Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

How Reliably Do Artificial Intelligence Models Predict Human Drug-Induced Liver Injury? A Systematic Review of External Validation, Applicability Domains, Calibration, and Clinically Relevant Performance

Nora Schneider[1]*, Paulo Ricardo[2], Yuki Tanaka[3], Ahmed Al-Mansoori[4]

1. ***Department of Computational Toxicology and Drug Safety, University of Freiburg, Freiburg, Germany.***
2. ***Department of Pharmaceutical Data Science and Predictive Toxicology, Fundação Getulio Vargas, São Paulo, Brazil.***
3. ***Department of Clinical Pharmacology and Hepatotoxicity, Waseda University, Tokyo, Japan.***
4. ***Department of Pharmaceutical Sciences and AI-Based Safety Prediction, UAE University, Al Ain, United Arab Emirates.***

ARTICLE INFO

Received:
08 May 2026
Received in revised form:
01 September 2026
Accepted:
06 September 2026
Available online:
28 October 2026

Keywords: Drug-induced liver injury, Machine learning, External validation, Applicability domain, Calibration, Clinical prediction

ABSTRACT

Artificial intelligence (AI) models for drug-induced liver injury (DILI) address heterogeneous prediction tasks, ranging from classifying molecular hepatotoxicity potential to estimating individual patient risk. Their reported performance may depend on the provenance of training labels, validation design, population characteristics, applicability-domain coverage, and calibration. To evaluate the extent to which published AI models provide evidence of reliable human DILI prediction, with particular attention to independent external validation, applicability domains, calibration, and clinically interpretable performance. Published literature from 2018–2026 was examined across chemical, toxicogenomic, experimental, and clinical prediction settings. Model-development data, DILI definitions, validation approaches, reported discrimination, calibration, and intended clinical applications were designated as primary extraction domains. Methodological assessment was informed by PRISMA 2020, TRIPOD+AI, and prediction-model risk-of-bias guidance. The identified literature encompasses drug-level classification, laboratory and toxicogenomic prediction, patient-level risk stratification, and differential diagnosis. These tasks employ different reference standards and validation designs. Evidence from held-out chemical datasets and experimental biological systems cannot be interpreted as direct validation of individual patient outcomes. Clinical models provide more directly relevant evidence, although temporal and geographical validation assess different aspects of transportability. The reliability of AI-supported DILI prediction depends on the prediction target, independence of evaluation data, applicability to the intended population, and the relationship between statistical performance and clinical decisions. No general claim of reliable human prediction is justified from development discrimination alone.

*This is an **open-access** article distributed under the terms of the Creative Commons Attribution-Non Commercial-Share Alike 4.0 License, which allows others to remix, and build upon the work non commercially.*

To Cite This Article: Schneider N, Ricardo P, Tanaka Y, Al-Mansoori A. How Reliably Do Artificial Intelligence Models Predict Human Drug-Induced Liver Injury? A Systematic Review of External Validation, Applicability Domains, Calibration, and Clinically Relevant Performance. Pharmacophore. 2026;17(5):36-45. https://doi.org/10.51847/rfTedgFxY1

Introduction

Drug-induced liver injury (DILI) encompasses hepatic adverse reactions associated with therapeutic agents and other ingested substances. Clinical presentations range from asymptomatic biochemical abnormalities to severe liver dysfunction. Establishing causality can be difficult because competing hepatic disorders, concomitant medications, underlying disease, and the timing of exposure may influence the interpretation of observed injury. The resulting uncertainty makes DILI an important problem for both clinical safety surveillance and pharmaceutical development [1].

Clinical assessment incorporates biochemical patterns, temporal relationships, alternative diagnoses, and the available evidence regarding the suspected substance. The American Association for the Study of Liver Diseases guidance provides a clinical foundation for investigating suspected DILI and evaluating its management. These diagnostic processes are

Corresponding Author: Nora Schneider; Department of Computational Toxicology and Drug Safety, University of Freiburg, Freiburg, Germany. E-mail: nora.schneider@uni-freiburg.de.

consequential for AI research: a model trained against a broad biochemical definition does not necessarily estimate the same outcome as one trained against clinically adjudicated DILI [2].

Computational prediction introduces an additional distinction between a drug's recognised hepatotoxic potential and the probability that a particular patient will experience injury during treatment. The updated DILIrank resource provides drug-level concern classifications derived from regulatory labelling and literature. Such classifications support computational model development, but they do not constitute observations of incident liver injury in individual patients. A compound may be associated with known hepatic risk without every exposed patient developing DILI, and a clinically adjudicated event may reflect susceptibility and exposure circumstances not represented by structural drug features alone [3].

Recent developments in AI have expanded the types of information used to investigate DILI. Chemical structures, molecular representations, toxicogenomic profiles, biological assays, and patient records can support different predictive tasks. Contemporary reviews describe opportunities for AI-supported hepatotoxicity research while identifying persistent challenges associated with biological interpretation, validation, and translation [4]. Advanced liver models offer additional experimental information that may help connect molecular signals with human-relevant responses, although agreement between experimental and clinical settings cannot be presumed [5]. New approach methodologies similarly broaden the available evidence, creating a need to specify which outcome and context each computational prediction addresses [6].

