. https://doi.org/10.1038/s41746-024-01106-8.
6. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71.
7. Whiting PF, Tomlinson E, Rutjes AWS, et al. QUADAS-3: a revised tool for the quality assessment of diagnostic test accuracy studies. Ann Intern Med. 2026. https://doi.org/10.7326/ANNALS-25-02104.
8. Preechathammawong N, Charoenpitakchai M, Wongsason N, et al. Development of a diagnostic support system for the fibrosis of nonalcoholic fatty liver disease using artificial intelligence and deep learning. Kaohsiung J Med Sci. 2024;40(8):757–65. https://doi.org/10.1002/kjm2.12850.
9. Yang Q, Chen Y, Meng Y, et al. Development of a lightweight deep learning model for accurate assessment of liver fibrosis in biliary atresia. Pediatr Surg Int. 2026;42(1):77. https://doi.org/10.1007/s00383-026-06296-8.
10. Panzeri D, Laohawetwanit T, Akpinar R, et al. Assessing the diagnostic accuracy of ChatGPT-4 in the histopathological evaluation of liver fibrosis in MASH. Hepatol Commun. 2025;9(5):e0695. https://doi.org/10.1097/HC9.0000000000000695.
11. Naik SN, Forlano R, Manousou P, Goldin R, Angelini ED. Fibrosis severity scoring on Sirius red histology with multiple-instance deep learning. Biol Imaging. 2023;3:e17. https://doi.org/10.1017/S2633903X23000144.
12. Pan H, Zhang R, Zhou S, et al. A Transformer-Residual Network-Integrated Deep Learning Model for Digital Pathology-Based Grading and Risk Stratification of Metabolic Dysfunction-Associated Steatotic Liver Disease-Related Liver Fibrosis. Clin Transl Gastroenterol. 2026;17(7):e01017. https://doi.org/10.14309/ctg.0000000000001017.
13. Shabanian M, Taylor Z, Woods C, et al. Liver fibrosis classification on trichrome histology slides using weakly supervised learning in children and young adults. J Pathol Inf. 2025;16:100416. https://doi.org/10.1016/j.jpi.2024.100416.
14. Levy JJ, Azizgolshani N, Andersen MJ Jr, et al. A large-scale internal validation study of unsupervised virtual trichrome staining technologies on nonalcoholic steatohepatitis liver biopsies. Mod Pathol. 2021;34(4):808–22. https://doi.org/10.1038/s41379-020-00718-1.
15. Heinemann F, Gross P, Zeveleva S, et al. Deep learning-based quantification of NAFLD/NASH progression in human liver biopsies. Sci Rep. 2022;12:19236. https://doi.org/10.1038/s41598-022-23905-3.
16. Ratziu V, Francque S, Behling CA, et al. Artificial intelligence scoring of liver biopsies in a phase II trial of semaglutide in nonalcoholic steatohepatitis. Hepatology. 2024;80(1):173–85. https://doi.org/10.1097/HEP.0000000000000723.
17. Travis N, Billaut V, Pokkalla H, et al. Machine learning models accurately interpret liver histology and are associated with disease progression in patients with primary sclerosing cholangitis. J Hepatol. 2020;73(Suppl 1):S485–6. https://doi.org/10.1016/S0168-8278(20)31451-3.
18. Zhan H, Chen S, Gao F, et al. AutoFibroNet: a deep learning and multi-photon microscopy-derived automated network for liver fibrosis quantification in MAFLD. Aliment Pharmacol Ther. 2023;58(6):573–84. https://doi.org/10.1111/apt.17635.
19. Jana A, Qu H, Rattan P, Minacapelli CD, Rustgi V, Metaxas D. Deep Learning based NAS Score and Fibrosis Stage Prediction from CT and Pathology Data. In: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE). 2020. pp. 981–986. https://doi.org/10.1109/BIBE50027.2020.00166
20. Qu H, Minacapelli CD, Tait C, et al. Training of computational algorithms to predict NAFLD activity score and fibrosis stage from liver histopathology slides. Comput Methods Programs Biomed. 2021;207:106153. https://doi.org/10.1016/j.cmpb.2021.106153.
21. Yin C, Liu S, Lyu F et al. XFibrosis: Explicit Vessel-Fiber Modeling for Fibrosis Staging from Liver Pathology Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. pp. 11282–11291. https://doi.org/10.1109/CVPR52733.2024.01072
22. Klimov S, Farris AB, Chen Y, Chen H, Jiang Y. Classifying cirrhosis on H&E slides via a novel 2-step machine learning pipeline. Hepatology. 2019;70(Suppl 1):1240. Conference abstract/poster.
23. Abedin N, Hunyady P, Becker N, Bankov K, Zeuzem S, Schulze F, Wild PJ. Automated NAS scoring of Masson stained liver biopsies in NASH. Hepatology. 2020;72(Suppl 1):895A-896A. Conference abstract.
24. Naglah A, Khalifa F, El-Baz A, Gondim D. Conditional GANs based system for fibrosis detection and quantification in Hematoxylin and Eosin whole slide images. Med Image Anal. 2022;81:102537. https://doi.org/10.1016/j.media.2022.102537.
25. Wojciechowska M, Thing M, Hu Y, et al. Decoding fibrosis: transcriptomic and clinical insights via AI-derived collagen deposition phenotypes in MASLD. Hepatology. 2026. https://doi.org/10.1097/HEP.0000000000001811. Advance online publication.
26. Wojciechowska M, Malacrino S, Windell D, et al. Enhancing liver fibrosis measurement: deep learning and uncertainty analysis across multi-center cohorts. J Pathol Inf. 2026;21:100653. https://doi.org/10.1016/j.jpi.2026.100653.

