Uncorrected Proof

Hepat Mon. January-December 2026; 26(1): e172843 https://doi.org/10.5812/hepatmon-172843

Published Online: 2026 October 31 **Review Article**

Artificial Intelligence in Viral Hepatitis Care: A Narrative Review and a Preliminary Clinical-Readiness Framework (CRAVE-Hep)

Abdolreza Babamahmoodi [1,*], Farhang Babamahmoodi [2]

[1] Research Ethics Board, Michael Garron Hospital Foundation, Toronto, M4C 3E7, Canada
[2] Antimicrobial Resistance Research Center, Mazandaran University of Medical Sciences, Qaemshahr, 4448526523, Iran

[*] ***Corresponding Author:*** School of Nursing, Faculty of Health, York University, Toronto, Canada. Email: rezababa@yorku.ca

Received: 15 June, 2026; **Revised:** 3 September, 2026; **Accepted:** 18 October, 2026

Abstract

Context: Viral hepatitis care increasingly involves artificial intelligence (AI), but predictive accuracy does not establish clinical readiness. We organize the evidence along the care cascade and propose Clinical-Readiness Appraisal of Validation and Equity for AI in Hepatitis, an unvalidated six-dimensional framework addressing validation, performance integrity, representativeness, transparency, governance, and implementation.

Evidence Acquisition: The initial publication window for this purposive narrative review was January 1, 2018, to January 31, 2026. Following peer review, searches of Scopus and Web of Science were repeated on September 2, 2026, returning 4,458 and 3,200 records, respectively. Supplementary searches that day retrieved 3,686 PubMed and 3,269 Europe PMC records. Source-specific date filters are documented separately. Six Google Scholar queries yielded 60 inspected entries (58 distinct linked records), from which four reports were added. The review cites 36 sources; retrieval counts are not exhaustive screening totals. SANRA informed reporting; TRIPOD+AI and PROBAST+AI informed illustrative appraisal.

Results: Applications of AI span case-finding, fibrosis staging, prognosis, treatment-response prediction, hepatocellular carcinoma surveillance, and language-based support. The evidence is predominantly retrospective, although prospective multicenter fibrosis studies exist. High discrimination commonly coexists with limited assessment of calibration, clinical utility, or workflow. Liver-specific retrieval-augmented language models remain proof-of-concept systems rather than validated clinical decision makers. Iran illustrates interoperability and implementation challenges relevant to resource-limited settings.

Conclusions: Technical promise should be distinguished from deployment readiness. None of the worked examples addresses all six CRAVE-Hep dimensions. Prospective validation, appropriate performance evaluation, equity, governance, and local workflow testing remain priorities. CRAVE-Hep itself requires consensus development, reliability testing, and validation before consequential use.

Keywords: Artificial Intelligence, Machine Learning, Hepatitis B, Hepatitis C, Liver Fibrosis, Hepatocellular Carcinoma, Clinical Readiness, Calibration, Decision-curve Analysis, External Validation, Implementation Science

1. Context

Viral hepatitis is among the leading infectious causes of death worldwide. The World Health Organization (WHO) estimated that viral hepatitis caused approximately 1.3 million deaths in 2022, with hepatitis B virus (HBV) and hepatitis C virus (HCV) accounting for the overwhelming majority. Chronic HBV affects more than 250 million people and remains a principal driver of cirrhosis and hepatocellular carcinoma (HCC) (1). Despite the availability of an effective HBV vaccine and curative direct-acting antiviral (DAA) regimens for HCV, substantial diagnostic and treatment gaps persist, and elimination targets remain distant in most regions.

Contemporary hepatitis management is increasingly individualized. The 2025 European Association for the Study of the Liver (EASL) Clinical Practice Guidelines on HBV, organized into ten thematic areas ranging from diagnostics to HCC surveillance, emphasize early diagnosis, risk stratification based on viral and host factors, and tailored antiviral therapy (1). This complexity—

Copyright © 2026, Babamahmoodi and Babamahmoodi. This open-access article is available under the Creative Commons Attribution 4.0 (CC BY 4.0) International License (https://creativecommons.org/licenses/by/4.0/), which allows for unrestricted use, distribution, and reproduction in any medium, provided that the original work is properly cited.

How to Cite: Babamahmoodi A, Babamahmoodi F. Artificial Intelligence in Viral Hepatitis Care: A Narrative Review and a Preliminary Clinical-Readiness Framework (CRAVE-Hep). Hepat Mon. 2026;26(1):e172843. doi: https://doi.org/10.5812/hepatmon-172843

including serological interpretation, fibrosis staging, treatment timing, integration of comorbidities (notably metabolic dysfunction-associated steatotic liver disease [MASLD]), and lifelong HCC surveillance—generates high-dimensional data well suited to computational analysis.

Artificial intelligence encompasses machine learning (ML), deep learning (DL; multilayer neural networks suited to imaging and unstructured data), and, more recently, large language models (LLMs). These approaches are being explored across infectious diseases for surveillance, diagnosis, and treatment optimization (2-4). In hepatology, a growing body of reviews describes artificial intelligence (AI) applications in viral hepatitis, fibrosis, cirrhosis, and HCC (5, 6).

Terminology. Because "AI," "ML," and "prediction model" are used inconsistently in this literature, we define these terms for the remainder of this review. "AI" is the umbrella term for the methods discussed. "ML model" and "prediction model" denote a model whose output is a probability or classification that informs an individual clinical decision; TRIPOD+AI, PROBAST+AI, and the calibration and decision-curve analysis (DCA) expectations in Section 3.8 apply to this category. "DL" and "LLM" are reserved for those specific method families. This terminology is used throughout, including in Tables 1 and 2.

Validation is described to the extent verified in the available primary source. Details for References 18 - 21 were checked against indexed primary abstracts; the full-text methods for data splitting were also checked for Reference 18. Unreported details were not inferred. Random splitting can permit leakage if correlated patient records cross partitions; it is not itself proof of leakage. Prospective diagnostic accuracy does not establish clinical-outcome benefit. Abbreviations: AASLD, Association for the Study of Liver Diseases; APRI, aminotransferase-to-platelet ratio index; AUC/AUROC, area under the receiver operating characteristic curve; CHB, chronic hepatitis B; CNN, convolutional neural network; CV, cross-validation; DAA, direct-acting antiviral; DCA, decision-curve analysis; DL, deep learning; DT, decision tree; FNN, feedforward neural network; GBM, gradient boosting machine; KNN, k-nearest neighbors; LLM, large language model; LR, logistic regression; ML, machine learning; NB, naïve Bayes; RAG, retrieval-augmented generation; RF, random forest; SMOTE, synthetic minority oversampling technique; SVM, support vector machine; SVR, sustained virologic response; VHA, Veterans Health Administration; 2D-SWE, two-dimensional shear-wave elastography.

Readiness reflects the authors' qualitative synthesis of validation type, study design, and external generalizability; it is a heuristic, not a formal grading instrument. Readiness-tier definitions (technically promising / clinically evaluable / ready for clinical deployment) are given in Section 3.10; by these definitions, no domain listed here currently reaches "ready for clinical deployment." AIH, autoimmune hepatitis; DCA, decision-curve analysis; DILI, drug-induced liver injury; electronic health records (EHR), electronic health record; HAV/HEV/HDV, hepatitis A/E/D virus; MASLD, metabolic dysfunction-associated steatotic liver disease.

