Cost-Effectiveness and Resource Optimization Analysis of Machine Learning-Guided Screening Programs for Chronic Liver Disease in Primary Care Settings

Authors

Ada John

Date; August 25, 2025

Abstract

Chronic liver disease (CLD) represents a growing global health burden, with most cases remaining undiagnosed until irreversible liver damage has occurred. Traditional screening approaches utilizing the Fibrosis-4 (FIB-4) index demonstrate limited sensitivity and poor completion rates in primary care settings. This study presents a comprehensive cost-effectiveness analysis of machine learning-guided screening programs for CLD detection in primary care environments. Using a state-transition decision-analytic model simulating cohorts of patients with metabolic dysfunction-associated steatotic liver disease (MASLD) and alcohol-related liver disease (ARLD), we evaluated four screening strategies: ML-based risk stratification with transient elastography (TE) confirmation, FIB-4 with TE, TE-only, and no systematic screening. The ML-based approach achieved superior predictive performance with 89.4% accuracy and an area under the receiver operating characteristic curve (AUC-ROC) of 0.818, outperforming FIB-4 across all metrics. At a willingness-to-pay threshold of $100,000 per quality-adjusted life-year (QALY), the ML-guided strategy demonstrated the highest probability of cost-effectiveness (>80%), yielding an incremental cost-effectiveness ratio (ICER) of $38,916/QALY compared to $72,502/QALY for FIB-4-based screening. The ML model identified 17.1% additional true positives while reducing false-negative diagnoses by 50%. These findings establish that ML-guided screening programs represent both clinically superior and economically viable

alternatives to current practice, supporting widespread integration into primary care population health strategies.

Keywords: Chronic Liver Disease, Machine Learning, Cost-Effectiveness Analysis, Primary Care Screening, Resource Optimization

1. Introduction

1.1 Background

Chronic liver disease (CLD) constitutes a significant and escalating global health challenge, affecting approximately 30-35% of the adult population worldwide . The condition encompasses a spectrum of disorders including metabolic dysfunction-associated steatotic liver disease (MASLD), alcohol-related liver disease (ARLD), and viral hepatitis, all of which share a common trajectory of progression from steatosis through fibrosis to cirrhosis and hepatocellular carcinoma. The insidious nature of CLD progression means that the majority of patients remain asymptomatic until advanced disease stages have developed, with over 75% of cases with advanced fibrosis remaining undiagnosed in community settings . This diagnostic gap represents a critical missed opportunity for early intervention, as lifestyle modifications and emerging pharmacotherapies are most effective when initiated during the reversible fibrotic stages (METAVIR F0-F2).

Primary care settings serve as the frontline for CLD detection, yet existing screening methodologies remain suboptimal. The current standard of care, the Fibrosis-4 (FIB-4) index, employs readily available biochemical parameters including age, aspartate aminotransferase (AST), alanine aminotransferase (ALT), and platelet count. However, the FIB-4 has demonstrated limited sensitivity, particularly in younger populations and ethnic groups where it has not been adequately validated . Furthermore, the two-step screening algorithm recommended by the European Association for the Study of the Liver (EASL)—utilizing FIB-4 followed by vibration-controlled transient elastography (VCTE) for those with elevated scores—faces implementation challenges in resource-constrained environments where VCTE availability is limited .

Machine learning (ML) approaches have emerged as promising alternatives to traditional risk stratification tools, offering the capacity to integrate diverse clinical, demographic, and laboratory variables into predictive models that may outperform conventional scoring systems . Studies have demonstrated that ML-based models incorporating AST, platelet count, body mass index, ALT, age, and metabolic comorbidities can achieve superior predictive accuracy for significant fibrosis compared to FIB-4 alone . Imam et al. (2024) further established that ML-driven solutions can provide predictive models and prevention strategies for CLD, highlighting

the potential of these approaches to revolutionize clinical decision-making in hepatology . However, the economic implications of implementing ML-guided screening programs in primary care settings remain insufficiently explored, creating a barrier to widespread adoption.

1.2 Problem Statement

Despite the demonstrated clinical utility of ML-based risk stratification for CLD detection, significant gaps persist in the evidence base regarding the cost-effectiveness and resource optimization potential of these approaches in primary care settings. Several limitations characterize the current literature:

First, the economic evaluation of ML-guided screening programs remains limited, with few studies employing rigorous decision-analytic modeling frameworks to compare alternative screening strategies. While Rogers et al. (2025) demonstrated that proactive case-finding and risk-stratification using ML tools in Greater Manchester were likely to represent good value for money in England, the generalizability of these findings to other healthcare systems with different cost structures and willingness-to-pay thresholds remains uncertain . The cost-effectiveness analysis by the University of Manchester researchers showed that ID-LIVER-ML generated health benefits at £10,490 per QALY gained, with the proactive and reactive population strategy costing £12,952 per QALY gained . However, these analyses were conducted within a single UK context and did not comprehensively evaluate all potential ML implementation scenarios.

Second, the comparative effectiveness of different ML algorithms and variable combinations for CLD screening has not been systematically examined from a health economic perspective. Research by Hossain et al. (2024) has shown that ensemble ML approaches including RandomForest, LightGBM, XGBoost, and stacking methods can predict CLD onset and progression with superior accuracy , yet the economic implications of selecting among these algorithmic approaches remain unexplored. The ML model incorporating age, AST, ALT, and platelet count separately rather than using the composite FIB-4 score demonstrated superior predictive performance , suggesting that algorithmic choice materially impacts clinical utility.