27. Ercan C, Kordy K, Knuuttila A, et al. A deep-learning-based model for assessment of autoimmune hepatitis from histology: AI(H). Virchows Arch. 2024;485(6):1095–105. https://doi.org/10.1007/s00428-024-03841-5.
28. Mizuochi S, Kondo R, Kawamura S, et al. Deep-Learning-Aided Quantification of Steatohepatitis-Associated Pathological Findings in Liver Specimens. Hepatol Res. 2025;55(11):1462–70. https://doi.org/10.1111/hepr.70015.
29. Zou J, Arshad MR, Wang Z. Comparative Analysis of YOLO Architectures for Automated Detection of Liver Disease in Histopathological Images. In: ICBSP 2024 - Proceedings of the 2024 9th International Conference on Biomedical Imaging, Signal Processing. Association for Computing Machinery; 2025. pp. 33–38. https://doi.org/10.1145/3707172.3707177
30. Younossi ZM, Pokkalla H, Pethia K et al. Machine learning fibrosis models based on liver histology images accurately characterize the heterogeneity of cirrhosis due to nonalcoholic steatohepatitis (NASH). Hepatology. 2019;70(Suppl 1). Abstract 1718. Conference abstract/poster.
31. Mistry P, Stanford-Moore A, Egger R, et al. Characterizing the histologic implications of resmetirom-induced liver volume reduction using artificial intelligence-powered digital pathology. J Hepatol. 2023;78(Suppl 1):S790. https://doi.org/10.1016/S0168-8278(23)02229-8. Abstract OS-029.
32. Pokkalla H, Pethia K, Taylor-Weiner A, et al. Machine learning models identify novel histologic features predictive of clinical disease progression in patients with advanced fibrosis due to non-alcoholic steatohepatitis. J Hepatol. 2020;73(Suppl 1):S402. https://doi.org/10.1016/S0168-8278(20)31289-7. Abstract FRI003.
33. Taylor-Weiner A, Pokkalla H, Han L et al. Validation of a Machine Learning-Based Approach (DELTA Liver Fibrosis Score) for the Assessment of Histologic Response in Patients With Advanced Fibrosis Due to NASH. Hepatology. 2020;72(Suppl 1). Conference poster, AASLD The Liver Meeting 2020.
34. Panzeri D, Akpinar R, Inverso D, Di Tommaso L, Sironi L. AI-driven collagen segmentation and unsupervised feature extraction for liver fibrosis evaluation. Hepatology. 2024;80(Suppl 1):S1531. https://doi.org/10.1097/HEP.0000000000001077. Abstract 3586.
35. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527. https://doi.org/10.1136/bmj.h5527.
36. Wojciechowska M, Malacrino S, Garcia Martin N, Fehri H, Rittscher J. Early Detection of Liver Fibrosis Using Graph Convolutional Networks. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2021. Lect Notes Comput Sci. 2021;12908:217–26. https://doi.org/10.1007/978-3-030-87237-3_21.

