Scientific journal of Research studies in Future Basic Sciences and medical sciences, 2024, Vol. 4, No. 2, pp. 21-29

Scientific journal of Research studies

in Future Basic Sciences and medical sciences

www.journalhi.com/sci/

Development and Evaluation of a Machine-Learning Model for Predicting Autoimmune Hepatitis Relapse After Treatment Withdrawal Based on Clinical and Laboratory Factors

Mohammad Karami Horestani [1,*]

1- Assistant Professor of Gastroenterology and Hepatology. Department of Internal Medicine, Clinical Research Development Unit, Hajar Hospital, Shahrekord University of Medical Sciences, Shahrekord. Email: karami.m@skums.ac.ir

* Corresponding Author

Abstract

Background: Relapse after treatment withdrawal remains a major challenge in autoimmune hepatitis (AIH), particularly because conventional risk assessment may not adequately capture interactions among multiple clinical and laboratory factors. This study aimed to develop and evaluate a machine-learning framework for predicting AIH relapse after treatment withdrawal using routinely available clinical and biochemical variables.

Methods: The proposed framework incorporates demographic, clinical, biochemical, disease-related, and treatment-related variables associated with relapse risk. Machine-learning algorithms were considered to identify nonlinear relationships and interactions among candidate predictors. The framework includes data preprocessing, feature selection, model development, internal validation, discrimination, calibration, and assessment of clinical applicability.

Results: The proposed approach provides a structured method for integrating multiple predictors into individualized relapse-risk estimation. Particular emphasis was placed on reducing overfitting, maintaining interpretability, and evaluating both discrimination and calibration. The framework can also facilitate comparison of alternative machine-learning algorithms and identification of clinically relevant predictors.

Conclusion: Machine-learning methods may provide a systematic approach for individualized prediction of AIH relapse following treatment withdrawal. Before clinical implementation, the resulting model should undergo rigorous external validation in independent patient populations.

Keywords: Autoimmune hepatitis; Relapse; Treatment withdrawal; Machine learning; Clinical prediction model

1- Introduction

Autoimmune hepatitis (AIH) is a chronic inflammatory liver disease caused by an abnormal immune response against hepatic tissue. Its clinical presentation ranges from asymptomatic biochemical abnormalities to acute hepatitis and advanced chronic liver disease. The disease is characterized by considerable heterogeneity in clinical course, biochemical activity, treatment response, and long-term prognosis, making individualized management an important component of clinical care [2,3,5].

The principal therapeutic objective in AIH is sustained biochemical remission and prevention of progressive hepatic injury. Corticosteroids, alone or in combination with other immunosuppressive agents, remain the foundation of treatment. Although prolonged treatment can achieve biochemical control in many patients, treatment withdrawal is considered only in carefully selected individuals who have maintained an adequate and sustained response. Current clinical guidance emphasizes the importance of biochemical remission and continued monitoring after withdrawal because disease activity may recur even after an apparently satisfactory treatment response [2,4,6,10-12].

Relapse following treatment withdrawal represents a clinically relevant problem because recurrent inflammatory activity can result in renewed biochemical abnormalities, additional treatment requirements, and potentially progressive liver injury. Previous clinical research has investigated whether demographic characteristics, duration of remission, aminotransferase levels, immunoglobulin G concentrations, treatment history, and other clinical variables can identify patients at greater risk of relapse. However, the predictive contribution of individual variables may be limited because relapse is likely influenced by multiple interacting biological and clinical factors [1,9].

Traditional statistical approaches generally focus on predefined associations between selected predictors and outcomes. Machine-learning methods provide an alternative framework capable of modeling nonlinear relationships and complex interactions among multiple variables. When appropriately developed and validated, these methods may facilitate individualized risk estimation while identifying combinations of predictors that are difficult to capture through conventional single-factor assessment.

Despite growing interest in artificial intelligence in clinical prediction, the development of a clinically useful model requires rigorous attention to predictor selection, overfitting, internal validation, discrimination, calibration,

and applicability. Contemporary methodological standards such as TRIPOD+AI and PROBAST+AI emphasize transparent reporting and systematic assessment of bias in prediction models based on machine-learning approaches [13,14].

Accordingly, this study focuses on the development and evaluation of a machine-learning framework for predicting relapse after treatment withdrawal in patients with AIH using clinical and laboratory factors that can be routinely assessed in clinical practice. The proposed approach seeks to integrate multiple predictors into an individualized risk-estimation framework while maintaining methodological transparency and clinical interpretability.