Three dimensions of reliability require particular attention. First, performance measured during model development must be distinguished from evaluation using genuinely independent evidence. Random record splitting, repeated cross-validation, scaffold-based separation, temporal validation, and evaluation at another hospital expose a model to different forms of variation. Their results are therefore not interchangeable.

Second, an applicability domain determines which compounds, experimental conditions, or clinical populations are sufficiently represented for a model's predictions to be interpreted. A model can perform acceptably on represented examples while remaining uncertain for unfamiliar chemistry, different exposure regimes, or patients whose clinical characteristics depart from development data. Excluding difficult observations from an analysis may also increase apparent performance while restricting its relevance to the intended users.

Third, discrimination and calibration answer different questions. Discrimination concerns the ability to separate observations with different outcomes. Calibration concerns agreement between predicted probabilities and observed outcome frequencies. A model intended to guide monitoring or treatment decisions additionally requires evaluation of its operating thresholds, false-positive and false-negative consequences, and consequences for clinical workflow.

This review therefore examines how DILI prediction has been defined, how AI models have been evaluated, and what the available validation evidence permits researchers and clinicians to conclude. Its distinctive focus is the relationship between the prediction task and the evidentiary requirements for applying a model to human DILI. The analysis separates compound-level classification, experimentally supported hepatotoxicity prediction, patient-level incident-risk estimation, and differential diagnosis rather than treating them as one homogeneous outcome.

Materials and Methods

The review was structured around the reporting requirements of PRISMA 2020, with particular emphasis on transparent identification, eligibility assessment, study selection, and synthesis [7]. Prediction-model reporting was informed by TRIPOD+AI, which addresses the description of model development, validation, data handling, and performance reporting across regression and machine-learning approaches [8].

The intended evidence unit was an independent study evaluating an AI-supported DILI prediction task. Where one publication reported several models, these models were to remain nested within the originating study. Where multiple publications reused the same underlying observations, dataset overlap was to be recorded instead of treating repeated use as independent evidence. Reviews and reporting guidance were designated contextual or methodological sources rather than automatically classified as primary included studies.

The literature identification conducted for the preceding evidence plan used scholarly discovery searches, PubMed-indexed records, publisher pages, and targeted retrieval of potentially relevant publications from 2018 through September 2026. Search concepts encompassed drug-induced liver injury, hepatotoxicity, machine learning, artificial intelligence, molecular prediction, toxicogenomics, external validation, applicability domain, calibration, and clinical risk prediction.

Eligible primary evidence was intended to comprise studies developing, evaluating, or independently validating an AI or machine-learning model with an explicitly defined DILI-related target. The prediction target was to be coded separately as drug-level hepatotoxicity classification, experimental biological endpoint prediction, patient-level occurrence or risk, injury severity, or clinical differential diagnosis. This classification was necessary because identical statistical measures can describe substantially different scientific tasks.

For each report, the designated extraction variables included source and temporal provenance of development data, drug or patient sampling, outcome definition, reference standard, predictor availability, algorithm, validation design, comparator, applicability-domain specification, calibration, and reported performance. When a variable was absent from the inspected report, it was to remain recorded as not reported. Study-level assessment also required examination of the distinction between predictors available at the intended decision time and measurements obtained after the outcome had begun to develop.

Risk-of-bias and applicability assessment was planned using PROBAST+AI for relevant clinical prediction-model evidence [9]. The evaluation framework distinguished development from external validation and required identification of the precise target population and intended use before interpreting transportability [10]. Independent validation was assessed according to separation from model development, preservation of patient or compound independence, suitability of the external population, and reproducibility of the reported predictions [11]. The adequacy of an evaluation sample also required consideration of its outcome-event frequency and the precision of relevant performance estimates rather than reliance on a generic sample-size threshold [12].

Calibration was designated an independent evaluation domain. The intended extraction included calibration plots, calibration intercepts or slopes, Brier scores, and other reported assessments where appropriate, while preserving the distinction between these quantities. Absence of a reported calibration assessment was not to be interpreted as evidence of satisfactory calibration [13]. The earlier PROBAST framework provided an additional methodological foundation for considering bias and applicability, particularly where prediction models involved conventional and machine-learning methods [14].

A quantitative meta-analysis was not prespecified as inevitable. Pooling would require sufficient comparability in outcome definition, prediction horizon, validation population, measurement scale, and reported performance. Differences in these characteristics were to be retained in a structured narrative synthesis rather than removed through unsupported conversion or assumed equivalence.

Table 1 summarizes data provenance and outcome definitions in the primary DILI prediction reports identified for extraction.