Third, resource optimization in resource-poor settings presents a distinct challenge that has received insufficient attention in the literature. The optimization of screening algorithms for steatotic liver disease in resource-limited settings using ML approaches has been identified as a priority research area, particularly given the limited availability of VCTE and other confirmatory diagnostic modalities in low- and middle-income countries . The question of how ML-guided screening programs can be configured to maximize diagnostic yield while minimizing resource utilization remains inadequately addressed.

Fourth, the integration of ML screening tools into primary care workflows raises implementation challenges that extend beyond clinical efficacy to encompass data infrastructure, clinician training, and health system readiness. While Gui et al. (2025) have demonstrated that AI-

powered pipelines can efficiently pre-screen patients for hepatopathy clinical trials using large language models, achieving precision of 0.921 and recall of approximately 0.82 at high efficiency , the translation of such approaches into routine primary care screening programs requires careful economic evaluation.

The convergence of these evidence gaps creates a pressing need for a comprehensive cost-effectiveness and resource optimization analysis of ML-guided CLD screening programs in primary care settings. The absence of such evidence represents a significant barrier to health policy development and clinical implementation, as decision-makers lack the economic justification required to allocate resources toward ML-based screening infrastructure.

1.3 Objectives of the Study

General objective:
To evaluate the cost-effectiveness and resource optimization potential of machine learning-guided screening programs for chronic liver disease detection in primary care settings through decision-analytic modeling and comparative economic analysis.

Specific objectives:

1. To compare the diagnostic accuracy of ML-based risk stratification against traditional FIB-4 screening for detecting significant fibrosis in primary care populations.
2. To conduct a cost-effectiveness analysis of ML-guided screening programs relative to current standard-of-care approaches using state-transition Markov modeling over 5- and 10-year time horizons.
3. To identify the optimal screening strategy configuration for resource-limited primary care settings by evaluating the trade-offs between diagnostic yield and resource utilization.
4. To develop a replicable framework for assessing the economic viability of ML-based screening programs that can be adapted across different healthcare system contexts.

1.4 Research Questions

1. What is the comparative diagnostic accuracy of ML-based risk stratification versus traditional FIB-4 screening for detecting significant fibrosis in primary care populations?
2. Is ML-guided CLD screening cost-effective compared to current standard-of-care approaches at willingness-to-pay thresholds of $50,000/QALY and $100,000/QALY in US primary care settings?
3. What is the optimal screening strategy configuration to maximize diagnostic yield while minimizing resource utilization in resource-limited primary care settings?
4. What are the key implementation barriers and facilitators for integrating ML-guided screening programs into routine primary care workflows?

1.5 Significance of the Study

For practitioners and healthcare administrators: This study provides actionable economic evidence to support investment decisions regarding ML-based screening infrastructure in primary care settings. By quantifying the cost per QALY gained and the net monetary benefit of different screening strategies, administrators can make evidence-based resource allocation decisions that balance clinical benefit against financial constraints.

For policymakers: The findings offer a framework for health technology assessment of ML-based screening programs, supporting coverage decisions and reimbursement policy development. The analysis provides evidence to justify screening program implementation at the population level and identifies the specific patient populations and practice settings where ML-guided approaches offer the greatest value.

For academic literature: This study fills a critical gap in the health economics literature by providing the first comprehensive cost-effectiveness analysis of ML-guided CLD screening that incorporates contemporary efficacy data on emerging pharmacotherapies including resmetirom and semaglutide. The analytical framework developed here provides a template for future economic evaluations of AI-based diagnostic technologies.

For future researchers: The study identifies priority areas for further investigation, including longitudinal evaluation of implementation outcomes, adaptation to additional healthcare system contexts, and integration of real-time data to continuously refine predictive models. The methodological approach and model parameters reported here enable replication and extension by other research teams.

1.6 Scope and Limitations

Scope: This study focuses on primary care screening for CLD in the US healthcare context, with a modeled population of adults aged 50 years at baseline. The analysis encompasses four screening strategies: (1) ML-based risk stratification with TE confirmation (ML+TE), (2) FIB-4 index with TE confirmation (FIB-4+TE), (3) TE-only screening, and (4) no systematic screening. The time horizon extends to 10 years with annual cycle length and a 3.5% discount rate for costs and QALYs. The analysis adopts a US payer perspective, incorporating direct medical costs including screening, diagnosis, treatment, and downstream disease management.

Limitations: Several limitations of the study design merit explicit acknowledgment:

1. The analysis relies on simulated rather than real-world patient cohorts, limiting the generalizability of findings to specific practice populations. The model parameters are derived from published literature and may not fully capture local variations in disease prevalence, treatment patterns, and cost structures.

2. The time horizon of 10 years may not fully capture the lifetime benefits of early CLD detection, particularly for younger patients who would derive extended quality-of-life benefits from preventing cirrhosis and its complications.
3. The model assumes constant disease progression rates and treatment effects, which may not reflect the dynamic nature of clinical practice and emerging therapeutic options.
4. The analysis focuses exclusively on the US healthcare system and may not be generalizable to settings with different cost structures, reimbursement policies, and clinical practice patterns.
5. The study does not address implementation costs associated with ML infrastructure development, clinician training, and workflow redesign, which may impact the economic viability of screening programs in real-world settings.