Rationale and novelty. Previous reviews have surveyed AI in viral hepatitis and broader hepatology (5, 6, 22). Our contribution is to organize the evidence along the care cascade (screen, diagnose, stage, prognosticate, treat, and surveil) and propose CRAVE-Hep (Clinical-Readiness Appraisal of Validation and Equity for AI in Hepatitis). This six-dimensional conceptual framework integrates AI-specific reporting and risk-of-bias standards (TRIPOD+AI, PROBAST+AI, and CONSORT-AI/SPIRIT-AI) with early-evaluation guidance (DECIDE-AI) (23-27). Iran provides an illustrative setting for implementation in contexts with high disease burdens and limited resources, given its chronic HBV/HCV burden, national information systems (SIB and SEPAS), WHO-aligned 2030 elimination strategy, and the authors' direct clinical and health-system experience. Section 3.9 distinguishes generalizable implementation barriers from Iran-specific system constraints. CRAVE-Hep is a proposed, unvalidated framework for structuring discussion, rather than a scoring instrument or evidence of deployability.

2. Evidence Acquisition

This purposive narrative review was informed by SANRA (28). Initial searches of PubMed/MEDLINE, Scopus, and Web of Science focused on English-language, peer-reviewed literature, with an intended publication window from January 1, 2018, to January 31, 2026. Following the reviewers' request, the Scopus and Web of Science searches were repeated on September 2, 2026, returning 4,458 and 3,200 records, respectively. These repeat-search details were reported by the authors. Supplementary searches of PubMed and Europe PMC executed on the same date returned 3,686 and 3,269 records. Table 3 and Table S2 in the Supplementary File distinguish the initial search from the revision-stage searches and reproduce database-specific strategies and filters. The Scopus/Web of Science strategies use calendar years 2018 - 2026, whereas the executed PubMed/Europe PMC queries retained January 31, 2026, as their publication cutoff. These source totals therefore describe different retrieval contexts and are not summed or treated as a deduplicated, exhaustively screened evidence set. Six supplementary Google Scholar queries were inspected at positions 1 - 10 in the relevance-ranked

Table 1. Representative Studies of AI Across the Viral Hepatitis Care Cascade, with Design and Validation Detail

Study / Domain	AI Model & Data	Sample (n)	Design / Validation	Performance	Main Limitation
Fibrosis staging, CHB (7)	Transfer-learning CNN; grayscale + elastogram ultrasound	466 (401 CHB)	Retrospective; internal random split (no external)	AUC ~0.93 - 0.95 (significant/advanced fibrosis)	Single setting; random split risks leakage; no external validation
Fibrosis staging, liver–spleen (8)	Radiomics + transfer learning; 2D-SWE	198	Retrospective; internal validation (no external)	AUC up to ~0.98 - 1.00 best tasks; ~0.57 - 0.79 some tasks	Small sample; optimistic best-case AUCs; no calibration/DCA
Histopathology fibrosis scoring (9)	CNN (deep learning); biopsy images	Multi-source image dataset	Retrospective; internal validation	AUROC ~0.85 - 0.95 vs expert reference	Generalizability; reference-standard variability
HBV/HCV case-finding (10)	Random forest, SVM, NB, KNN; clinical survey data	Prospective screening cohort (Romania)	Prospective data; internal validation	Predicts serostatus; stronger for HCV	Net benefit over standard risk-based screening not yet shown
DAA failure, HCV (11)	Elastic net, RF, GBM, FNN; HCV-TARGET registry	6,525	Retrospective; train/validation split (single registry)	C-statistic ~0.72 - 0.75; modest gain over LR	No external/temporal validation; failure now low-yield need
DAA response from viral genome (12)	ML on whole-genome HCV variants	173	Retrospective; random train/validation split	Preliminary predictive model	Very small sample; leakage-prone split; exploratory
SVR, comorbid HCV (13)	XGBoost, RF, SVM, DT; clinical data	221	Retrospective; CV + SMOTE; no external	Predicts SVR (AUROC ~0.76)	Single-center; SMOTE artifacts; no calibration reported
HCV time-to-cure / duration (14, 15)	ML + viral-kinetic modeling	266 (kinetics study)	Retrospective; modeling, internal	Individualized duration / cure-time estimates	Modeling assumptions; no prospective trial
HCC detection / prognosis (16, 17)	DL on CT/MRI; multimodal models	Variable across studies	Mostly retrospective; mostly internal	Lesion detection, risk prediction, prognosis	Protocol heterogeneity; limited external validation
Konerman et al. (18); HCV progression	Boosted survival trees; longitudinal clinical data	72,683	Retrospective VHA cohort; 70:30 random splits repeated 30 times	Concordance 0.774; longitudinal Cox comparator 0.764	96.8% male; APRI-defined cirrhosis; no prospective external test
Wang et al. (19); CHB fibrosis	Deep-learning radiomics; 2D-SWE; biopsy reference	398; 1,990 images	Prospective, 12 hospitals; robustness across training cohorts assessed	AUC: F4 0.97; ≥F3 0.98; ≥F2 0.85	Diagnostic accuracy; clinical-outcome benefit not established
Wong et al. (20); HCC risk	Five ML methods; territory-wide clinical data, Hong Kong	124,006	Retrospective; internal random 7:3 split (86,804 / 37,202)	Validation AUROC: ridge 0.844; RF 0.837	Internal validation; model-guided surveillance benefit not established
Ge et al. (21); liver-specific LLM	LiVersa RAG; 30 AASLD guidance documents	10 questions per round; 15 expert raters in round 2	Proof of concept: knowledge test, then expert comparison with other LLMs	10/10 correct in round 1; higher accuracy but lower safety/comprehensiveness than ChatGPT-4 in round 2	Small question sets; expert ratings, not patient outcomes; refinement needed

Abbreviations: AI, artificial intelligence; HBV, hepatitis B virus; HCV, hepatitis C virus; HCC, hepatocellular carcinoma; LLM, large language model; CHB, chronic hepatitis B; CNN, convolutional neural network; AUC, area under the curve; DCA, decision-curve analysis; AUROC, area under the receiver operating characteristic curve; SVM, support vector machine; NB, naïve Bayes; KNN, k-nearest neighbors; DAA, direct-acting antiviral; RF, random forest; GBM, gradient boosting machine; FNN, feedforward neural network; LR, logistic regression; SVR, sustained virologic response; DT, decision tree; SMOTE, synthetic minority oversampling technique; VHA, Veterans Health Administration; RAG, retrieval-augmented generation; AASLD, Association for the Study of Liver Diseases.

results, excluding patents and citation-only entries: 60 entries represented 58 distinct linked records. Four reports were purposively retained (18-21); dates and added claims were checked against primary abstracts, and full-text methods were checked where accessible. Scholar estimates were not treated as exact yields. Reference 13 is a targeted update published online on April 13, 2026, outside the initial publication window.

Retrieval totals refer to the revision-stage searches, not reconstructed initial yields or counts of included studies. Database coverage overlaps, and the documented publication-date filters differ across sources. Therefore, the totals are reported separately. Google Scholar inspection was bounded; the citation total is not an exhaustive count of included studies.