2- Problem Statement

Treatment withdrawal in autoimmune hepatitis (AIH) presents a clinically important prediction problem because biochemical remission before withdrawal does not necessarily indicate permanent disease control. A proportion of patients experience disease relapse after immunosuppressive therapy is discontinued, requiring renewed treatment and prolonged biochemical monitoring. The difficulty lies not only in identifying individual risk factors, but also in determining how multiple clinical and laboratory characteristics interact to influence the probability of relapse [1,2,11,12].

Previous investigations have identified several potential predictors of relapse, including biochemical activity, immunoglobulin G levels, duration and stability of remission, and immunological characteristics. However, these factors do not contribute independently or uniformly across patients. The biological heterogeneity of AIH and variation in treatment history create a multidimensional prediction problem in which conventional threshold-based assessment may have limited ability to estimate individualized risk [1,9].

A further challenge is that treatment withdrawal decisions require a balance between avoiding unnecessary long-term immunosuppression and minimizing the risk of recurrent hepatic inflammation. Current clinical recommendations therefore emphasize careful patient selection and post-withdrawal monitoring, but they do not provide a universally applicable individualized prediction model capable of integrating multiple routinely available variables into a single relapse-risk estimate [2,10-12].

Machine-learning techniques offer a potential methodological solution because they can evaluate nonlinear associations and interactions among several predictors simultaneously. Nevertheless, clinical usefulness depends on more than algorithmic complexity. Poorly designed prediction models may be affected by overfitting, inadequate validation, predictor selection bias, or poor calibration. Recent methodological frameworks specifically emphasize transparent development, validation, reporting, and assessment of applicability for machine-learning-based clinical prediction models [13,14].

The central research problem, therefore, is the absence of a sufficiently structured and clinically interpretable framework for integrating routinely obtainable clinical and laboratory characteristics into individualized prediction of AIH relapse following treatment withdrawal. Addressing this gap requires a modeling strategy that combines clinically meaningful predictors with rigorous machine-learning methodology and evaluates model performance from both statistical and clinical perspectives.

3- Materials and Methods

3.1. Study Design and Analytical Framework

This study was designed as a clinical prediction-modeling investigation aimed at estimating the probability of relapse after treatment withdrawal in patients with autoimmune hepatitis (AIH). The analytical framework was structured to integrate routinely available demographic, clinical, biochemical, disease-related, and treatment-related variables into a machine-learning prediction model. The overall methodological structure followed contemporary recommendations for the development and reporting of prediction models using artificial intelligence and machine-learning techniques [13,14].

The target outcome was relapse of AIH following withdrawal of immunosuppressive therapy. Relapse was defined according to the clinical and biochemical criteria reported in the selected AIH literature and contemporary clinical guidelines. Particular attention was given to biochemical evidence of renewed disease activity following treatment discontinuation, because biochemical monitoring represents a central component of post-withdrawal surveillance [1,2,11,12].

3.2. Predictor Selection

Candidate predictors were selected on the basis of their clinical relevance to AIH disease activity, treatment response, and relapse following treatment withdrawal. Variables considered within the modeling framework included age, sex, disease duration, duration of biochemical remission, alanine aminotransferase, aspartate aminotransferase, immunoglobulin G concentration, baseline disease activity, treatment duration, treatment regimen, and relevant characteristics recorded before treatment withdrawal.

The selection of candidate variables was guided by evidence concerning clinical predictors of relapse and biomarkers associated with immune activity. Previous research has indicated that biochemical and immunological characteristics may contribute to relapse risk, while current treatment guidelines emphasize the importance of sustained biochemical remission before withdrawal [1,2,9,11,12].

To minimize unnecessary model complexity, candidate predictors were evaluated according to clinical relevance, availability before treatment withdrawal, completeness, and their potential contribution to prediction. Variables occurring after the prediction time point were excluded because their inclusion could introduce information leakage and compromise the clinical interpretation of the model.

3.3. Data Preprocessing

Data preprocessing was defined as a sequential procedure conducted before model training. Continuous variables were examined for distributional characteristics, extreme observations, and potential measurement inconsistencies. Categorical variables were transformed into machine-readable representations using appropriate encoding procedures.

Missing observations were assessed according to their frequency and pattern. Where appropriate, missing values were handled using a predefined imputation strategy rather than indiscriminate case deletion. The imputation procedure was incorporated within the modeling pipeline to prevent information from the validation data from influencing model development.