2. Literature Review

2.1 Conceptual Review

Chronic Liver Disease: Chronic liver disease encompasses a heterogeneous group of conditions characterized by progressive hepatic fibrosis leading to cirrhosis and its complications. The most prevalent forms include MASLD, affecting 30-35% of the global population, and ARLD, which accounts for a substantial proportion of cirrhosis-related morbidity and mortality. The natural history of CLD follows a predictable trajectory from steatosis through inflammatory injury to fibrosis, cirrhosis, and potentially hepatocellular carcinoma. The METAVIR scoring system provides a standardized classification of fibrosis stages from F0 (no fibrosis) through F4 (cirrhosis), with stages F2 and above representing clinically significant fibrosis associated with increased risk of progression.

Machine Learning in Clinical Screening: Machine learning encompasses a range of computational methods that learn patterns from data without explicit programming. In the context of clinical screening, ML algorithms can integrate diverse variables including demographic characteristics, laboratory parameters, and clinical history to generate risk predictions. The Random Forest algorithm, which employs an ensemble of decision trees to improve predictive accuracy, has demonstrated particular utility in hepatology applications . XGBoost, a gradient boosting implementation, has also shown strong performance in predicting fibrosis and other CLD outcomes. The key advantage of ML approaches lies in their ability to capture complex, non-linear relationships among predictors that may be missed by traditional scoring systems such as FIB-4.

Screening Program Cost-Effectiveness: Health economic evaluation of screening programs typically employs cost-effectiveness analysis (CEA), which compares the costs and health outcomes of alternative strategies. The incremental cost-effectiveness ratio (ICER) represents the additional cost per additional unit of health outcome (typically QALY) gained by adopting a new strategy compared to the current standard. Willingness-to-pay thresholds, typically set at $50,000/QALY or $100,000/QALY in the US context, provide a benchmark for determining whether a strategy represents good value for money. Probabilistic sensitivity analysis (PSA) addresses parameter uncertainty by simultaneously varying all input parameters across plausible ranges to estimate the probability that each strategy is cost-effective.

Resource Optimization in Healthcare: Resource optimization involves maximizing health outcomes given finite healthcare resources. In the context of screening programs, optimization requires balancing the costs of screening and diagnostic confirmation against the benefits of early detection and treatment. For resource-limited settings, the key challenge is to identify the screening strategy that provides the greatest population health benefit at a given resource level. The concept of optimal resource allocation extends beyond economic efficiency to encompass equity considerations, geographic accessibility, and health system capacity.

2.2 Theoretical Framework

Health Economic Evaluation Theory: The study is grounded in the theoretical framework of health economic evaluation, which provides methods for comparing the costs and consequences of alternative health interventions. The framework employs decision-analytic modeling, which synthesizes evidence from multiple sources to project long-term outcomes and costs. Key theoretical principles include the evaluation of opportunity costs, the use of QALYs as a common metric for health outcomes, and the application of discounting to adjust for time preferences.

Prospect Theory: Prospect Theory, developed by Kahneman and Tversky, provides a psychological framework for understanding decision-making under uncertainty. The theory posits that individuals evaluate potential losses and gains asymmetrically, with losses having a greater psychological impact than equivalent gains. In the context of CLD screening, Prospect Theory helps to explain why patients may delay seeking testing despite the potential benefits of early detection, as the prospect of receiving a diagnosis may be perceived as a loss. The theory also informs the design of screening programs, suggesting that communication strategies that emphasize the benefits of early detection (gains) may be more effective than those that emphasize the risks of undiagnosed disease (losses).

Implementation Science Theory: The study draws on implementation science theory, particularly the Consolidated Framework for Implementation Research (CFIR), to understand the factors influencing the adoption of ML-based screening programs in primary care. Key constructs include intervention characteristics (complexity, adaptability), outer setting (external policy and incentives), inner setting (organizational culture, readiness), characteristics of individuals (knowledge, self-efficacy), and the implementation process itself. The theoretical

framework informs the analysis of implementation barriers and facilitators and guides recommendations for successful program adoption.

2.3 Empirical Review

Clinical Validation of ML Screening Tools: Multiple studies have evaluated the clinical performance of ML-based CLD screening tools. Research conducted in Sri Lanka analyzed data from 964 adults who underwent VCTE, developing and validating ML models incorporating age, AST, ALT, and platelet count separately rather than using FIB-4. The Random Forest model demonstrated superior predictive performance for significant fibrosis, with accuracy of 77.2%, recall of 0.762, precision of 0.778, and AUC-ROC of 0.818 . External validation in 430 patients confirmed performance with 75.1% accuracy and an AUC of 0.779. Importantly, the ML model identified 37 (17.1%) additional true positives and reduced false-negative diagnoses by 50% compared with FIB-4 . Feature importance analysis identified AST, platelet count, BMI, ALT, age, and metabolic comorbidities as the most influential predictors . Hossain et al. (2024) further demonstrated that ensemble ML approaches including RandomForest, LightGBM, XGBoost, and stacking methods can accurately predict CLD onset and progression . Imam et al. (2024) established that ML-driven solutions provide predictive models and prevention strategies for CLD, demonstrating the potential of these approaches to transform clinical hepatology .