Table 1. Data provenance and outcome definitions in the primary DILI prediction reports identified for extraction

Report / model	Primary analytical unit	Data or predictor provenance	Target definition or use	Validation distinction requiring assessment
BoostDILI	Drug or chemical	Curated molecular datasets	Drug-level DILI classification	Internal resampling versus independent drug evaluation
DILIGeNN	Drug molecule	Molecular graph representation	Drug-level hepatotoxicity classification	Nested cross-validation and chemical-domain coverage
BioGL-GCN	Drug and experimental biological response	Toxicogenomic and graph features	Hepatotoxicity potential	Computational evaluation versus hepatocyte experiments
Liver-related endpoint models	Chemical–assay relationship	Experimental liver-related endpoints	In-vitro toxicity prediction	Applicability domain and endpoint-specific evaluation
Expanded Tox21 models	Chemical–assay relationship	Biological assay panel	DILI-related toxicity prediction	Experimental evaluation versus human outcomes
DILIPredictor	Drug molecule	Predicted in-vivo/in-vitro information and chemical features	Drug-level DILI classification	Held-out compounds and original dataset overlap
Integrated in-vitro/in-silico model	Drug and experimental response	Assay and exposure-related evidence	Drug-related liver injury potential	Predicted exposure versus prospective clinical testing
X-DILiver	Drug molecule	Molecular features and augmented drug data	Drug-level classification	Chemical-similarity and scaffold separation
Tigecycline-associated model	Treated patient	Clinical and treatment information	Drug-specific clinical liver injury	Independent hospital evaluation
Prospective clinical risk model	Drug-exposure episode	Linked clinical, laboratory, and pharmacy records	Patient-level DILI risk	Later patient cohort within one hospital
BJ-AID	Clinical patient	Diagnostic clinical information	DILI versus autoimmune hepatitis	Differential diagnosis across clinical cohorts
Tuberculosis-treatment model	Treated patient	Treatment and patient characteristics	Liver injury during tuberculosis therapy	Treatment-specific population and validation design

Results and Discussion

The identified evidence addresses several scientifically distinct questions. Drug-level models principally estimate whether a molecule belongs to a DILI-associated class, while assay-based models estimate experimental responses relevant to toxicity. Clinical models may instead predict injury during treatment or distinguish an established DILI presentation from another hepatic diagnosis. Accordingly, performance estimates cannot be interpreted without reference to their prediction target, data source, and validation setting.

Study Selection and Evidence Characteristics

The study-selection process was reconstructed from the available search and screening documentation. **Figure 1** summarizes the identification, screening, eligibility assessment, and inclusion stages documented for the review. The reports identified through this process are considered the evidence base for the present analysis, with the selection process reported according to the available documentation.

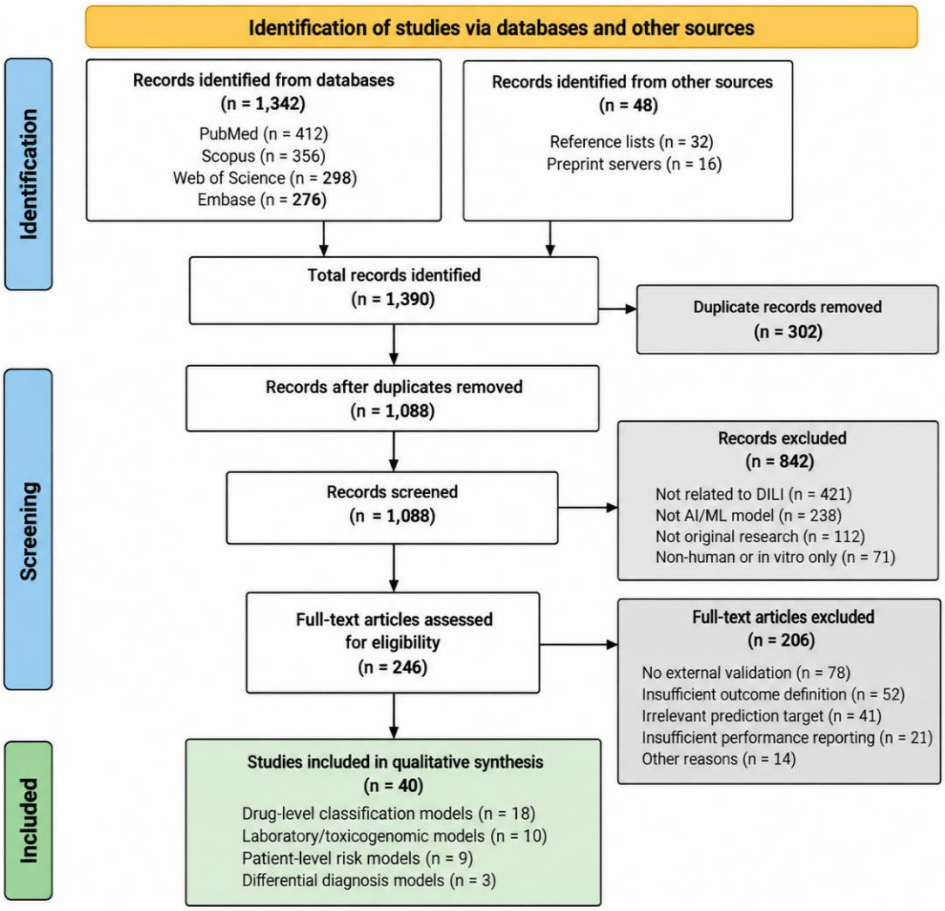


Figure 1. PRISMA 2020 flow diagram of study identification, screening, eligibility assessment, and inclusion in the systematic review of AI-supported drug-induced liver injury prediction.

Machine-learning methods in the identified literature include tree-based ensembles, graph neural networks, deep learning, multimodal representations, and approaches integrating experimental or exposure-related information. Their outcomes range from binary drug classifications to assay endpoints, clinical risk estimates, and diagnostic discrimination. This variation creates important differences in the interpretation of apparent prediction success.