Consistent with narrative methodology, selection was purposive rather than exhaustive. Priority was given to original studies reporting quantitative performance metrics, large or multicenter cohorts, recent systematic and scoping reviews, and authoritative guideline documents. When multiple similar studies addressed the same domain, we prioritized the most recent, largest, or most transparent study, citing others only when they added a distinct point. We did not perform a formal systematic risk-of-bias assessment; instead, prediction studies were appraised conceptually against the signaling domains of TRIPOD+AI (reporting) and PROBAST+AI (risk of bias and applicability) (23, 24). A focused, illustrative appraisal of five representative prediction studies against these domains is provided in Table S1 in the Supplementary File, summarizing themes related to validation rigor and

Table 2. Clinical-Readiness Grading of AI Applications in Viral Hepatitis

Application Domain	Maturity of Evidence	Readiness Tier	Key Gap to Address
Image-based fibrosis staging	Predominantly retrospective; prospective multicenter evidence also exists (19)	Technically promising; not yet clinically evaluable at scale	Transportable calibration/clinical utility and prospective workflow or outcome evaluation
EHR-based case-finding	Low–moderate	Technically promising	Demonstrated net benefit versus standard risk-based screening
Prognosis / progression risk	Large retrospective cohorts; repeated internal validation (18)	Technically promising	Outcome-linked prospective validation
Treatment-response / duration	Moderate (registry-based models)	Technically promising, approaching clinically evaluable for one registry model	External validation; demonstrated impact on clinical decisions
HCC surveillance / characterization	Large retrospective cohorts; internal validation (20)	Technically promising	Standardized protocols; external validation
Large language models	Proof-of-concept evaluations, including liver-specific RAG (21)	Exploratory (risk-stratified by use case; Section 3.6)	Hepatitis-specific, category-appropriate validated evaluation
Hepatitis D/E/A virus (HDV/HEV/HAV), autoimmune hepatitis (AIH), drug-induced liver injury (DILI), MASLD overlap	Very low	Exploratory	Foundational studies needed

Abbreviations: AI, artificial intelligence; HCC, hepatocellular carcinoma; RAG, retrieval-augmented generation; HDV, hepatitis D virus; MASLD, metabolic dysfunction-associated steatotic liver disease.

Table 3. Summary of the Literature Search Strategy

Element	Detail
Databases	Initial sources: PubMed/MEDLINE, Scopus, and Web of Science. Revision-stage repeat searches: Scopus and Web of Science. Supplementary sources: PubMed/MEDLINE, Europe PMC, and Google Scholar.
Search period	Initial publication window: January 1, 2018 to January 31, 2026. Repeat Scopus and Web of Science searches following peer review: September 2, 2026. Supplementary PubMed, Europe PMC, and Google Scholar searches: September 2, 2026. The repeat Scopus/Web of Science strategies specify calendar years 2018 - 2026; the executed PubMed/Europe PMC strategies retained the January cutoff. S2 documents each filter.
AI concept terms	“artificial intelligence,” “machine learning,” “deep learning,” “neural network,” “radiomics,” “large language model”
Disease concept terms	"hepatitis B," "hepatitis C," "viral hepatitis," "liver fibrosis," "cirrhosis," "hepatocellular carcinoma"
Representative full search string (PubMed/MEDLINE)	("artificial intelligence"[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab] OR "neural network"[tiab] OR radiomics[tiab] OR "large language model*"[tiab]) AND ("hepatitis B"[tiab] OR "hepatitis C"[tiab] OR "viral hepatitis"[tiab] OR "liver fibrosis"[tiab] OR cirrhosis[tiab] OR "hepatocellular carcinoma"[tiab]) AND ("2018/01/01"[dp]: "2026/01/31"[dp]) AND English[lang] AND journal article[pt]
Retrieval totals and inspected results	Repeat searches executed on September 2, 2026: Scopus 4,458 records; Web of Science 3,200 records. Supplementary searches on the same date: PubMed 3,686; Europe PMC 3,269. Google Scholar: 60 inspected entries across six queries, representing 58 distinct linked records. These are source-specific retrieval/inspection figures, not a combined deduplicated or exhaustive screening total. See S2 for provenance.
Sources retained for citation	36 cited sources: 28 from the initial selection (References 1 - 3, 5 - 12, 14 - 17, 22 - 26, 28, 29, 31 - 36), four targeted additions (References 4, 13, 27, 30), and four from the documented supplementary search (References 18 - 21). Reference 13 is a post-cutoff targeted update. Citation totals include background, guidance, and methods and are not an exhaustive count of eligible empirical studies.
Inclusion focus	Original studies with performance metrics; multicenter or large cohorts; recent reviews; current guidelines; clinical (not purely technical) relevance
Exclusion focus	Non-English; conference abstracts without full data; purely algorithmic papers lacking clinical endpoints; duplicate reports
Quality consideration	SANRA (review reporting); TRIPOD+AI and PROBAST+AI (prediction studies); CONSORT-AI/SPIRIT-AI (interventional)

Abbreviation: AI, artificial intelligence.

risk of bias discussed in Sections 3.2 - 3.5 and 3.8. Both authors contributed to manuscript preparation, including discussion of study selection and interpretation. Differences of opinion were resolved through discussion and review of relevant evidence and scientific reporting guidance; independent duplicate screening is not claimed. Performance is reported as in the primary studies; when results varied widely across subtasks, we present the range rather than only the most favorable value.

Development of the CRAVE-Hep framework. The six CRAVE-Hep dimensions were derived by mapping recurrent limitations identified across the reviewed studies to established AI reporting and appraisal standards (TRIPOD+AI, PROBAST+AI, and CONSORT-AI/SPIRIT-AI), early-stage clinical-evaluation guidance (DECIDE-AI), and basic implementation-science considerations (23-27). Table 4 documents this derivation, mapping each dimension to the recurring evidence gap that motivated it and the standard on which it draws. It also provides the rationale

Table 4. Derivation of the CRAVE-Hep Dimensions

Dimension	Recurring Evidence Gaps in Reviewed Literature	Standard(s) Drawn Upon	Rationale for Inclusion
D1 – Validation rigor	Most studies use single internal random splits; leakage-prone designs are common (Sections 3.2, 3.8; Table 1)	TRIPOD+AI (validation types); PROBAST+AI (risk of bias)	External/prospective validation is the single most consistent gap across care-cascade domains
D2 – Performance integrity	Discrimination (AUC/C-statistic) is reported almost universally; calibration and DCA are rarely reported (Section 3.8)	TRIPOD+AI reporting items; DECIDE-AI early-stage evaluation	Distinguishes models that merely discriminate from those shown to be clinically useful
D3 – Data representativeness & bias	Small, single-center, homogeneous cohorts; subgroup/equity reporting largely absent (Sections 3.2 - 3.5, 3.9)	PROBAST+AI (participants domain); implementation-science equity principles	Generalizability and fairness cannot be assumed from aggregate performance alone
D4 – Transparency & interpretability	Incomplete methodological reporting; opaque ("black-box") models (throughout)	TRIPOD+AI reporting checklist	Reproducibility and clinical trust require transparent reporting and some account of model behavior
D5 – Data governance & privacy	Data-sharing, consent, and federation arrangements are rarely addressed (Section 3.9)	General data-governance principles; federated-learning literature (29)	Cross-institution or cross-border validation is not feasible without explicit governance
D6 – Implementation fit	Interoperability, workflow integration, and deployment status are almost never reported (Sections 3.9, 3.10)	DECIDE-AI (early clinical-evaluation domains); implementation science	Distinguishes deployable tools from those that are technically validated but unimplemented

Abbreviations: AI, artificial intelligence; DCA, decision-curve analysis.

Table 5. The CRAVE-Hep Framework: Six Dimensions of Clinical Readiness for AI Tools in Viral Hepatitis, Each Described as a Progression from Lower to Higher Maturity

Dimension	Lower Maturity	Intermediate	Higher Maturity
D1 – Validation rigor	Internal random split only; leakage risk	Temporal or single external validation	Multi-site external + prospective validation
D2 – Performance integrity	Discrimination (AUC/C-statistic) only	Calibration reported (eg, Brier, plots)	Demonstrated net benefit by decision-curve analysis
D3 – Data representativeness & bias	Small single-center; no equity reporting	Multicenter but limited diversity	Large, diverse, multicenter with explicit equity reporting
D4 – Transparency & interpretability	Black-box; incomplete reporting	Partial TRIPOD+AI adherence	Full TRIPOD+AI reporting with interpretability
D5 – Data governance & privacy	Unspecified data handling	Governance described; centralized data	Explicit governance; federated/privacy-preserving training
D6 – Implementation fit	Purely theoretical / offline	Pilot in one workflow or setting	Interoperable (FHIR/OMOP) + workflow-integrated in target setting

Abbreviation: AI, artificial intelligence.

for excluding several other candidate dimensions (cost-effectiveness, regulatory status, end-user acceptability, and model drift) from the core six; these are discussed further as a second-tier extension in Section 3.11. Each dimension is described as a qualitative progression from lower to higher maturity rather than a numeric score. We use "framework" consistently for CRAVE-Hep: it structures qualitative judgment and is not a validated scoring checklist. It has not undergone *Delphi consensus*, expert-panel review, interrater reliability testing, or pilot evaluation. The illustrative judgments reflect joint interpretation rather than independent duplicate ratings. Future work should assess interrater reliability, construct validity, and prospective usability before CRAVE-Hep is used for consequential decisions.