Figures

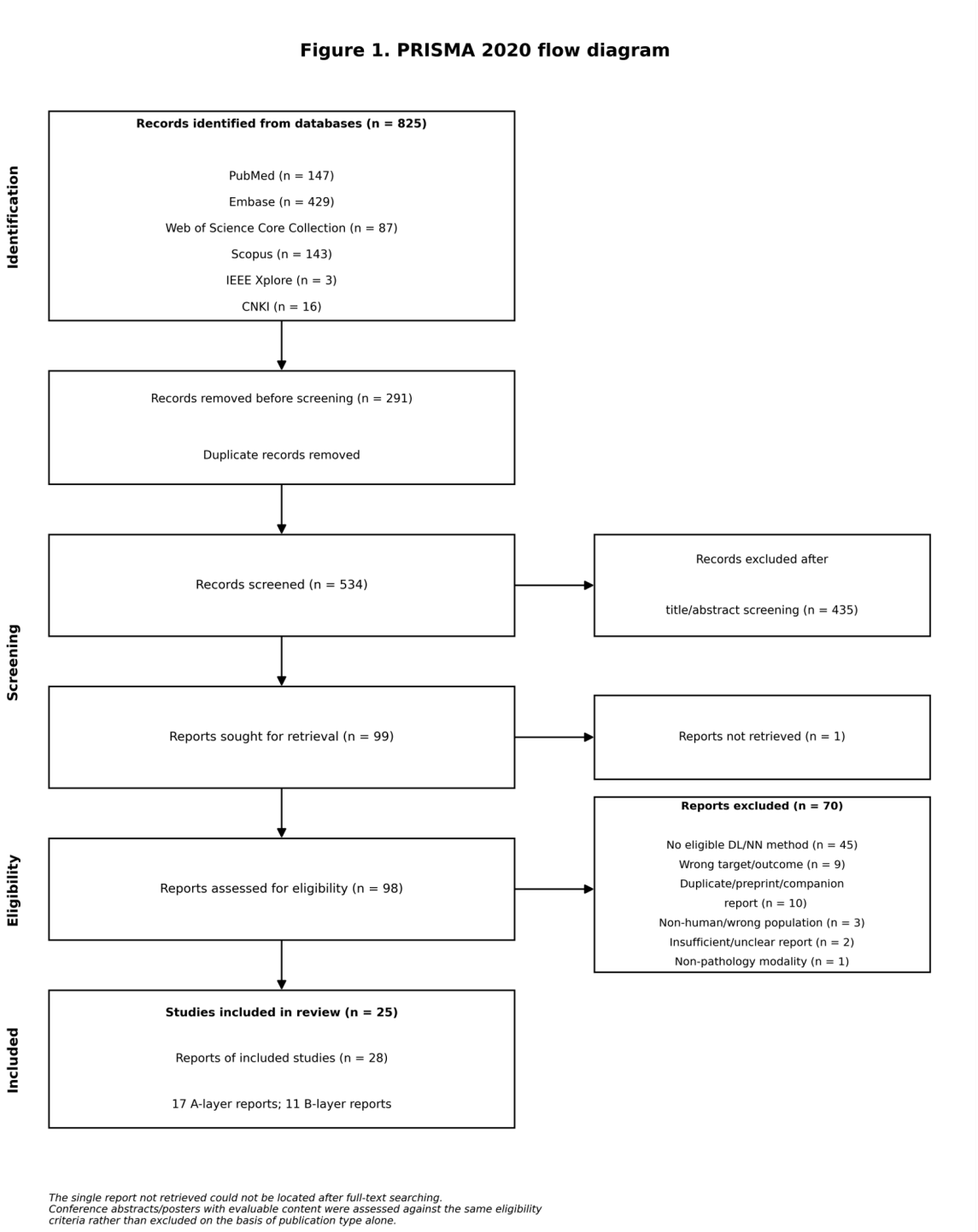


Figure 1

PRISMA 2020 flow diagram of study selection. The single report not retrieved could not be located after full-text searching. Conference abstracts and posters with evaluable content were assessed against the same eligibility criteria rather than excluded by publication type alone.