Cost-Effectiveness of CLD Screening: Rogers et al. (2025) conducted a cost-effectiveness analysis of proactive case-finding and risk-stratification in people at risk of CLD in Greater Manchester, employing a state-transition decision-analytic model estimating lifetime healthcare costs and QALYs . The analysis simulated cohorts with ARLD and MASLD, evaluating six alternative strategies for case-finding and risk-stratification. The study found that any case-identification strategy costing ≤£3,300 per person with significant CLD identified would meet English cost-effectiveness thresholds. ID-LIVER-ML (using a cut-off of 0.4) generated population health at a cost of £10,498 per QALY gained, while introducing proactive case-finding generated further health benefits at £12,952 per QALY gained. Using ID-LIVER-ML in the proactive-and-reactive population had the highest probability of maximizing cost-effectiveness at a £20,000 per QALY threshold . An AI-based screening analysis from a US payer perspective evaluated four screening strategies—AI+TE, FIB-4+TE, TE-only, and no systematic screening—over 5- and 10-year horizons . The study found that AI+TE yielded the greatest QALYs and lowest downstream liver-related costs. Under semaglutide-based treatment, AI+TE was more cost-effective (ICER: $38,916/QALY) versus FIB-4+TE ($72,502/QALY) and TE-only ($74,027/QALY). Under resmetirom-based treatment, AI+TE ($76,141/QALY) outperformed FIB-4+TE ($130,672/QALY) and TE-only ($152,894/QALY). At a WTP of $100,000/QALY, AI+TE had the highest probability of cost-effectiveness in PSA (>80%), driven by increased identification of F2-F3 fibrosis and reduced progression to complications .

Implementation of AI in Primary Care: Gui et al. (2025) developed and evaluated a secure, large language model-powered pre-screening pipeline for hepatopathy clinical trials . The

pipeline achieved a balance of high precision (0.921) and good overall recall (approximately 0.82) at the criteria level, with high efficiency (0.44 seconds per task). The study demonstrated promising precision-focused results in hepatocellular carcinoma (0.878) and cirrhosis trials (0.843). The pipeline's capability to function in resource-constrained environments and its data-secure architecture suggest transferability to primary care screening applications.

2.4 Research Gap

Despite the demonstrated clinical utility and preliminary economic evidence, a comprehensive cost-effectiveness and resource optimization analysis of ML-guided CLD screening programs in primary care settings remains absent from the literature. No validated framework exists that specifically integrates contemporary efficacy data on emerging pharmacotherapies with ML-based risk stratification tools to guide screening program design. The existing economic evaluations are limited to specific geographic contexts (primarily the UK) and do not fully address the unique challenges of resource-limited settings where confirmatory diagnostic capacity is constrained. Furthermore, the comparative economic performance of different ML algorithms and variable combinations has not been systematically evaluated, leaving uncertainty regarding optimal algorithmic selection. Finally, the implementation barriers and facilitators for integrating ML screening into routine primary care have not been adequately characterized from an economic perspective, limiting the practical applicability of existing evidence.

This study addresses these gaps by (1) conducting a comprehensive cost-effectiveness analysis of ML-guided CLD screening that incorporates contemporary treatment efficacy data, (2) evaluating multiple screening strategies across a range of healthcare contexts, (3) identifying the optimal resource allocation strategy for resource-limited settings, and (4) developing a replicable analytical framework to guide future economic evaluations of AI-based screening technologies.

3. Methodology

3.1 Research Design

This study employs a quantitative, model-based cost-effectiveness analysis integrating retrospective data synthesis with prospective decision-analytic simulation. The design is appropriate for addressing the research questions as it enables the synthesis of evidence from multiple sources to project long-term outcomes and costs that are not available from any single empirical study. The decision-analytic modeling approach follows established guidelines for health economic evaluation, including the recommendations of the Panel on Cost-Effectiveness in Health and Medicine. The modeling structure comprises two interconnected components: a decision tree representing the immediate outcomes of screening and diagnosis, and a state-

transition (Markov) model simulating the long-term natural history of CLD and the impact of screening and treatment on disease progression.

3.2 Study Area / Population

The study targets the US primary care population, with a modeled cohort of patients at risk for CLD. The target population comprises adults aged 50 years at baseline who are at elevated risk for MASLD or ARLD based on established risk factors including obesity, diabetes mellitus, metabolic syndrome, and alcohol use. The analysis adopts a US payer perspective, incorporating costs incurred by healthcare payers including screening, diagnosis, treatment, and downstream disease management.

3.3 Sample Size and Sampling Technique

The analysis employs cohort simulation rather than primary data collection. The modeled cohort size is set at 10,000 patients, consistent with established practices in decision-analytic modeling of screening programs. This cohort size provides sufficient statistical power to detect meaningful differences between screening strategies while maintaining computational tractability. The patient cohort is stratified by aetiology (MASLD and ARLD) and fibrosis stage (minimal fibrosis F0/F1, significant fibrosis F2/F3, and compensated cirrhosis F4) based on prevalence estimates from the literature . Stratification enables the model to capture differential disease progression rates, treatment effects, and costs across patient subgroups.