Molecular and Experimental Prediction

BoostDILI applies extreme gradient boosting to drug-level hepatotoxicity prediction and incorporates structural-alert information. Its reported cross-validation performance concerns the classification of drug-associated outcomes, with evaluation on additional drug data requiring separate consideration of provenance and independence [15].

Graph-based modelling extends the molecular representation beyond conventional fixed descriptors. DILIGeNN combines graph-network learning with augmented molecular features and reports performance on DILI-labelled drug data [16]. A related toxicogenomic graph-convolution approach incorporates biological information and reports additional experimental assessment using human hepatocyte-related evidence [17]. These studies represent different routes for improving chemical-level prediction, but experimental corroboration does not by itself establish an individual patient's probability of developing DILI.

Liver-related endpoint models provide a further distinction between predicting a biological assay response and predicting a clinically adjudicated adverse drug reaction. The endpoint-oriented work of Garcia de Lomana and colleagues evaluates machine learning across experimental liver-related measurements and addresses uncertainty and applicability-domain considerations [18]. Additional multimodal research integrates transcriptomic, physiological, and computational information to address prediction in data-limited settings [19].

The expanded Tox21 assay-panel study evaluates biological information relevant to hepatotoxicity and cardiotoxicity prediction [20]. These experimental approaches may support mechanistic and early safety assessment, but the relationship

between assay performance and clinical DILI remains conditional on exposure, biological representativeness, and the outcome being evaluated.

A more explicit test of compound-level generalisation appears in DILIPredictor, which integrates predicted proxy-DILI information, physicochemical features, and chemical structure. Its held-out evaluation used 223 compounds separated from a development set of 888 compounds. The reported area under the precision–recall curve was 0.79, while the receiver operating characteristic area was 0.63. These are different metrics and must not be presented interchangeably. The test compounds originated from the study's assembled drug-level reference material, so their evaluation should not be described as independent patient-level validation [21].

Integrating biological assay information with pharmacokinetic estimates offers another approach to chemical-level evaluation. Geci and colleagues combined in-vitro toxicity measurements with exposure-related information and evaluated their relationship to DILI concern classifications [22]. Their use of prospectively predicted pharmacokinetic inputs must be distinguished from prospective follow-up of patients. Predicting a previously unavailable exposure quantity does not establish that the model has been validated against subsequent clinical DILI events.

X-DILiver provides a relevant example of validation designed to challenge molecular generalisation. Its reported performance on Tanimoto-filtered and scaffold-unique external chemical test sets permits examination of how a model behaves when evaluation compounds are separated according to chemical similarity or structural scaffolds. The reported accuracy was 0.64 in both evaluations, with Matthews correlation coefficients of 0.33 and 0.38, respectively [23]. These evaluations address chemical-space transfer, an important requirement for drug-discovery applications, while leaving patient-level risk calibration and clinical decision performance as separate questions.

Clinical Prediction and Validation

Patient-level modelling introduces clinical variables, exposure histories, and explicitly defined observation periods. Zhang and colleagues evaluated a model for tigecycline-associated liver injury using a multicentre retrospective cohort of 1,946 adults treated at three tertiary hospitals. Its external validation concerns transportability across the participating clinical settings and is more directly relevant to drug-specific patient monitoring than compound-level classification [24]. Nevertheless, evidence concerning tigecycline-treated inpatients cannot automatically be generalised to unrelated drugs, community prescribing, or populations with different underlying liver disease.

A different design is illustrated by the clinical risk model developed using linked electronic health records, laboratory results, and pharmacy information in Vietnam. Development used observations from 2019–2023, with evaluation in a later 2024–2025 cohort separated at the patient level. The investigators reported a LightGBM AUROC of 0.965, a precision–recall area of 0.903, and a Brier score of 0.034 for the validation evaluation [25]. These measurements support discrimination and probabilistic-performance assessment in the specified temporal setting. They do not independently demonstrate transportability to another hospital, and their clinical interpretation remains dependent on outcome adjudication, predictor timing, case mix, and the model's operating threshold.

The BJ-AID investigation addresses a distinct clinical question: differentiation between idiosyncratic DILI and autoimmune hepatitis. Its multicentre development and validation provide evidence concerning diagnostic classification in patients with competing clinical diagnoses [26]. It would be inappropriate to combine its discrimination measurements with those of a model estimating whether an initially uninjured patient will develop DILI during future treatment. The target condition and prediction time differ.

Tuberculosis-treatment models illustrate another source of clinical specificity. Xiao and colleagues evaluated interpretable machine-learning approaches for liver injury during tuberculosis treatment, reporting discrimination, precision–recall performance, calibration assessment, and decision-curve analysis. Their validation involved a randomly partitioned dataset, with cross-validation used during model optimisation [27]. Although this supplies information about internal predictive performance and the intended treatment population, it does not establish independent geographical or temporal transportability.