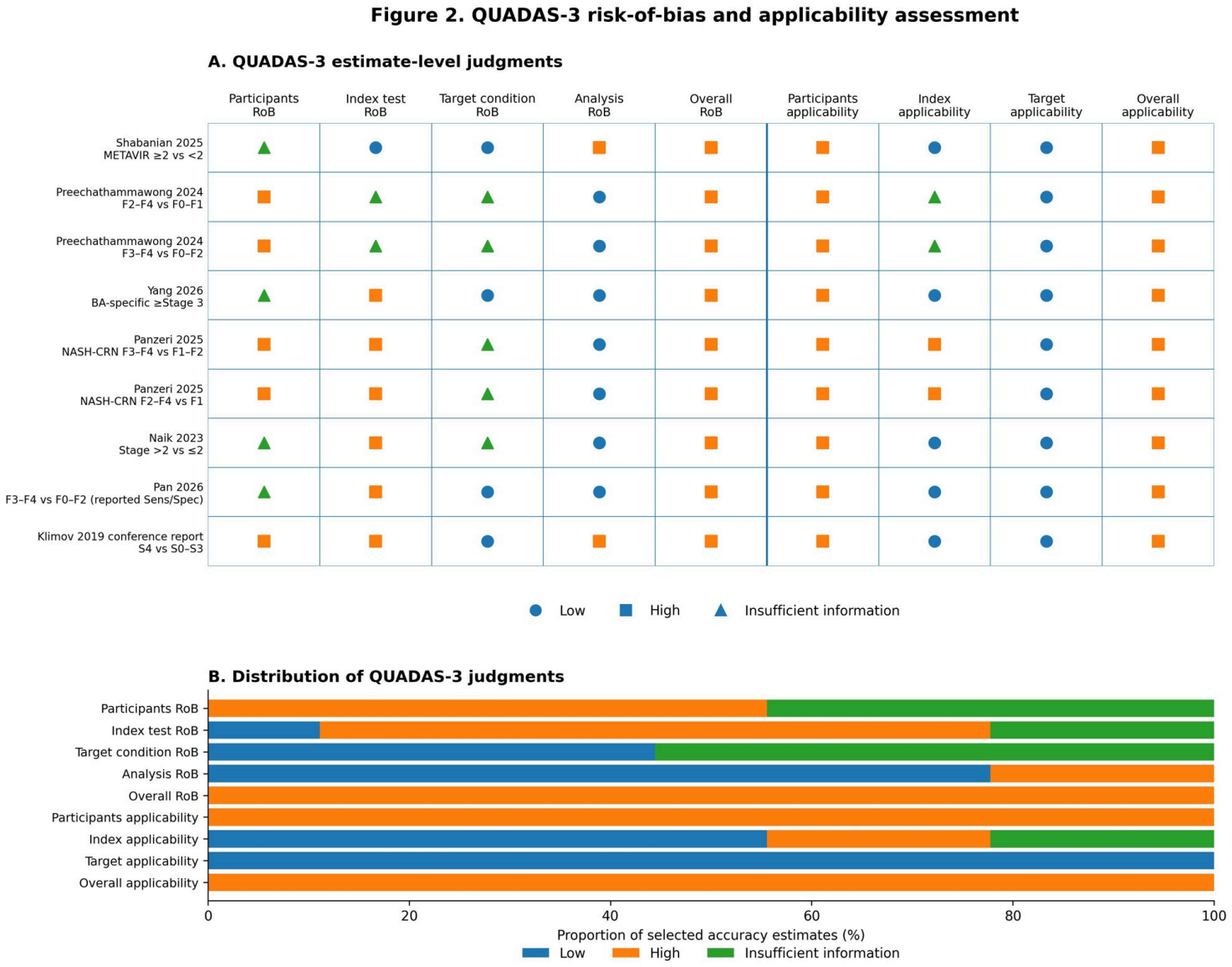