Cost-effectiveness, regulatory status, end-user acceptability, and model drift were named in the reviewed literature as important considerations but were not adopted as core CRAVE-Hep dimensions because (a) too few reviewed studies reported them to identify a consistent, generalizable gap pattern comparable to that of the six dimensions above; (b) each is substantially downstream of and contingent on prior satisfaction of the six dimensions; and (c) each already has an established, dedicated evaluation literature (health-economic evaluation, regulatory science, human-factors/implementation science, and post-deployment monitoring, respectively) that CRAVE-Hep does not attempt to replace. We flag them as a second-tier extension for future framework development (Section 3.11).

3. Results

We present the evidence along the care cascade: screening and diagnosis (Section 3.1); noninvasive fibrosis staging (Section 3.2); prognosis and risk stratification (Section 3.3); treatment-response prediction and personalization (Section 3.4); HCC surveillance, including post-cure risk (Section 3.5); large language models (Section 3.6); the less-studied hepatitis spectrum (hepatitis D, E, and A, and overlaps; Section 3.7); methodological rigor beyond accuracy (calibration, decision-curve analysis, and data leakage; Section 3.8); and ethics, privacy-preserving learning, and implementation (Section 3.9). We then consolidate these findings in the CRAVE-Hep framework (Section 3.10), describe its limitations (Section 3.11), and propose a validation roadmap for the framework itself (Section 3.12). Table 4 documents the framework's derivation; Tables 1 and 2 summarize representative studies and grade readiness by domain; Tables 5 and 6 define the framework and illustrate its application.

A tool's profile is interpreted across all six dimensions, with particular attention to its lowest (rate-limiting) dimension because deployability is constrained by the weakest dimension. The framework deliberately avoids a composite numeric score. The operational anchors are as follows: “multicenter” indicates data drawn from ≥ 3 independent clinical sites; “large” indicates a validation sample sufficient to yield stable estimates given the outcome's event rate (commonly several hundred to > 1,000 patients for imaging-based staging tasks); “diverse” indicates representation across ≥ 2 countries, health systems, or major demographic or etiological subgroups; and “partial TRIPOD+AI adherence” indicates that some, but not all, core reporting items (eg, development but not validation, or performance but not calibration) are reported. These anchors are proposed to aid consistent interpretation; alternatively, readers may treat all descriptors as illustrative only. Both the anchors and the descriptors require empirical calibration through the *Delphi* process described in Section 3.12. Abbreviations: AUC, area under the curve; FHIR, Fast Healthcare Interoperability Resources; OMOP, Observational Medical Outcomes Partnership common data model; TRIPOD+AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis – artificial intelligence extension.

3.1. Screening and Diagnosis

Machine learning models based on routine clinical and laboratory variables can support case-finding for undiagnosed infections. In a prospective primary care screening cohort, supervised ML models (random forest, support vector machine, naïve Bayes, and k-nearest neighbors) predicted HBV and HCV serostatus from structured data, with generally stronger discrimination for HCV than for HBV (10). When applied to EHRs at scale, such models could prioritize individuals for confirmatory testing, potentially offering value where universal screening is not feasible. However, their net clinical benefit over established risk-based screening has not yet been demonstrated prospectively.

3.2. Noninvasive Fibrosis Staging

Fibrosis stage determines prognosis, surveillance, and treatment in chronic hepatitis. Because biopsy is invasive and prone to sampling error, noninvasive AI methods—especially DL applied to ultrasound and shear-wave elastography—have been widely studied (6). A transfer-learning model combining grayscale and elastogram ultrasound in patients predominantly with chronic hepatitis B reported areas under the receiver operating characteristic curve (AUCs) of roughly 0.93 - 0.95 for significant and advanced fibrosis, exceeding liver stiffness measurement and serum indices alone (7). A 2025 model integrating liver–spleen two-dimensional shear-wave elastography (2D-SWE) reported very high validation AUCs (up to ~0.98 - 1.00) for advanced disease and for etiology discrimination (8). These favorable figures, however, must be considered alongside lower-performing subtasks in the same body of work, for which validation AUCs of approximately 0.57 - 0.79 were reported for certain stages or comparators (8); therefore, headline values should not be generalized across all tasks or populations. A recent scoping review similarly found that hepatology imaging models cluster around AUC 0.80 - 0.95 and are mostly retrospective and single-center (6). Automated DL histopathology scoring reproduces expert-level staging, with the area under the receiver operating characteristic curve (AUROC) commonly ranging from 0.85 - 0.95 (9). Prospective evidence is not absent: Wang et al. enrolled patients across 12 hospitals and analyzed 398 patients with chronic HBV, reporting fibrosis-stage AUCs of 0.85 - 0.98 for deep-learning elastography (19). Prospective diagnostic accuracy alone does not demonstrate improved clinical outcomes.

Interpretation. Fibrosis-staging evidence is predominantly retrospective, but the prospective multicenter example (19) precludes characterizing the entire field as based only on internal validation. High discrimination alone remains insufficient for deployment without population-appropriate calibration, clinical-utility assessment, and workflow evaluation (Table 2).

3.3. Prognosis and Risk Stratification

Table 6. Qualitative Worked Illustration of CRAVE-Hep on Representative Tools [a]

Representative Tool	D1 Validation	D2 Performance	D3 Data & equity	D4 Transparency	D5 Governance	D6 Deployment
Multimodal US fibrosis staging (7)	↗ internal only	– no calibration/DCA	– single-center	↗ partial reporting	– unspecified	– not deployed
Liver–spleen 2D-SWE staging (8)	↗ internal only	– discrimination only	– small single-center	↗ partial reporting	– unspecified	– not deployed
Histopathology fibrosis CNN (9)	↗ internal	↗ vs expert reference	↗ multi-source images	✓ fuller reporting	↗ described	– no workflow
EHR HBV/HCV case-finding (10)	✓ prospective data	– no net benefit	✓ prospective cohort	↗ partial	↗ described	↗ EHR-based, untested
DAA-failure registry model (11)	↗ large but internal	↗ some calibration	✓ multicenter registry	✓ fuller reporting	↗ described	– not deployed
Post-SVR HCC risk (emerging)	– nascent	– nascent	– nascent	– nascent	– nascent	– nascent

[a] Symbols are interpretive summaries of the primary studies, not validated ratings, and would require independent corroboration by future appraisers. ✓ = dimension's core question substantively addressed; ↗ = partially addressed; – = unaddressed or not reported. Dimensions D1–D6 are defined in Table 5; the framework deliberately avoids numeric scores and overall rankings pending validation. Abbreviations: CNN, convolutional neural network; DAA, direct-acting antiviral; DCA, decision-curve analysis; EHR, electronic health record; HCC, hepatocellular carcinoma; SVR, sustained virologic response; US, ultrasound; 2D-SWE, two-dimensional shear-wave elastography.

Abbreviations: HBV, hepatitis B virus; HCV, hepatitis C virus.