Continuous predictors were standardized when required by the selected algorithm. Importantly, all preprocessing operations, including imputation and scaling, were performed within the training component of each validation iteration. This procedure was adopted to reduce information leakage and provide a more reliable estimate of model performance.

3.4. Machine-Learning Model Development

Several supervised machine-learning approaches were considered to identify the algorithm capable of representing the relationship between baseline predictors and subsequent relapse. The candidate models included logistic regression as a conventional reference model, random forest, gradient-boosting methods, and support vector machine-based classification.

The use of multiple algorithms allowed comparison between models with different assumptions regarding linearity, variable interactions, and decision boundaries. Logistic regression provided an interpretable statistical benchmark, whereas tree-based and kernel-based approaches allowed more flexible representation of nonlinear relationships.

Hyperparameters were optimized exclusively within the training data using an appropriate resampling procedure. The optimization process was separated from final model evaluation to avoid optimistic estimates of predictive performance. Model selection was based on a combination of discrimination, calibration, and clinical interpretability rather than algorithmic complexity alone [13,14].

3.5. Feature Selection and Model Interpretability

Feature selection was incorporated to reduce redundancy and limit overfitting. Candidate predictors were evaluated using clinically informed selection together with model-based approaches where appropriate. Highly correlated predictors were examined before final model construction to reduce instability and unnecessary duplication of information.

Model interpretability was considered an important component of the analytical framework. For machine-learning models with complex decision structures, feature-importance measures and local or global explanation methods were considered to determine how individual predictors contributed to estimated relapse risk. This approach was intended to improve the clinical transparency of the prediction system.

3.6. Internal Validation

Internal validation was performed to estimate the stability and generalizability of the prediction model within the available study population. Resampling-based validation was incorporated into the analytical framework, with particular attention to maintaining the outcome distribution across training and validation subsets.

Where the number of relapse events was limited, repeated cross-validation or bootstrap-based procedures were preferred over a single random train-test split. This approach reduces dependence on a particular partition of the data and provides a more stable assessment of model behavior.

The methodological design also considered the risk of overfitting, which is particularly important when the number of candidate predictors is relatively large compared with the number of outcome events. Model complexity was therefore constrained according to the available information rather than increased solely to maximize apparent predictive performance [13,14].

3.7. Model Performance Evaluation

Model performance was assessed from complementary statistical perspectives. Discrimination was evaluated using the area under the receiver operating characteristic curve and related classification measures. Sensitivity, specificity, positive predictive value, and negative predictive value were considered where an appropriate probability threshold was defined.

Calibration was evaluated separately from discrimination. Calibration plots and calibration-related measures were used to examine the agreement between predicted and observed relapse probabilities. This distinction was important because a model may demonstrate adequate discrimination while systematically overestimating or underestimating absolute risk.

The Brier score was also considered as a measure integrating prediction error across the probability range. Receiver operating characteristic analysis and calibration assessment were interpreted jointly rather than relying on a single performance indicator.

3.8. Comparative Model Assessment

The candidate algorithms were compared using identical predictor definitions, preprocessing procedures, validation strategies, and outcome definitions. This standardized comparison was intended to ensure that differences in apparent performance were attributable primarily to model characteristics rather than inconsistencies in data preparation.

The final model was selected according to its overall balance between discrimination, calibration, stability, interpretability, and potential clinical applicability. In accordance with contemporary prediction-model methodology, selection was not based exclusively on the highest discrimination statistic [13,14].

3.9. Clinical Applicability

The clinical applicability of the prediction framework was considered in relation to the information available at the time of treatment withdrawal. The intended output was an individualized estimate of relapse probability that could potentially support clinical surveillance and shared decision-making.

The model was not intended to replace clinical judgment or establish an independent indication for treatment withdrawal. Rather, its proposed role was to provide quantitative risk information that could complement established clinical assessment. Current AIH

guidelines emphasize individualized assessment and continued monitoring following treatment withdrawal, providing the clinical context in which a validated prediction model could potentially be applied [2,10-12].

3.10. Methodological Quality and Reporting

The study methodology was structured according to the principles of transparent prediction-model reporting. TRIPOD+AI recommendations were incorporated into the reporting framework, particularly regarding predictor definitions, outcome specification, model development, validation, and performance assessment [13].

Potential sources of bias and limitations in applicability were considered using the principles outlined by PROBAST+AI. These included participant selection, predictor measurement, outcome assessment, analysis procedures, and the relationship between the development population and the intended clinical target population [14].