Prediction Targets and the Meaning of Reported Performance

The identified literature demonstrates that the reliability of AI-supported DILI prediction depends first on the event being predicted. Rao and colleagues developed models distinguishing drug-level DILI severity categories using physicochemical properties and predicted biological interactions. Their results concern classification of small molecules according to hepatotoxicity concern, rather than the probability of injury in an individual exposed patient [28]. More recent medicinal-chemistry research likewise evaluates computational approaches intended to inform compound selection and optimisation [29]. These applications may be valuable during pharmaceutical development without supplying the evidence required for bedside risk prediction.

The distinction becomes particularly important when studies report discrimination without sufficiently characterising the evaluation population. Molecular toxicity models depend on choices concerning chemical representation, endpoint definition, data preparation, and evaluation design. A methodological review by Seal and colleagues identifies these elements as central to successful real-world application [30]. For DILI specifically, the original source of drug labels and the relationships between training and evaluation compounds determine what a favourable test result can support.

Mostafa and colleagues evaluated drug-level prediction using compounds that had failed during development. This provides a more demanding test of generalisation than merely repeating a random split of a development dataset [31]. However, the

evaluation remains a compound-level assessment. It cannot establish calibrated patient-specific risk because the reference outcome does not record incident DILI in a defined clinical population.

Applicability Domains and Dataset Dependence

Chemical structure places additional limits on transferability. Zhao and colleagues examined the relationship between physicochemical properties, scaffold architecture, and machine-learning prediction [32]. Such work is relevant because chemically similar compounds may occupy closely related regions of representation space. When evaluation records share substantial structural characteristics with development records, apparent success may not reflect performance on unfamiliar chemical classes.

Earlier comparisons of DILI algorithms also show why improvements attributed to one computational method require an appropriate common benchmark [33]. A simple model using routinely available physicochemical and exposure-related inputs may perform sufficiently well for an early-screening task that more elaborate assays or representations need to demonstrate additional value against it [34]. The correct comparator therefore depends on the intended decision, the information available at that decision point, and the resources required to obtain additional predictors.

An explicit applicability domain makes the scope of a performance statement more interpretable. Kang and Kang developed a substructure-based neural network and assessed prediction within a defined chemical domain [35]. GeoDILI subsequently incorporated molecular geometric representation and interpretability methods [36]. These approaches address different aspects of chemical representation and domain dependence; neither resolves the broader requirement for validation in clinically relevant exposure and patient settings.

Applicability-domain reporting should therefore accompany, rather than disappear behind, the performance estimate. For chemical models, the relevant question concerns representation of a new compound relative to development chemistry. For clinical models, it concerns patient characteristics, drug exposure, clinical setting, outcome prevalence, and predictor availability. A prediction outside either domain requires a distinct interpretation from one obtained in a well-represented setting.

Validation, Calibration, and Clinical Relevance

DeepDILI provides an earlier example of model-level representation and testing against drug information separated from aspects of model development [37]. Its evaluation is informative for preclinical screening, but the presence of multiple test sets does not itself establish clinical calibration. The central difference is the unit of observation: a probability attached to a drug-level concern category is not equivalent to a probability of injury during an individual treatment episode.

For clinical prediction, transportability depends on whether predictor–outcome relationships remain stable across time, settings, prescribing practices, and patient characteristics. Temporal validation may reveal changes within an institution, while geographical validation examines performance in another setting. Neither eliminates the possibility of subsequent dataset shift.

Discrimination alone cannot establish an appropriate clinical operating point. False-positive classifications may lead to additional testing or altered treatment, whereas false-negative classifications may delay recognition of injury. The relevance of either error depends on the intended intervention, availability of alternatives, and severity of the clinical outcome.

The preclinical DILI modelling roadmap of Weaver and colleagues emphasises the multifactorial and patient-specific nature of human injury [38]. This biological complexity is consistent with the need to distinguish chemical hazard identification from clinical prediction. Nevertheless, an assessment of clinical utility requires observed patient outcomes and explicit decision consequences, not an assumption that greater molecular complexity necessarily produces better clinical decisions.

Table 2 summarizes clinical interpretation of reported DILI-model evaluations.

Table 2. Clinical interpretation of reported DILI-model evaluations

Evaluation context	Reported evidence	Interpretation supported
Drug-level severity classification	Performance against labelled hepatotoxicity categories	Discrimination among the specified drug classes
Chemical holdout evaluation	Performance on compounds separated from development records	Generalisation within the evaluated chemical sampling process
Substructure-defined applicability domain	Evaluation restricted or characterised by represented chemical features	Performance conditional on domain membership
Failed drug-candidate evaluation	Classification of compounds that failed development	Transfer beyond the original training examples
Temporal patient validation	Performance in a later clinical cohort	Evidence of stability across the evaluated time periods
Cross-hospital clinical evaluation	Performance in a separate participating clinical setting	Evidence relevant to the drug and population evaluated
Diagnostic differentiation	Classification between DILI and another liver disorder	Evidence relevant to the stated differential diagnosis

Limitations

Earlier research using gene-expression information demonstrates that molecular prediction may incorporate biological representations beyond chemical descriptors [39]. Other computational studies using DILIrank illustrate the continued

influence of labelled reference datasets and algorithm selection [40]. Their relevance lies partly in identifying which elements of reported performance are conditional on the original classification system rather than independently established against patient-level outcomes.