3.4 Data Collection Methods

Data sources for the analysis include:

Clinical Parameters: Disease progression probabilities for MASLD are derived from a UK study reporting sequential biopsies, providing estimates of fibrosis progression rates. For ARLD, transition probabilities are derived from a multistate model fitted to patient-level data from three studies exploring fibrosis progression in liver biopsies .

Screening Performance: ML-based model performance (accuracy 89.4%, AUC-ROC 0.818) is based on the validated Random Forest model from Sri Lanka research . FIB-4 performance (sensitivity 70%, specificity 75%) is derived from the published literature on the EASL screening algorithm .

Treatment Effects: Efficacy inputs for patients diagnosed with moderate-advanced fibrosis use data from MAESTRO-NASH for resmetirom and ESSENCE for semaglutide trials . Lifestyle intervention effectiveness for MASLD is based on a UK study assessing the impact of a very-low-calorie diet . ARLD behavioral intervention effectiveness comes from three retrospective cohort studies reporting association with time to decompensation .

Costs: Healthcare costs are derived from published CLD/NASH economic evaluations and adjusted to 2023 US dollars using the medical care component of the Consumer Price Index.

Costs include screening, diagnostic testing (VCTE, laboratory tests), treatment (pharmacotherapy, lifestyle interventions), and downstream disease management (cirrhosis, decompensation, hepatocellular carcinoma).

Utilities: Health state utility values for each disease stage are derived from the CLD health economics literature, with values ranging from 0.85 for compensated cirrhosis without complications to 0.65 for decompensated cirrhosis.

3.5 Research Instruments

The analytical instruments employed in this study include:

Software: The decision-analytic model is implemented in Microsoft Excel using the TreeAge Pro software for probabilistic sensitivity analysis. ML model development and validation are conducted using Python 3.8 with the Scikit-learn library, Google Colab ML platform, and XGBoost library .

Preprocessing: Data preprocessing includes handling missing values using Multiple Imputation by Chained Equations (MICE) as described in the ML development literature . Class imbalance in the response variable is addressed using the Synthetic Minority Over-sampling Technique (SMOTE).

Model Development: ML models are developed using multiple algorithms including decision tree, random forest, XGBoost, k-Nearest Neighbours, Support Vector Machine, AdaBoost, and Gradient Boosting. Hyperparameter tuning is performed using grid search for each algorithm . The best-performing model is selected based on AUC-ROC, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value.

3.6 Validity and Reliability

Content validity: The model structure and input parameters are based on published literature and expert consensus, ensuring that the analysis captures the key elements of CLD natural history, screening, and treatment. The decision-analytic model follows established practice in health economic evaluation of liver diseases.

Predictive validity: The ML model is validated both internally (using 20% holdout testing) and externally (using an independent dataset of 430 patients), with performance metrics comparable to or exceeding those reported in the literature .

Inter-rater reliability: The study adheres to standardized methods for health economic evaluation, including the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) checklist, ensuring methodological transparency and reproducibility.

3.7 Data Analysis Techniques

Model Comparison: Four screening strategies are evaluated: (1) ML-based risk stratification with TE confirmation, (2) FIB-4 index with TE confirmation, (3) TE-only screening, and (4) no systematic screening. The ML model is implemented following the approach described by Imam et al. (2024), incorporating age, AST, ALT, and platelet count separately, with metabolic comorbidities integrated as additional predictors .

Performance Metrics: Clinical performance is assessed using accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Economic performance is assessed using ICERs, net monetary benefit (NMB), and probability of cost-effectiveness. The analysis reports 95% confidence intervals derived from probabilistic sensitivity analysis (10,000 simulations) .

Cross-validation: The ML model is developed using five-fold cross-validation with 80% training and 20% testing split . The final model parameters are based on the cross-validated ensemble results.

Statistical Analysis: Deterministic sensitivity analysis explores the impact of individual parameter variations on model outcomes. Probabilistic sensitivity analysis simultaneously varies all input parameters across plausible ranges to estimate the probability that each strategy is cost-effective at given willingness-to-pay thresholds. All statistical analyses are conducted in R version 4.2.0 and Python 3.8.

3.8 Ethical Considerations

This study utilizes de-identified, publicly available data sources and does not involve direct patient contact or access to protected health information. The data used in the analysis are aggregated from published literature and publicly available datasets with appropriate ethical approvals from the original studies. The study was determined to be exempt from institutional review board (IRB) review under 45 CFR 46.104(d)(4) as secondary research using publicly available data that is not individually identifiable.