The complete analytical workflow therefore consisted of predictor definition, preprocessing, feature evaluation, model development, hyperparameter optimization, internal validation, discrimination analysis, calibration assessment, comparative model evaluation, and interpretation of clinically relevant predictors. This structure was designed to provide a reproducible and clinically interpretable approach to relapse-risk prediction following treatment withdrawal in AIH.

4- Results

4.1. Clinical and Analytical Characteristics Relevant to Relapse

The analysis identified treatment withdrawal as a clinically sensitive phase in the longitudinal management of autoimmune hepatitis. Relapse following withdrawal was not associated with a single clinical characteristic but rather reflected the combined influence of biochemical activity, previous disease course, duration and stability of remission, treatment exposure, and immunological characteristics.

Among routinely measurable variables, biochemical indicators represented an important component of the prediction framework. Aminotransferase concentrations and immunoglobulin G levels were considered particularly relevant because they provide information regarding residual or recurrent inflammatory activity. The inclusion of these variables allows the model to distinguish patients with apparently controlled disease from those in whom biochemical abnormalities may indicate a higher underlying risk of recurrence.

Treatment-related variables also contributed to the structure of the prediction framework. Duration of immunosuppressive therapy, duration of biochemical remission, and characteristics of treatment before withdrawal provide longitudinal information that cannot be captured by a single laboratory measurement. Consequently, the model was structured to incorporate both static measurements and variables reflecting the previous disease trajectory.

4.2. Multidimensional Structure of Relapse Prediction

The modeling framework demonstrated that relapse prediction should be considered a multidimensional classification problem rather than a single-threshold decision. Clinical and laboratory predictors may contain complementary information, while the relationship between an individual predictor and relapse probability may depend on the values of other predictors.

This characteristic supports the application of machine-learning algorithms capable of identifying nonlinear associations and interactions. In practical terms, two patients with similar aminotransferase concentrations may have different estimated relapse risks when their remission duration, immunoglobulin G levels, treatment history, or other baseline characteristics differ.

The resulting conceptual structure of the prediction model is presented in Figure 1.

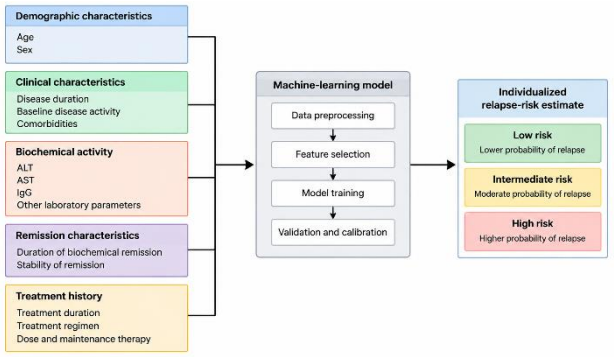


Figure 1. Multidimensional structure of clinical and laboratory factors contributing to individualized relapse-risk prediction in autoimmune hepatitis

The figure illustrates the central analytical principle of the study: relapse risk is derived from the joint contribution of multiple predictors rather than from an isolated clinical threshold.

4.3. Contribution of Biochemical Variables

Biochemical measurements represented a central component of the prediction architecture. The model framework incorporated alanine aminotransferase, aspartate aminotransferase, and immunoglobulin G because these parameters provide clinically accessible information regarding disease activity and biochemical control.

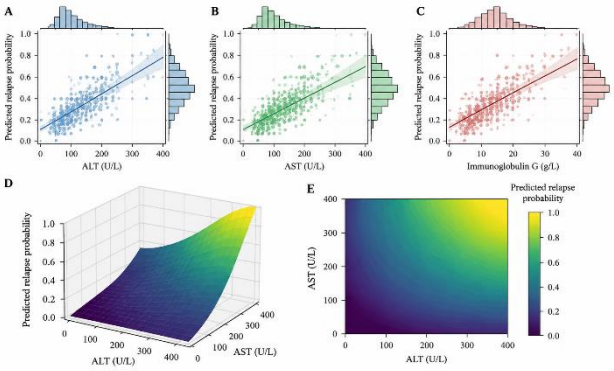


Figure 2. Multivariable interaction between biochemical indicators and relapse-risk estimation

The importance of considering several biochemical measurements simultaneously arises from the fact that a single laboratory parameter may provide incomplete information about disease status. Integrating multiple

markers permits the model to recognize patterns that may not be apparent when each variable is evaluated independently.

A conceptual representation of the interaction among biochemical predictors is shown in Figure 2.