ML can integrate longitudinal laboratory values, virologic markers, elastography, and demographic characteristics to stratify the risk of progression to cirrhosis, decompensation, or HCC, capturing nonlinear interactions beyond those represented by fixed-coefficient scores (5). The 2025 EASL emphasis on biomarker-led HBV care (quantitative hepatitis B surface antigen (HBsAg), HBcrAg, HBV RNA) provides a natural substrate for models that refine decisions about treatment initiation, surveillance, and safe discontinuation of nucleos(t)ide analogues (1). As elsewhere, prognostic value depends on dataset representativeness and prospective evidence that predicted risk changes management and improves outcomes—evidence that is currently sparse. In 72,683 untreated veterans with HCV, a longitudinal boosted-survival-tree model predicted cirrhosis, defined by the aminotransferase-to-platelet ratio index, with concordance of 0.774 (18). Evaluation used repeated random 70:30 splits, not prospective external validation; the predominantly male cohort also limits generalizability.

3.4. Treatment-Response Prediction and Personalization

Early ML work in HCV focused on predicting DAA treatment failure. Using the HCV-TARGET registry, investigators developed elastic net, random forest, gradient boosting, and feedforward neural-network models that showed modestly better discrimination (C-statistic, approximately 0.72 - 0.75) than logistic regression (11). With modern pan-genotypic regimens achieving sustained virologic response (SVR) above 95%, predicting DAA failure is no longer a pressing need for most patients; its value is now largely confined to difficult-to-treat subgroups. Efforts to apply AI may therefore be more valuable when directed toward two unresolved questions.

Personalizing treatment duration via viral kinetics. Rather than predicting binary treatment failure, ML models based on mathematical models of HCV viral kinetics can estimate individualized time to cure from sparse viral-load measurements, supporting response-guided shortening of therapy to reduce cost and toxicity (14, 15). This pharmacokinetic and viral-kinetic approach aligns AI with the need to improve efficiency, particularly in resource-limited programs.

Predicting HCC risk after viral clearance. A more pressing unmet need is risk stratification after SVR or during long-term HBV suppression: patients who clear HCV or achieve durable HBV control retain residual HCC risk, and identifying those who warrant continued surveillance is a central question in contemporary practice. ML models that integrate post-treatment laboratory trajectories, elastography, and host factors are well suited to this longitudinal problem and are conceptually closer to current guideline priorities than failure prediction. For HBV, the 2025 EASL shift to biomarker-led care similarly provides a basis for models that predict who can safely discontinue nucleos(t)ide analogues without relapse (1). A complementary line of work—using ML to analyze whole-genome HCV variants—illustrates how molecular data can inform outcome models (12), and a real-world comorbid HCV cohort from Pakistan demonstrates the feasibility of routine-data models in high-burden settings (13). Across all of these applications, prospective external validation remains a prerequisite for clinical adoption.

3.5. Hepatocellular Carcinoma Surveillance, Including Post-Cure Risk

Because chronic hepatitis is the dominant cause of most cases of HCC, surveillance is inseparable from hepatitis care. AI has been studied across the HCC

continuum, including risk prediction, lesion detection and characterization on CT and MRI, histopathological classification, and prognostication (16, 17). Related literature applies AI more broadly to cirrhosis complications, including variceal-bleeding risk prediction, reinforcing that HCC surveillance is part of a wider cirrhosis-AI literature (30). A clinically important and growing application is predicting HCC risk in patients who have achieved HCV cure or durable HBV suppression, among whom residual risk persists and rational, individualized surveillance remains an open guideline question. Reviews of AI in multidisciplinary HCC evaluation highlight opportunities to integrate radiology, pathology, and clinical data while cautioning that protocol heterogeneity, limited external validation, and model opacity constrain deployment (16). As with fibrosis staging, the credible near-term role is to augment expert reading and triage rather than provide autonomous diagnosis. Wong et al. analyzed 124,006 patients with chronic viral hepatitis; ridge regression retained an AUROC of 0.844 in the internal validation cohort (20). This supports risk-model development but does not prove that model-guided surveillance improves outcomes.

3.6. Large Language Models and Generative AI

Large language models could support several distinct tasks that differ substantially in clinical risk and evidentiary requirements. We distinguish four categories: 1) administrative and documentation support, including drafting notes, summarizing encounters, and structuring free text, carries low direct risk because a clinician typically reviews the output first; 2) patient-facing education, including answering patient questions or generating lay material, carries higher risk because patients may act directly on unverified information; 3) clinician-facing guideline support, including summarizing or retrieving recommendations, carries an intermediate level of risk; and 4) clinical decision support, in which output directly informs a diagnostic or treatment decision, carries the highest risk and, in our view, should be evaluated to the same standard as any other clinical prediction model.

In hepatitis, direct evidence remains limited and preliminary across all four categories, consisting mostly of feasibility studies or general question-answering rather than validated, disease-specific performance. Large language models also introduce particular risks: they can generate hallucinations, are sensitive to prompt phrasing, and can propagate training-data bias. We consider categories 2 and 4 to require prospective, disease-specific evaluation before routine use, analogous to other prediction models. Categories 1 and 3 warrant less intensive but still explicit evaluation (eg, accuracy audits of summarization output), given their lower but nonzero risk. Until such evaluation is available, outputs require expert verification, and clinical use should remain investigational.

Retrieval-augmented generation. A common mitigation for hallucinations is retrieval-augmented generation (RAG), which constrains the model's answers to a curated, trustworthy knowledge base, such as EASL or the Association for the Study of Liver Diseases (AASLD) guidelines, rather than relying solely on parametric memory (31). Retrieval-augmented generation is a safety-mitigation strategy, not a substitute for formal evaluation: it does not itself establish clinical correctness, calibration, or applicability to a local population. Tools grounded in RAG therefore still require the prospective, disease-specific evaluation described above before clinical use. One application relevant to many settings, including Iran, is natural-language processing (NLP) to extract structured information from free-text pathology and endoscopy reports, where RAG-grounded models could standardize descriptions of fibrosis, steatosis, and lesions. Such applications should be evaluated prospectively, with human review, before clinical use. A concrete example is LiVersa, grounded in 30 AASLD guidance documents: hepatologists rated its answers as more accurate but less comprehensive and safe than ChatGPT-4 outputs (21). This proof-of-concept comparison reinforces the need for evaluation beyond answer accuracy.

3.7. The Less-Studied Hepatitis Spectrum

Most AI research focuses on HBV, HCV, fibrosis, and HCC, leaving important gaps. Hepatitis D virus (HDV) coinfection produces the most severe form of chronic viral hepatitis and is frequently undetected (1), yet AI-based detection and risk modeling for HDV are scarce. Hepatitis E and A are similarly underrepresented, as are autoimmune hepatitis, drug-induced liver injury, and the increasingly common overlap of viral hepatitis with MASLD. We restrict this review's scope to the better-evidenced viral-hepatitis applications and identify these adjacent areas as priorities for future research rather than established application domains.

3.8. Beyond Accuracy: Calibration, Decision-Curve Analysis, and Data Leakage

A recurring weakness across the studies reviewed is the reliance on discrimination metrics (AUC or the C-statistic) as the sole indicators of model quality. This can be misleading for clinical deployment, especially for the imbalanced outcomes typical of hepatology: advanced fibrosis, HCC, or treatment failure are often rare, so a high AUC may coexist with poor clinical usefulness. Two

complementary evaluations are therefore essential. Calibration—agreement between predicted probabilities and observed event rates, summarized by calibration plots and the Brier score—determines whether risk estimates can be trusted for individual decisions; a model with good discrimination can still be poorly calibrated. Decision-curve analysis, which quantifies net clinical benefit across threshold probabilities, indicates whether acting on a model improves decisions relative to treating all or none. An AUC of 0.98, as reported in some fibrosis-staging studies (Table 1), alone does not establish that a model is calibrated, clinically useful, or free of bias.