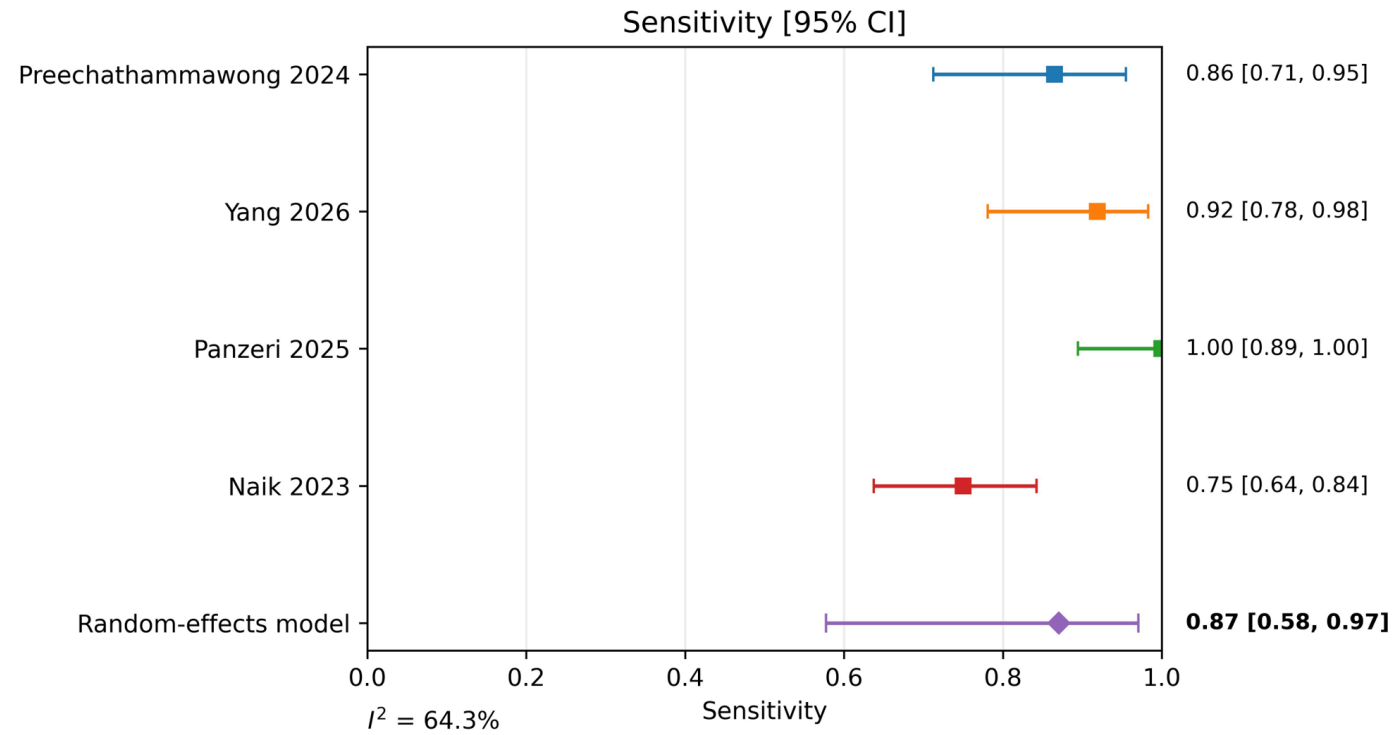