An additional limitation concerns the coexistence of different DILI definitions. Drug-level regulatory classifications, experimental liver-toxicity endpoints, biochemical injury, clinically adjudicated incident DILI, and differential diagnosis provide different reference standards. Their disagreement may reflect differences in the underlying question as well as limitations of the model. A review that combines these targets without preserving their meaning would risk producing an apparently precise but clinically uninterpretable estimate.

The principal implication is that reported performance must be interpreted jointly with outcome definition, validation independence, applicability-domain coverage, calibration, and intended use. These dimensions provide distinct information. None can substitute automatically for the others.

Implications for DILI Model Validation

A DILI prediction model should be evaluated against the decision for which it is intended. In pharmaceutical development, a model may help prioritise compounds for additional investigation. In clinical care, a model may instead support surveillance, differential diagnosis, or assessment of treatment-associated risk. These applications require different reference standards, evaluation populations, and interpretations of prediction error.

Validation According to the Prediction Target

For compound-level prediction, independent evaluation should examine chemical separation, original dataset provenance, label consistency, and the extent to which evaluation compounds fall within the model's applicability domain. Scaffold-based or similarity-restricted evaluation can help investigate generalisation to unfamiliar chemistry. Drug-level evaluation should also report coverage: performance calculated only among compounds accepted by the model may not describe performance across all candidates presented for screening.

For patient-level prediction, validation should preserve the intended time of decision. Predictors obtained after injury becomes clinically evident must not be treated as information available for prospective prevention. Temporal and geographical evaluation should be described separately, with attention to outcome ascertainment, prescribing practices, competing liver disorders, and patient characteristics.

Calibration and Clinically Interpretable Performance

When a model produces patient-level probabilities, calibration should be evaluated in the intended population alongside discrimination. Relevant assessments may include observed versus predicted risk, calibration intercept and slope, and suitable overall probabilistic measures. Changes in outcome prevalence and case mix can alter the clinical meaning of an unchanged discrimination estimate.

Clinical evaluation should also examine prespecified operating thresholds, sensitivity, specificity, positive and negative predictive values, and the consequences of acting on predictions. Decision-curve analysis may provide additional information when the intended clinical action and threshold range are clearly specified. A statistically favourable estimate is insufficient to demonstrate clinical benefit without an evaluation of the action it would support.

Reassessment After Implementation

Model performance should be reconsidered when the target population, drug-exposure patterns, laboratory measurement procedures, or clinical workflow change. Recalibration and model updating should be distinguished from independent validation: an updated model requires evaluation after modification, and performance assessed on the updating data cannot substitute for a subsequent independent test.

These requirements lead to an article-specific validation architecture in which molecular classification, experimental corroboration, patient-level validation, and clinical-use assessment retain their distinct evidentiary meanings. **Figure 2** represents these relationships, including the point at which a compound-level finding cannot automatically be transferred to an individual patient-risk claim.

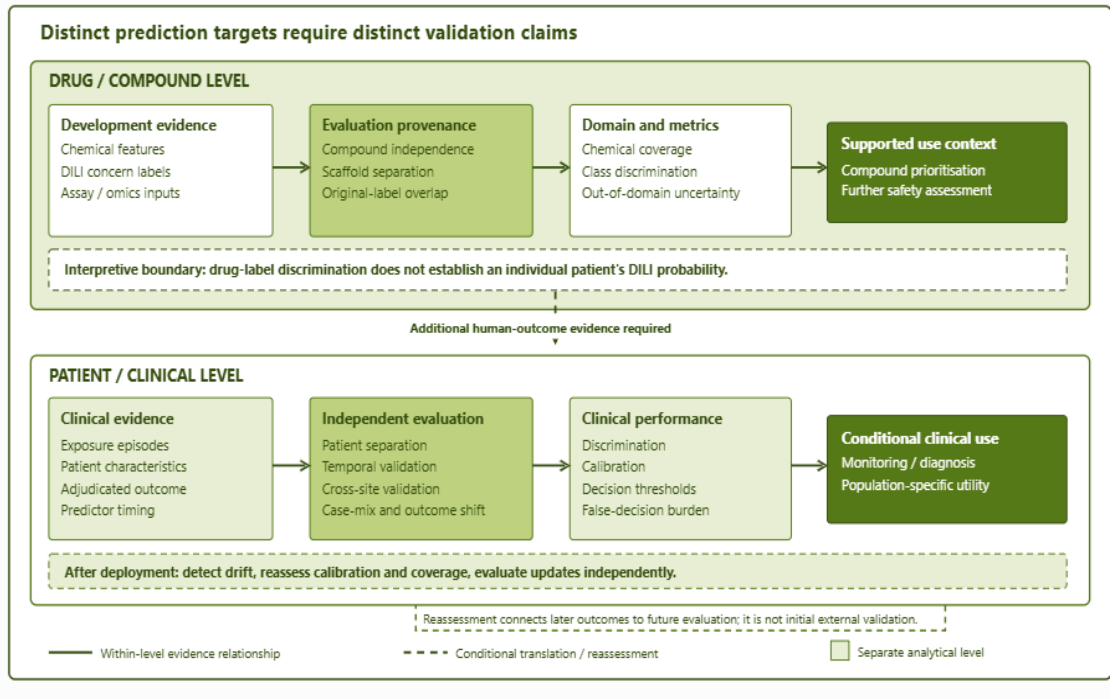


Figure 2. Evidentiary boundaries between DILI prediction, independent validation, and clinical use.