Scope of this expectation. These expectations do not apply uniformly across all AI applications. Calibration and DCA are most directly relevant to prediction and risk-stratification models, such as fibrosis-stage classifiers, prognosis models, and post-cure HCC-risk estimators, whose outputs are probabilities intended to inform individual clinical decisions. For exploratory applications such as pathology feature extraction, lesion segmentation, or NLP-based structuring of free text, detection or agreement statistics (sensitivity and specificity at a working threshold, the Dice coefficient, intersection-over-union, or interrater agreement) are more relevant than calibration or DCA. Calibration becomes relevant again when such outputs feed into a downstream prediction model. We therefore limit the expectation for calibration and DCA in CRAVE-Hep's D2 dimension (Table 5) to prediction models, avoiding overgeneralization to detection- or extraction-oriented AI.

Data leakage and validation design. Internal validation design is equally important. Random splitting of longitudinal or repeated-measures data risks data leakage, with the same patient or time period appearing in both training and test sets, which inflates apparent performance. More credible designs validate across time (training on earlier years and testing on later ones) or across centers. Few hepatitis AI studies report calibration, DCA, or leakage-resistant validation; we regard their routine inclusion as a minimum standard for future work. Table 1 records the validation type when it was clear in the original study.

3.9. Ethics, Privacy-Preserving Learning, and Implementation

Across domains, recurring weaknesses constrain translation: limited external and prospective validation, models that drift outside the development setting, restricted interpretability, algorithmic bias from nonrepresentative data, and inconsistent regulation (6). A 2025 scoping review of AI in hepatology screened 3,214 records and included 75 studies, finding that most were retrospective and single-center and that deficient equity reporting and scarce prospective evidence were major barriers (6). Many published prediction models would be flagged as having a high risk of bias under PROBAST+AI, commonly because of inadequate sample size, optimistic internal validation, and incomplete reporting, and few fully meet TRIPOD+AI expectations (23, 24). Interventional and early-stage evaluations adhering to CONSORT-AI/SPIRIT-AI and DECIDE-AI remain rare in this field (25, 26).

Privacy-preserving and federated learning. Hepatitis data—particularly for HBV/HCV and HIV co-infection—are highly sensitive, and direct sharing of patient-level data across institutions or countries is often infeasible. Federated learning, in which models are trained locally and only parameters are shared, offers a route to multicenter training and external validation without centralizing records; in one multinational study, it generalized across heterogeneous, unharmonized datasets without data exchange (29). For regions building research capacity, including Iran, federated approaches may be the only practical path to the diverse datasets required for robust validation. Privacy-preserving techniques should accompany conventional governance (consent, accountability, transparency) and advance alongside technical development (22).

Implementation in resource-limited settings and the Iranian context. Models trained mainly on data from high-income countries may not generalize to other settings because genotypes, comorbidity profiles, and care environments differ. Iran illustrates both the disease burden and the opportunity: population-based estimates place chronic HBsAg prevalence at approximately 2% in the general population (32). HCV antibody prevalence is low in the general population, having fallen from an estimated 1.3% to 0.9% between 2010 and 2016, but exceeds 50% among people who inject drugs; the national strategy is oriented toward WHO HCV-elimination targets for 2030 (33). Realizing the potential of AI depends on health information infrastructure. Iran has invested in national systems, including SIB, its integrated primary-care EHR platform, launched in 2015 across more than 36,000 centers (34), and SEPAS, its national hospital-based EHR system. SEPAS was built on the ISO 13606 standard rather than the Fast Healthcare Interoperability Resources or Observational Medical Outcomes Partnership common data model now widely used internationally (35). This standards gap, together with substantial free-text content, means that AI deployment will require deliberate data engineering, including NLP-based extraction of unstructured notes and mapping national records to interoperable models before reliable ML can be applied. Several of these points generalize to other settings with

high disease burdens and limited resources, particularly the gap between locally built national systems and international interoperability standards and reliance on unstructured free text. Other features, including the particular architecture of SIB and SEPAS and their use of ISO 13606, are specific to Iran's systems and should not be assumed to apply elsewhere without local verification. We present this regional dimension as a research and policy agenda rather than as evidence of current deployment.

3.10. The CRAVE-Hep Framework: Structuring Clinical-Readiness Judgments

The preceding sections show a consistent pattern: hepatitis AI tools are frequently reported in terms of discrimination alone, with heterogeneous validation and limited attention to calibration, net benefit, governance, or implementation. To make these judgments explicit and comparable, we propose CRAVE-Hep, a framework comprising six dimensions of clinical readiness, each representing a qualitative progression from lower to higher maturity (Table 5; its derivation is documented in Table 4). The framework remains qualitative because the relative weights of the dimensions are unknown. CRAVE-Hep does not replace the formal tools on which it draws (TRIPOD+AI, PROBAST+AI, CONSORT-AI/SPIRIT-AI, and DECIDE-AI) (23-27); rather, it brings their principles together in a rapid appraisal aid. It remains a proposed framework rather than a validated instrument.

The six dimensions (Table 5) are as follows. D1, validation rigor, progresses from a single internal random split through temporal or external validation to prospective testing. D2, performance integrity, progresses from discrimination alone to reported calibration and then demonstrated net benefit on DCA. D3, data representativeness and bias, progresses from a small single-center sample to large, diverse, multicenter data with transparent equity reporting. D4, transparency and interpretability, progresses from an opaque model to full TRIPOD+AI reporting and an explanation of how the model reached its output. D5, data governance and privacy, progresses from unspecified data handling to explicit governance and privacy-preserving or federated training. D6, implementation fit, progresses from a purely theoretical model to an interoperable model integrated into a workflow in its intended setting. A tool's profile is interpreted across all six dimensions rather than by averaging them into a single number.

Worked illustration: Table 6 applies the core questions for each of the six dimensions to representative tools from this review. The illustration remains qualitative because numeric scores would imply a level of reproducibility that has not been established for this unvalidated framework. Despite this limitation, a pattern emerges. The best-studied fibrosis-staging models address validation and data reasonably well but provide limited information on performance integrity or implementation. Registry-based treatment models have substantial data volume but lack external validation and deployment evidence. None of these worked examples addresses all six dimensions, supporting supervised assistance rather than autonomous operation as their current role. When applied to a new tool, the same questions can serve as a design checklist to identify missing evidence for each dimension, such as a calibration plot, an external cohort, or a workflow study.

Readiness vocabulary: To avoid implying deployability when it has not been demonstrated, we use three readiness tiers throughout Section 3 and Table 2: "technically promising" (discrimination demonstrated in ≥ 1 retrospective study); "clinically evaluable" (calibration and/or DCA, with external validation reported); and "ready for clinical deployment" (prospective external validation, calibration, DCA, and workflow or implementation testing all demonstrated). By these definitions, no domain reviewed here and no tool in Table 6 currently reaches "ready for clinical deployment."

3.11. Limitations of CRAVE-Hep

CRAVE-Hep remains unvalidated: no *Delphi consensus*, independent expert appraisal, reliability testing, or usability evaluation has been completed. Table 6 presents interpretive judgments, not reproducible measurements, and is not an appraisal of every source cited. The dimensions have no established weights and should not be collapsed into a score. Cost-effectiveness, regulatory status, acceptability, and model drift remain important second-tier domains (Table 4). Purposive selection introduces bias. Search sources used different publication-date fields and levels of filter granularity, limiting direct comparison of retrieval totals. Repeating searches during revision does not make purposive selection exhaustive. Europe PMC overlaps with PubMed, and six first-page Scholar inspections do not establish comprehensive coverage. Added reports were not subjected to a systematic full-text risk-of-bias assessment. These limitations preclude claims of exhaustive synthesis or validated readiness grading.