Figure 2

Risk-of-bias and applicability assessment of selected diagnostic-accuracy estimates using QUADAS-3. Panel A shows estimate-level judgments and Panel B summarizes the distribution of judgments across domains. Low, high, and insufficient-information categories are shown separately.



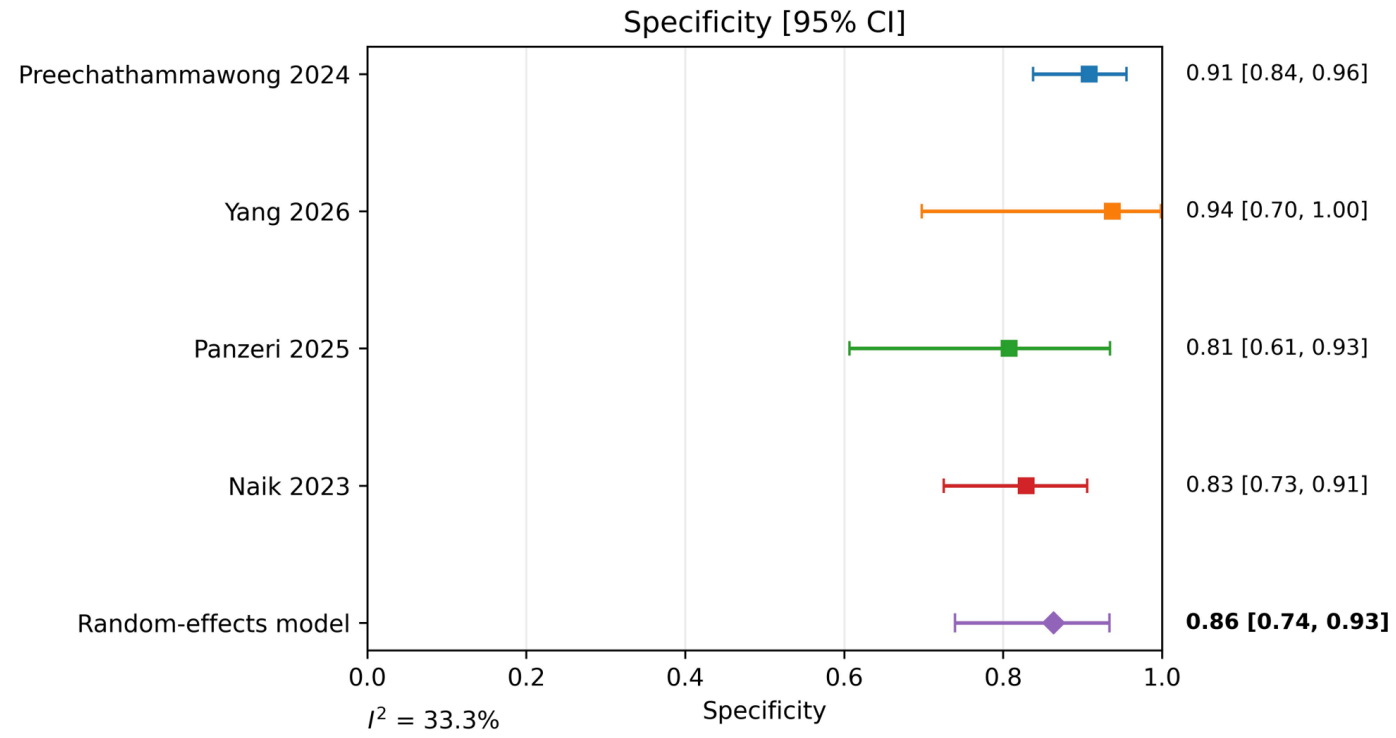


Figure 3

Forest plots of diagnostic performance for advanced liver fibrosis (F>=3 or an etiologically equivalent threshold). Panel A shows sensitivity and Panel B specificity for four independent study families with exact 2x2 data. Study-level proportions and exact binomial confidence intervals use uncorrected counts; pooled random-effects estimates use logit transformation, REML, and Knapp-Hartung confidence intervals. For transformed pooling only, a 0.5 correction was applied to all four cells of a study when any cell was zero.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1S2.xlsx
- SupplementaryTableS3.xlsx
- SupplementaryAppendixS1.docx
- SupplementaryFigureS2.tiff
- SupplementaryFigureS3.tiff
- SupplementaryFigureS1.tiff