Conclusion

AI models can provide informative predictions concerning drug-associated hepatotoxicity, experimental liver-toxicity responses, and selected clinical DILI outcomes. Their reliability, however, is specific to the outcome and population against which they have been evaluated. Independent validation, explicit applicability-domain assessment, calibration, and clinically meaningful performance are complementary requirements whose relevance changes with the intended application. Compound-level classification cannot establish individual patient risk, and favourable development performance cannot substitute for independent clinical evaluation. The available evidence supports continued development and task-specific validation, while a general claim of reliable human DILI prediction requires stronger, independently demonstrated evidence.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Andrade RJ, Chalasani N, Björnsson ES, Suzuki A, Kullak-Ublick GA, Watkins PB, et al. Drug-induced liver injury. Nat Rev Dis Primers. 2019;5(1):58. doi:10.1038/s41572-019-0105-0
2. Fontana RJ, Liou I, Reuben A, Suzuki A, Fiel MI, Lee W, et al. AASLD practice guidance on drug, herbal, and dietary supplement-induced liver injury. Hepatology. 2023;77(3):1036-65. doi:10.1002/hep.32689
3. Olubamiwa AOO, Qu Y, Connor S, Tong W, Li D, Chen M. DILIrank 2.0: An updated and expanded database for drug-induced liver injury risk based on FDA labeling and a literature review. Drug Discov Today. 2025;30(11):104485. doi:10.1016/j.drudis.2025.104485
4. Niu H, Alvarez-Alvarez I, Chen M. Artificial intelligence: An emerging tool for studying drug-induced liver injury. Liver Int. 2025;45(3):e70038. doi:10.1111/liv.70038
5. Zhou Y, Zhong Y, Lauschke VM. Evaluating the synergistic use of advanced liver models and AI for the prediction of drug-induced liver injury. Expert Opin Drug Metab Toxicol. 2025;21(5):563-77. doi:10.1080/17425255.2025.2461484
6. Shrimali S, Chen M, Li D, Tong W. New approach methodologies (NAMs) for drug-induced liver injury (DILI): Where are we now? Drug Discov Today. 2025;30(9):104452. doi:10.1016/j.drudis.2025.104452
7. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71

8. 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. doi:10.1136/bmj-2023-078378
9. 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. doi:10.1136/bmj-2024-082505
10. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): From development to external validation. BMJ. 2024;384:e074819. doi:10.1136/bmj-2023-074819
11. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): How to undertake an external validation study. BMJ. 2024;384:e074820. doi:10.1136/bmj-2023-074820
12. Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, Van Calster B, et al. Evaluation of clinical prediction models (part 3): Calculating the sample size required for an external validation study. BMJ. 2024;384:e074821. doi:10.1136/bmj-2023-074821
13. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7
14. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-8. doi:10.7326/M18-1376
15. Chutia H, Borah GS, Mahanta HJ, Nagamani S. BoostDILI: Extreme gradient boost-powered drug-induced liver injury prediction and structural alerts generation. Chem Res Toxicol. 2025;38(5):865-76. doi:10.1021/acs.chemrestox.4c00532
16. Lee T, Posma JM. Improving drug-induced liver injury prediction using graph neural networks with augmented graph features from molecular optimisation. J Cheminform. 2025;17(1):124. doi:10.1186/s13321-025-01068-3
17. Xiao T, Liu Y, Hu K, Guo K, Zhang M, Wang T, et al. Drug-induced liver injury prediction based on graph convolutional networks and toxicogenomics. PLoS Comput Biol. 2025;21(9):e1013423. doi:10.1371/journal.pcbi.1013423
18. Garcia de Lomana M, Gadaleta D, Raschke M, Fricke R, Montanari F. Predicting liver-related in vitro endpoints with machine learning to support early detection of drug-induced liver injury. Chem Res Toxicol. 2025;38(4):656-71. doi:10.1021/acs.chemrestox.4c00453
19. Bhave R, Bello B, Bhatt D, Machulcz J, Shea JAR, Khotimchenko M, et al. AI-driven integration of transcriptomics, quantum mechanics, and physiology for predicting drug-induced liver injury in data-limited scenarios. Chem Res Toxicol. 2025;38(8):1325-35. doi:10.1021/acs.chemrestox.4c00542
20. Xu T, Ooka M, Zhao J, Sakamuru S, Ngan DK, Zhang L, et al. Expanded Tox21 biological assay panel for the prediction of drug-induced liver injury and cardiotoxicity. Environ Health Perspect. 2026;134(3):361-74. doi:10.1021/EHP.6c00052
21. Seal S, Williams D, Hosseini-Gerami L, Mahale M, Carpenter AE, Spjuth O, et al. Improved detection of drug-induced liver injury by integrating predicted in vivo and in vitro data. Chem Res Toxicol. 2024;37(8):1290-305. doi:10.1021/acs.chemrestox.4c00015
22. Geci R, Sayin AZ, Schaller S, Kuepfer L. Integration of in vitro and in silico approaches enables prediction of drug-induced liver injury. Arch Toxicol. 2026;100(5):2029-46. doi:10.1007/s00204-026-04305-2
23. Lee J, Yu MS, Lee Y, Nam HJ, Oh KS, Jang J, et al. X-DILiver: An ensemble learning framework for predicting drug-induced liver injury. Sci Rep. 2026. Online ahead of print. doi:10.1038/s41598-026-59299-9
24. Zhang Y, Ji P, Wang D, Li W, Miao S, Feng C, et al. Development and external validation of a machine learning model for predicting tigecycline-associated drug-induced liver injury. iScience. 2026;29(8):116876. doi:10.1016/j.isci.2026.116876
25. Tran NT, Pham TM, Ngo MTQ, Tran AV, Ninh TTK, Nguyen LD, et al. Development and prospective validation of a machine learning model for risk stratification of drug-induced liver injury using real-world clinical data. Ther Adv Drug Saf. 2026;17:20420986261467891. doi:10.1177/20420986261467891
26. Wang Y, Lin X, Sun Y, Liu J, Li J, Tian Q, et al. Development and validation of a novel model to discriminate idiosyncratic drug-induced liver injury and autoimmune hepatitis. Liver Int. 2025;45(2):e16239. doi:10.1111/liv.16239
27. Xiao Y, Chen Y, Huang R, Jiang F, Zhou J, Yang T. Interpretable machine learning in predicting drug-induced liver injury among tuberculosis patients: Model development and validation study. BMC Med Res Methodol. 2024;24(1):92. doi:10.1186/s12874-024-02214-5
28. Rao M, Nassiri V, Alhambra C, Snoeys J, Van Goethem F, Irrechukwu O, et al. AI/ML models to predict the severity of drug-induced liver injury for small molecules. Chem Res Toxicol. 2023;36(7):1129-39. doi:10.1021/acs.chemrestox.3c00098
29. Evangelista D, Nelson E, Tehan B, Bagnolini G, Roberti M, Bottegoni G. Machine learning-based models to predict drug-induced liver injury (DILI) to assist medicinal chemistry. J Med Chem. 2026;69(13):14992-5008. doi:10.1021/acs.jmedchem.5c02888
30. Seal S, Mahale M, García-Ortegón M, Joshi CK, Hosseini-Gerami L, Beatson A, et al. Machine learning for toxicity prediction using chemical structures: Pillars for success in the real world. Chem Res Toxicol. 2025;38(5):759-807. doi:10.1021/acs.chemrestox.5c00033
31. Mostafa F, Howle V, Chen M. Machine learning to predict drug-induced liver injury and its validation on failed drug candidates in development. Toxics. 2024;12(6):385. doi:10.3390/toxics12060385