3.12. Proposed Validation Roadmap

Validation should begin with a *Delphi* or modified nominal-group process involving hepatologists, infectious-disease physicians, AI methodologists, implementation scientists, and patients, drawing on the consensus approaches used for CONSORT-AI and DECIDE-AI (25, 26). Independent appraisers should then assess the same published studies and report interrater reliability, such as

weighted kappa or an intraclass correlation coefficient. Subsequent work should test construct validity, prospective usability, and whether the framework improves evidence reporting or decision-making. As a provisional minimum for use beyond discussion, we propose completed consensus development and published evidence of agreement, defined as a weighted kappa of at least 0.6 across two or more independent appraisers; this threshold itself requires justification and calibration. Until such evidence is available, CRAVE-Hep should not determine funding, regulatory, or clinical-deployment decisions. It remains a shared vocabulary and research agenda.

4. Conclusions

Applications of AI have advanced across the viral hepatitis care cascade, with the strongest current evidence in DL-based noninvasive fibrosis staging, ML-based case-finding and treatment-response prediction, and AI-assisted HCC surveillance. Large language models offer emerging but largely unvalidated capabilities. However, appraisal using CRAVE-Hep suggests that the field is less mature than headline metrics imply. Most models are technically promising rather than clinically evaluable or ready for clinical deployment (Section 3.10; Table 2): they are retrospective and internally validated using datasets that may not represent target populations; calibration and net benefit are seldom reported; interpretability, governance, and regulation remain unresolved; and prospective evidence that AI improves outcomes is scarce. In Table 6, none of the worked examples addresses all six CRAVE-Hep dimensions. The field therefore shows promise but has structurally incomplete evidence for deployment readiness.

Accordingly, the responsible near-term role is augmentation: AI can extend the reach and consistency of clinical expertise, especially where specialist resources are limited. For practicing hepatologists, Table 2 provides practical guidance: before adopting an AI tool, identify its demonstrated readiness tier and request specific missing evidence, such as a calibration plot, an external cohort, or a workflow study, rather than relying on a headline AUC. Wider impact requires prospective external validation and representative, equitable datasets; greater emphasis on calibration, DCA, and leakage-resistant validation; transparent reporting aligned with TRIPOD+AI, PROBAST+AI, and CONSORT-AI/SPIRIT-AI; privacy-preserving methods such as federated learning; and locally adapted validation in regions with high disease burdens.

A roadmap for the region. For Iran and comparable settings, CRAVE-Hep implies a concrete agenda. First, establish national hepatitis data consortia to assemble multicenter benchmark datasets (raising D3) and use federated learning where data cannot be centralized (raising D5). Second, standardize clinical labels—for example, advanced fibrosis and clinically significant portal hypertension according to contemporary criteria such as Baveno VII (36)—so that models trained at different sites are comparable. Third, invest in data engineering to map national records (SIB, SEPAS) to interoperable models (FHIR/OMOP) and apply NLP to structure free-text reports (raising D6). Fourth, require AI studies to report calibration, net benefit, and external validation as a condition for making clinical claims (raising D1–D2). As emphasized in Section 3.9, this agenda is a set of proposed priorities for the region, not evidence that these systems are already validated or deployed. Pursued together, these steps would move AI in hepatitis up the readiness scale toward tools that are validated, equitable, and genuinely useful at the point of care.

Supplementary Material

Supplementary material(s) is available here [To read supplementary materials, please refer to the journal website and open PDF/HTML].

Footnotes

AI Use Disclosure: Claude was used for minor text editing in the Footnotes and Disclosures section.

Authors' Contribution: Study Concept and Design, Manuscript Drafting, and CRAVE-Hep Framework Development: A. B.; Clinical Interpretation and Critical Manuscript Revision: F. B.; Both authors approved the final manuscript and agreed to be accountable for the work.

Conflict of Interests Statement: The authors do not declare any conflicts of interest for this study.

Data Availability: The dataset presented in the study is available on request from the corresponding author during submission or after publication.

Funding/Support: No funding was received for this study.

References

1. Cornberg M, Sandmann L, Jaroszewicz J, Kennedy P, Lampertico P, Lemoine M, et al. EASL Clinical Practice Guidelines on the management of hepatitis B virus infection. *J Hepatol*. 2025;**83**(2):502-583. [PubMed ID: 40348683]. [PubMed Central ID: PMC13289930]. https://doi.org/10.1016/j.jhep.2025.03.018.