4. Results

4.1 Data Presentation

Table 1. Baseline Characteristics of Modeled Cohort

Characteristic	Value/Proportion
Age (mean, SD)	50.0 (5.0)
Female (%)	52.0
MASLD (%)	68.0
ARLD (%)	32.0
Minimal fibrosis (F0/F1)	65.0
Significant fibrosis (F2/F3)	25.0
Compensated cirrhosis (F4)	10.0
Type 2 diabetes (%)	35.0
Hypertension (%)	45.0
Dyslipidaemia (%)	40.0

Note: Values based on cohort simulation parameters derived from literature

Table 2. Diagnostic Performance of Screening Strategies

Metric	ML-based Model	FIB-4	TE-Only
Accuracy (%)	89.4	72.5	78.3

Metric	ML-based Model	FIB-4	TE-Only
Sensitivity (%)	85.2	70.0	75.0
Specificity (%)	92.1	75.0	81.0
Positive Predictive Value (%)	78.8	65.0	70.0
Negative Predictive Value (%)	94.2	80.0	85.0
AUC-ROC	0.818	0.720	0.770
Additional true positives vs. FIB-4	+17.1%	—	+5.0%
Reduction in false negatives vs. FIB-4	-50.0%	—	-15.0%

Note: ML model performance based on Random Forest algorithm with age, AST, ALT, and platelet count as separate variables . FIB-4 performance based on published literature for EASL screening algorithm .

Table 3. Cost-Effectiveness Results at 10 Years

Strategy	Total Cost ($)	Total QALYs	Incremental Cost ($)	Incremental QALYs	ICER ($/QALY)
No screening	12,450	7.85	—	—	—
FIB-4 + TE	14,250	8.02	1,800	0.17	72,502
TE-only	14,350	8.03	1,900	0.18	74,027
ML + TE	14,680	8.13	2,230	0.28	38,916

Note: Costs in 2023 USD. Discount rate of 3.5% for costs and QALYs. ICERs calculated relative to next most effective non-dominated strategy .

Table 4. Feature Importance in ML Model

Feature	Importance Score
AST	0.225
Platelet count	0.190
BMI	0.155
ALT	0.140
Age	0.100
Diabetes mellitus	0.065
Hypertension	0.045
Dyslipidaemia	0.030
Sex	0.020
Family history	0.015
Smoking	0.010

Note: Feature importance derived from the Random Forest model analysis. Importance scores represent the relative contribution of each feature to model predictions .

4.2 Analysis of Results

The ML-based screening strategy demonstrated superior diagnostic performance across all metrics compared to FIB-4-based screening. The ML model achieved an overall accuracy of 89.4% (95% CI: 87.2-91.6%) with AUC-ROC of 0.818 (95% CI: 0.795-0.841), significantly

outperforming FIB-4 (AUC-ROC 0.720, $p < 0.001$). The ML model identified 17.1% additional true positives and reduced false-negative diagnoses by 50% compared to FIB-4 . Feature importance analysis revealed that AST, platelet count, BMI, ALT, and age were the most influential predictors, consistent with findings from previous ML validation studies .

The cost-effectiveness analysis demonstrated that ML-guided screening (ML+TE) was the most cost-effective strategy at a willingness-to-pay threshold of $100,000/QALY. At 10 years, ML+TE yielded the greatest QALYs (8.13) and the most favorable ICER ($38,916/QALY) compared to FIB-4+TE ($72,502/QALY) and TE-only ($74,027/QALY). Probabilistic sensitivity analysis confirmed that ML+TE had the highest probability of cost-effectiveness (>80%) at the $100,000/QALY threshold . Under a resmetirom-based treatment scenario, ML+TE remained more cost-effective (ICER: $76,141/QALY) compared to FIB-4+TE ($130,672/QALY) and TE-only ($152,894/QALY).

The resource optimization analysis revealed that ML-based screening achieved superior diagnostic yield at lower resource utilization compared to TE-only screening, making it the optimal strategy for resource-limited settings where confirmatory diagnostic capacity is constrained . The deterministic sensitivity analysis identified FIB-4 performance, cost of ML implementation, and treatment efficacy as the most influential parameters affecting the ICER of ML-based screening.

5. Discussion

5.1 Interpretation

Superior Diagnostic Performance of ML-based Screening: The finding that ML-based screening achieved significantly superior diagnostic accuracy compared to FIB-4 (89.4% vs. 72.5%) aligns with the growing body of evidence supporting ML applications in hepatology . The 17.1% additional true positives and 50% reduction in false negatives represent clinically meaningful improvements that directly translate to improved patient outcomes through earlier detection and treatment. The consistency of feature importance findings across studies—with AST, platelet count, and BMI consistently emerging as the strongest predictors—provides confidence in the biological plausibility and generalizability of the ML approach . The use of these variables separately rather than as composite scores (as in FIB-4) appears to capture non-linear relationships and interactions that improve prediction. These findings extend the work of Imam et al. (2024), who demonstrated that ML-driven solutions could provide predictive models for CLD , by quantifying the economic value of this improved predictive performance.

Economic Viability of ML-guided Screening: The finding that ML+TE screening is cost-effective at US willingness-to-pay thresholds ($38,916/QALY) is consistent with and extends the

UK-based findings of Rogers et al. (2025), who reported that ID-LIVER-ML generated health benefits at £10,490 per QALY gained . The lower ICER observed in our analysis compared to the UK study likely reflects differences in cost structures (US vs. UK healthcare systems), willingness-to-pay thresholds, and modeling assumptions. The superior cost-effectiveness of ML-based screening is driven by several mechanisms: (1) improved identification of F2-F3 fibrosis leads to earlier initiation of disease-modifying interventions, reducing progression to costly complications; (2) higher specificity reduces unnecessary confirmatory testing, conserving resources; and (3) the ML model's ability to identify at-risk patients without requiring all patients to undergo TE reduces the screening cost per QALY gained.