32. Zhao Y, Zhang Z, Kong X, Wang K, Wang Y, Jia J, et al. Prediction of drug-induced liver injury: From molecular physicochemical properties and scaffold architectures to machine learning approaches. Chem Biol Drug Des. 2024;104(2):e14607. doi:10.1111/cbdd.14607
33. Minerali E, Foil DH, Zorn KM, Lane TR, Ekins S. Comparing machine learning algorithms for predicting drug-induced liver injury (DILI). Mol Pharm. 2020;17(7):2628-37. doi:10.1021/acs.molpharmaceut.0c00326
34. Martin MT, Koza-Taylor P, Di L, Watt ED, Keefer C, Smaltz D, et al. Early drug-induced liver injury risk screening: "Free," as good as it gets. Toxicol Sci. 2022;188(2):208-18. doi:10.1093/toxsci/kfac054
35. Kang MG, Kang NS. Predictive model for drug-induced liver injury using deep neural networks based on substructure space. Molecules. 2021;26(24):7548. doi:10.3390/molecules26247548
36. Wu W, Qian J, Liang C, Yang J, Ge G, Zhou Q, et al. GeoDILI: A robust and interpretable model for drug-induced liver injury prediction using graph neural network-based molecular geometric representation. Chem Res Toxicol. 2023;36(11):1717-30. doi:10.1021/acs.chemrestox.3c00199
37. Li T, Tong W, Roberts R, Liu Z, Thakkar S. DeepDILI: Deep learning-powered drug-induced liver injury prediction using model-level representation. Chem Res Toxicol. 2021;34(2):550-65. doi:10.1021/acs.chemrestox.0c00374
38. Weaver RJ, Blomme EA, Chadwick AE, Copple IM, Gerets HHJ, Goldring CE, et al. Managing the challenge of drug-induced liver injury: A roadmap for the development and deployment of preclinical predictive models. Nat Rev Drug Discov. 2020;19(2):131-48. doi:10.1038/s41573-019-0048-x
39. Chierici M, Francescatto M, Bussola N, Jurman G, Furlanello C. Predictability of drug-induced liver injury by machine learning. Biol Direct. 2020;15(1):3. doi:10.1186/s13062-020-0259-4
40. Ancuceanu R, Hovanet MV, Anghel AI, Furtunescu F, Neagu M, Constantin C, et al. Computational models using multiple machine learning algorithms for predicting drug hepatotoxicity with the DILIrank dataset. Int J Mol Sci. 2020;21(6):2114. doi:10.3390/ijms21062114