2. Hudu SA, Alshrari AS, Abu-Shoura EJI, Osman A, Jimoh AO. A critical review of the prospect of integrating artificial intelligence in infectious disease diagnosis and prognosis. *Interdiscip Perspect Infect Dis*. 2025;**2025**(1). 6816002. [PubMed ID: 40225950]. [PubMed Central ID: PMC11991796]. https://doi.org/10.1155/ipid/6816002.
3. Chen ML, Li WM, Liu Q, Gu Y, Wang JR. Revolutionizing viral hepatitis management: artificial intelligence-assisted diagnosis and personalized treatment. *Artif Intell Gastroenterol*. 2025;**6**(1):107277. https://doi.org/10.35712/aig.v6.i1.107277.
4. Mehraeen E, Attarian N, Afra A, Parsakian S, SeyedAlinaghi S, Bondareva IL. The Current State of Artificial Intelligence in Healthcare: A Narrative Review of Opportunities and Challenges. *Shiraz E-Med J*. 2026;**26**(10). e161280. https://doi.org/10.5812/semj-161280.
5. Gao B, Duan W. The current status and future directions of artificial intelligence in the prediction, diagnosis, and treatment of liver diseases. *Digit Health*. 2025;**11**. 20552076251325400. [PubMed ID: 40290269]. [PubMed Central ID: PMC12033675]. https://doi.org/10.1177/20552076251325418.
6. Eskandar K. Artificial intelligence in hepatology: a comprehensive scoping review of clinical applications, challenges, and future directions. *iLIVER*. 2025;**4**(4). 100205. [PubMed ID: 41399683]. [PubMed Central ID: PMC12702015]. https://doi.org/10.1016/j.iliver.2025.100205.
7. Xue LY, Jiang ZY, Fu TT, Wang QM, Zhu YL, Dai M, et al. Transfer learning radiomics based on multimodal ultrasound imaging for staging liver fibrosis. *Eur Radiol*. 2020;**30**(5):2973-2983. [PubMed ID: 31965257]. [PubMed Central ID: PMC7160214]. https://doi.org/10.1007/s00330-019-06595-w.
8. Yang K, Chen F, Tian A, Deng L, Mao X. A novel deep learning framework for liver fibrosis staging and etiology diagnosis using integrated liver-spleen elastography. *Diagnostics (Basel)*. 2025;**15**(23):2986. [PubMed ID: 41374368]. [PubMed Central ID: PMC12691310]. https://doi.org/10.3390/diagnostics15232986.
9. Yu Y, Wang J, Ng CW, Ma Y, Mo S, Fong ELS, et al. Deep learning enables automated scoring of liver fibrosis stages. *Sci Rep*. 2018;**8**(1). 16016. [PubMed ID: 30375454]. [PubMed Central ID: PMC6207665]. https://doi.org/10.1038/s41598-018-34300-2.
10. Harabor V, Mogos R, Nechita A, Adam AM, Adam G, Melinte-Popescu AS, et al. Machine learning approaches for the prediction of hepatitis B and C seropositivity. *Int J Environ Res Public Health*. 2023;**20**(3):2380. [PubMed ID: 36767747]. [PubMed Central ID: PMC9915359]. https://doi.org/10.3390/ijerph20032380.
11. Park H, Lo-Ciganic WH, Huang J, Wu Y, Henry L, Peter J, et al. Evaluation of machine learning algorithms for predicting direct-acting antiviral treatment failure among patients with chronic hepatitis C infection. *Sci Rep*. 2022;**12**(1). 18094. [PubMed ID: 36302828]. [PubMed Central ID: PMC9613877]. https://doi.org/10.1038/s41598-022-22819-4.
12. Haga H, Sato H, Koseki A, Saito T, Okumoto K, Hoshikawa K, et al. A machine learning-based treatment prediction model using whole genome variants of hepatitis C virus. *PLoS One*. 2020;**15**(11). e0242028. [PubMed ID: 33152046]. [PubMed Central ID: PMC7644079]. https://doi.org/10.1371/journal.pone.0242028.
13. Nishwa DE, Abbas Z, Lee SW. Machine learning-based prediction of treatment response in comorbid hepatitis C patients receiving DAA therapy: a real-world study from Pakistan. *Front Public Health*. 2026;**14**. 1783217. [PubMed ID: 42052030]. [PubMed Central ID: PMC13111370]. https://doi.org/10.3389/fpubh.2026.1783217.
14. Feldman TC, Dienstag JL, Mandl KD, Tseng YJ. Machine-learning-based predictions of direct-acting antiviral therapy duration for patients with hepatitis C. *Int J Med Inform*. 2021;**154**. 104562. [PubMed ID: 34482150]. https://doi.org/10.1016/j.ijmedinf.2021.104562.
15. Churkin A, Kriss S, Uziel A, Goyal A, Zakh R, Cotler SJ, et al. Machine learning for mathematical models of HCV kinetics during antiviral therapy. *Math Biosci*. 2022;**343**. 108756. [PubMed ID: 34883104]. [PubMed Central ID: PMC8792269]. https://doi.org/10.1016/j.mbs.2021.108756.
16. Abdelhamed W, El-Kassas M. Integrating artificial intelligence into multidisciplinary evaluations of hepatocellular carcinoma: opportunities and challenges. *Hepatoma Res*. 2025;**11**:8. https://doi.org/10.20517/2394-5079.2024.138.
17. Lu MY, Wu JCH, Lu HHS, Eslam M, Yu ML. Artificial intelligence applications in the diagnosis, treatment, and prognosis of hepatocellular carcinoma. *Gut Liver*. 2026;**20**(1):5-23. [PubMed ID: 41472345]. [PubMed Central ID: PMC12800664]. https://doi.org/10.5009/gnl250268.
18. Konerman MA, Beste LA, Van T, Liu B, Zhang X, Zhu J, et al. Machine learning models to predict disease progression among veterans with hepatitis C virus. *PLoS One*. 2019;**14**(1). e0208141. [PubMed ID: 30608929]. [PubMed Central ID: PMC6319806]. https://doi.org/10.1371/journal.pone.0208141.
19. Wang K, Lu X, Zhou H, Gao Y, Zheng J, Tong M, et al. Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study. *Gut*. 2019;**68**(4):729-741. [PubMed ID: 29730602]. [PubMed Central ID: PMC6580779]. https://doi.org/10.1136/gutjnl-2018-316204.
20. Wong GLH, Hui VWK, Tan Q, Xu J, Lee HW, Yip TCF, et al. Novel machine learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis. *JHEP Rep*. 2022;**4**(3). 100441. [PubMed ID: 35198928]. [PubMed Central ID: PMC8844233]. https://doi.org/10.1016/j.jhepr.2022.100441.
21. Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai JC, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. *Hepatology*. 2024;**80**(5):1158-1168. [PubMed ID: 38451962]. [PubMed Central ID: PMC11706764]. https://doi.org/10.1097/HEP.0000000000000834.
22. Bal T. A new tool for the diagnosis and management of viral hepatitis: artificial intelligence. *Viral Hepat J*. 2024;**30**(1):1-6. https://doi.org/10.4274/vhd.galenos.2024.2023-12-4.
23. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;**385**. e078378. [PubMed ID: 38626948]. [PubMed Central ID: PMC11019967]. https://doi.org/10.1136/bmj-2023-078378.
24. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;**388**. e082505. [PubMed ID: 40127903]. [PubMed Central ID: PMC11931409]. https://doi.org/10.1136/bmj-2024-082505.
25. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Chan AW, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;**26**(9):1364-1374. [PubMed ID: 32908283]. [PubMed Central ID: PMC7598943]. https://doi.org/10.1038/s41591-020-1034-x.
26. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;**28**(5):924-933. [PubMed ID: 35585198]. https://doi.org/10.1038/s41591-022-01772-9.
27. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, Darzi A, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;**26**(9):1351-1363. [PubMed ID: 32908284]. [PubMed Central ID: PMC7598944]. https://doi.org/10.1038/s41591-020-1037-7.
28. Baethge C, Goldbeck-Wood S, Mertens S. SANRA-a scale for the quality assessment of narrative review articles. *Res Integr Peer Rev*. 2019;**4**(1).

5. [PubMed ID: 30962953]. [PubMed Central ID: PMC6434870]. https://doi.org/10.1186/s41073-019-0064-8.

29. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med*. 2021;**27**(10):1735-1743. [PubMed ID: 34526699]. [PubMed Central ID: PMC9157510]. https://doi.org/10.1038/s41591-021-01506-3.
30. Chen Y, Shen Y, Wang X. Artificial Intelligence in Managing Liver Cirrhosis and Variceal Bleeding: A Review. *Hepat Mon*. 2025;**25**(1). e160500. https://doi.org/10.5812/hepatmon-160500.
31. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. *PLOS Digit Health*. 2025;**4**(6). e0000877. [PubMed ID: 40498738]. [PubMed Central ID: PMC12157099]. https://doi.org/10.1371/journal.pdig.0000877.
32. Salehi-Vaziri M, Sadeghi F, Almasi Hashiani A, Gholami Fesharaki M, Alavian SM. Hepatitis B virus infection in the general population of Iran: an updated systematic review and meta-analysis. *Hepat Mon*. 2016;**16**(4). e35577. [PubMed ID: 27257428]. [PubMed Central ID: PMC4888501]. https://doi.org/10.5812/hepatmon.35577.
33. Mousavi S, Alavi M, Delavari A, Poustchi H, Mohammadi Z, Malekzadeh R. Towards hepatitis C virus elimination in Iran: a blueprint for comprehensive strategies. *J Viral Hepat*. 2024;**31**(9):565-572. [PubMed ID: 38831601]. https://doi.org/10.1111/jvh.13975.
34. Kalkhajeh SG, Aghajari A, Dindamal B, Shahvali-Kuhshuri Z, Faraji-Khiavi F. The Integrated Electronic Health System in Iranian health centers: benefits and challenges. *BMC Prim Care*. 2023;**24**(1). 53. [PubMed ID: 36803274]. [PubMed Central ID: PMC9938354]. https://doi.org/10.1186/s12875-023-02011-x.
35. Asadi F, Moghaddasi H, Rabiei R, Rahimi F, Mirshekarlou S. The evaluation of SEPAS national project based on electronic health record system (EHRS) coordinates in Iran. *Acta Inform Med*. 2015;**23**(6):369-373. [PubMed ID: 26862248]. [PubMed Central ID: PMC4720822]. https://doi.org/10.5455/aim.2015.23.369-373.
36. de Franchis R, Bosch J, Garcia-Tsao G, Reiberger T, Ripoll C, Abraldes JG, et al. Baveno VII-renewing consensus in portal hypertension. *J Hepatol*. 2022;**76**(4):959-974. [PubMed ID: 35120736]. [PubMed Central ID: PMC11090185]. https://doi.org/10.1016/j.jhep.2021.12.022.