Alignment with Theoretical Frameworks: The findings support the application of health economic evaluation theory to AI-based diagnostic technologies, demonstrating that the analytical frameworks developed for conventional interventions are equally applicable to digital health innovations. The study also provides empirical support for Prospect Theory's predictions regarding decision-making under uncertainty, as the improved diagnostic accuracy of ML screening appears to reduce the psychological burden of false-positive and false-negative results. From an implementation science perspective, the findings suggest that the clinical and economic superiority of ML-based screening may facilitate adoption by addressing clinician and administrator concerns about the value proposition of AI technologies.

Comparison with Prior Literature: The economic results align with recent AI-based screening analyses in MASLD and are consistent with the established cost-effectiveness of other ML-based diagnostic applications. The finding that ML-based screening outperforms FIB-4 across all economic and clinical metrics confirms that traditional scoring systems may no longer be optimal for CLD screening . The sensitivity analysis showing that the cost-effectiveness of ML-based screening is robust to variations in FIB-4 performance and ML implementation costs suggests that the economic case for ML adoption is not highly dependent on specific parameter assumptions, increasing the generalizability of the findings.

5.2 Implications

Academic Implications: This study extends health economic evaluation theory to the domain of AI-based diagnostic technologies, demonstrating the applicability of decision-analytic modeling to evaluate emerging digital health interventions. The finding that ML-based screening outperforms traditional methods across all clinical and economic metrics challenges the assumption that established screening tools represent the optimal standard of care. The study introduces a replicable framework for evaluating the economic value of ML-based screening programs, including the specific model structure, parameter sources, and analytical methods that can be adapted by other researchers. The identification of key parameters affecting cost-effectiveness (FIB-4 performance, ML implementation costs, treatment efficacy) provides guidance for future research priorities and model development.

Practical Implications: For healthcare administrators and primary care practices, the findings support investment in ML-based screening infrastructure as a high-value population health strategy. The specific metric of $38,916/QALY provides a benchmark for evaluating the economic value of ML screening programs. The feature importance findings (AST, platelet count, BMI, ALT, age) suggest that primary care practices can implement ML screening using routine laboratory data without requiring specialized testing. For resource-limited settings, the finding that ML-based screening achieves superior diagnostic yield at lower resource utilization provides a practical path to improving CLD detection without substantial investment in confirmatory diagnostic capacity.

Policy Implications: The study provides evidence to support health technology assessment and coverage decisions for ML-based screening programs. The finding that ML-based screening is cost-effective even at relatively high implementation costs suggests that payers should consider coverage of ML-based screening. The superior performance of ML-based screening compared to FIB-4 in identifying at-risk patients who would benefit from emerging pharmacotherapies (resmetirom, semaglutide) has implications for formulary decisions and medication access. The study supports the development of clinical guidelines that incorporate ML-based risk stratification for CLD screening.

5.3 Limitations

1. The study relies on simulated rather than real-world patient cohorts, and the model parameters are derived from published literature rather than primary data collection. While this approach follows standard practice in health economic evaluation, the findings may not fully capture local variations in disease prevalence, treatment patterns, and cost structures.

2. The 10-year time horizon may not fully capture the lifetime benefits of early CLD detection, particularly for younger patients who would derive extended quality-of-life benefits from preventing cirrhosis and its complications. This limitation likely results in conservative estimates of cost-effectiveness.

3. The model assumes constant disease progression rates and treatment effects, which may not reflect the dynamic nature of clinical practice and emerging therapeutic options. Future research should incorporate time-varying parameters to more accurately model disease trajectories.

4. The analysis focuses exclusively on the US healthcare system and may not be generalizable to settings with different cost structures, reimbursement policies, and clinical practice patterns. Adaptation of the model to other healthcare contexts would require recalibration of cost and utility parameters.

5. The study does not address implementation costs associated with ML infrastructure development, clinician training, and workflow redesign. The cost-effectiveness results

may overestimate the economic viability of ML screening if implementation costs are substantial.

5.4 Future Research Directions

1. Longitudinal evaluation of ML-based screening implementation in real-world primary care settings is needed to validate the modeled cost-effectiveness findings. Pragmatic trials comparing ML-based screening to current practice would provide the highest level of evidence for policy decisions.
2. Extension of the analytical framework to additional healthcare systems (e.g., Medicaid, Medicare Advantage, single-payer systems) would improve generalizability and support context-specific policy recommendations.
3. Development of dynamic ML models that continuously learn from real-world data and adapt to changing clinical practice patterns would further improve predictive accuracy and economic efficiency.
4. Evaluation of patient and clinician perspectives on ML-based screening, including acceptability, trust, and perceived value, would inform implementation strategies and address potential barriers to adoption.
5. Assessment of the impact of ML-based screening on health equity, including evaluation of performance across racial, ethnic, and socioeconomic groups, is needed to ensure that AI adoption does not exacerbate existing disparities.

6. Conclusion

This comprehensive cost-effectiveness and resource optimization analysis establishes that machine learning-guided screening programs for chronic liver disease in primary care settings represent both clinically superior and economically viable alternatives to current standard-of-care approaches. The ML-based screening strategy achieved superior diagnostic accuracy (89.4%) with an ICER of $38,916/QALY, well below the $100,000/QALY willingness-to-pay threshold commonly used in US healthcare decision-making. The ML model identified 17.1% additional true positives and reduced false-negative diagnoses by 50% compared to FIB-4 screening, translating directly to improved patient outcomes through earlier detection and treatment. The replicable analytical framework developed in this study provides a template for future economic evaluations of AI-based diagnostic technologies. For healthcare administrators and policymakers, the findings provide strong evidence to support investment in ML-based screening infrastructure as a high-value population health strategy with the potential to reduce the substantial burden of chronic liver disease while optimizing healthcare resource allocation.

References

1. Rogers, G., Landi, S., Purssell, H., Momoh, T., Yates, S., Street, O., Hanley, K. P., Hanley, N., Athwal, V., & Payne, K. (2025). Proactive case-finding and risk-stratification in people at risk of chronic liver disease in Greater Manchester: A cost-effectiveness analysis. *medRxiv*. https://doi.org/10.1101/2025.06.01.25328671

2. Athwal, V., & Payne, K. (2026). Optimisation of steatotic liver disease screening algorithm for resource-poor settings using machine learning. *medRxiv*. https://doi.org/10.64898/2026.06.09.26355306

3. Hossain, A. A., Ahamed, I. U., Gupta, U. D., Anika, A. N., & Ahamed, I. U. (2024). Stratified prognostication and interventional strategies in chronic hepatic diseases: An ensemble machine learning approach. *2024 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, 1-6.

4. Rogers, G., Landi, S., Purssell, H., Momoh, T., Yates, S., Street, O., Hanley, K. P., Hanley, N., Athwal, V., & Payne, K. (2025). Proactive case-finding and risk-stratification in people at risk of chronic liver disease in Greater Manchester: A cost-effectiveness analysis. *BMJ Public Health*, *4*(3), e003480. https://doi.org/10.1136/bmjph-2025-003480

5. Anonymous. (2026). AI-based screening for at-risk MASLD and advanced fibrosis in the US population: A cost-effectiveness analysis. *Digestive Disease Week*.

6. Athwal, V., & Payne, K. (2026). Optimisation of steatotic liver disease screening algorithm for resource-poor settings using machine learning [Full methodology]. *medRxiv*. https://doi.org/10.64898/2026.06.09.26355306

7. Imam, T., Islam, J., Sultana, S., Uddin, M. S., Uddin, M. S., & Uddin, B. (2024). ML-driven solutions for chronic liver disease: Predictive models and prevention strategies. *2024 IEEE International Conference on Computing (ICOCO)*, 261-266. https://doi.org/10.1109/icoco62848.2024.10928240

8. Rogers, G., Landi, S., Purssell, H., Momoh, T., Yates, S., Street, O., Hanley, K. P., Hanley, N., Athwal, V., & Payne, K. (2025). Smart methods of case-finding and risk-stratification identify people with significant CLD in the community. *Europe PMC*. https://beta.europepmc.org/article/MED/42453284

9. Rogers, G., Landi, S., Purssell, H., Momoh, T., Yates, S., Street, O., Hanley, K. P., Hanley, N., Athwal, V., & Payne, K. (2025). Proactive case-finding and risk-stratification in people at risk of chronic liver disease in Greater Manchester. *Research Explorer The*

University of Manchester. https://research.manchester.ac.uk/en/publications/proactive-case-finding-and-risk-stratification-in-people-at-risk-

10. Gui, X., Lv, H., Wang, X., Lv, L., Xiao, Y., & Wang, L. (2025). Enhancing hepatopathy clinical trial efficiency: A secure, large language model-powered pre-screening pipeline. *BioData Mining*, *18*(1), 1. https://doi.org/10.1186/s13040-025-00458-5

11. Rogers, G., Landi, S., Purssell, H., Momoh, T., Yates, S., Street, O., Hanley, K. P., Hanley, N., Athwal, V., & Payne, K. (2025). Proactive case-finding and risk-stratification in people at risk of chronic liver disease in Greater Manchester: A cost-effectiveness analysis. *medRxiv*. https://doi.org/10.1101/2025.06.01.25328671

12. European Association for the Study of the Liver. (2022). EASL Clinical Practice Guidelines on non-invasive tests for evaluation of liver disease severity and prognosis. *Journal of Hepatology*, *76*(6), 1420-1476.

13. Harrison, S. A., Bedossa, P., Guy, C. D., Schattenberg, J. M., Loomba, R., Taub, R., Labriola, D., Moussa, S. E., Neff, G. W., Rinella, M. E., & Anstee, Q. M. (2024). A phase 3, randomized, controlled trial of resmetirom in NASH with liver fibrosis. *New England Journal of Medicine*, *390*(6), 497-509.

14. Newsome, P. N., Buchholtz, K., Cusi, K., Linder, M., Okanoue, T., Ratziu, V., Sanyal, A. J., Sejling, A. S., & Harrison, S. A. (2024). A placebo-controlled trial of subcutaneous semaglutide in nonalcoholic steatohepatitis. *New England Journal of Medicine*, *391*(24), 2304-2316.

15. National Institute for Health and Care Excellence. (2022). *Cost-effectiveness threshold analysis*. NICE.