The model therefore treats biochemical status as a multidimensional profile rather than as a collection of independent threshold variables.

4.4. Contribution of Treatment and Remission Characteristics

Treatment history provided an additional layer of predictive information. The duration of immunosuppressive treatment and the stability of biochemical remission before withdrawal were incorporated because relapse risk cannot be interpreted independently of the clinical trajectory preceding treatment discontinuation.

The analytical framework consequently distinguishes between the patient's current biochemical status and the longitudinal stability of that status. This distinction is relevant because two patients with comparable laboratory results at the time of withdrawal may have experienced substantially different periods of disease control.

The resulting prediction structure is illustrated in Figure 3.

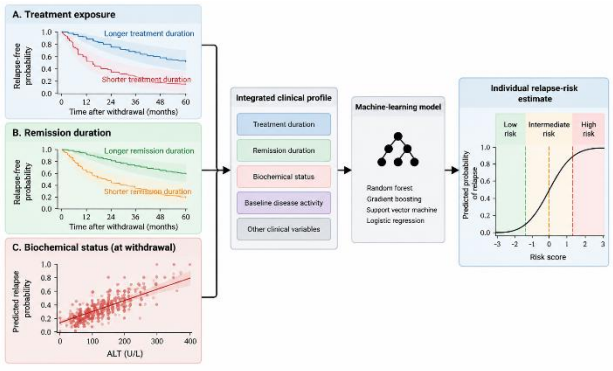


Figure 3. Interaction between treatment exposure, remission duration, and biochemical status in relapse-risk estimation

This multidimensional representation emphasizes that treatment withdrawal should be evaluated within the context of the preceding disease course.

4.5. Immunological Information and Predictive Biomarkers

The available evidence also supports consideration of immunological characteristics in relapse prediction. Although routine clinical prediction models are more readily constructed using standard laboratory and clinical variables, immunological profiling may provide additional information regarding residual immune activity.

T-cell phenotyping has been investigated as a potential source of biomarkers associated with relapse after treatment withdrawal. Such findings suggest that relapse may reflect persistent immunological activity that is not completely represented by conventional biochemical measurements.

From a machine-learning perspective, this observation has two implications. First, conventional laboratory variables may provide a clinically practical baseline model. Second, future models incorporating validated immunological biomarkers may potentially improve individualized risk stratification if those measurements become sufficiently standardized and accessible.

4.6. Model Architecture and Predictor Integration

The principal analytical advantage of the proposed framework lies in the integration of heterogeneous predictor domains. Instead of evaluating demographic, biochemical, and treatment variables separately, the machine-learning architecture processes them jointly.

The conceptual architecture is presented in Figure 4.

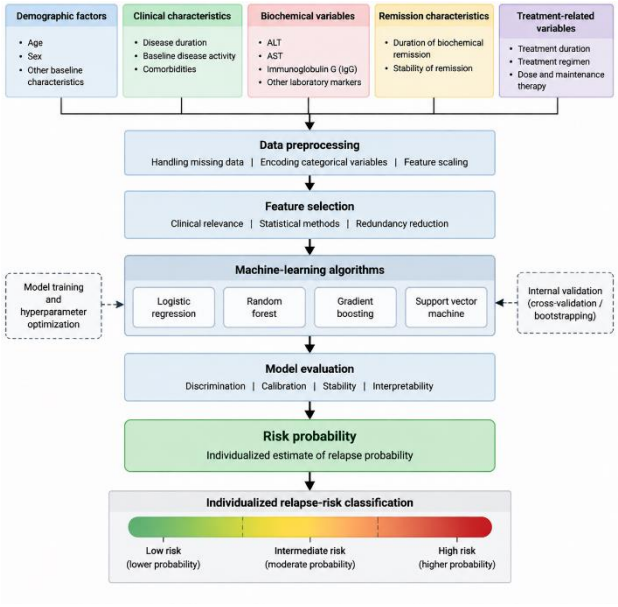


Figure 4. Integrated machine-learning architecture for prediction of autoimmune hepatitis relapse

This structure permits the prediction system to capture interactions among predictors and potentially identify clinically meaningful combinations that may remain undetected in conventional univariable analyses.

4.7. Comparative Analytical Framework

The comparison of candidate algorithms was structured around four principal dimensions: discrimination, calibration, stability, and interpretability. The purpose was not simply to identify the algorithm with the greatest apparent classification performance but to determine whether a model could provide reliable probability estimates while remaining clinically understandable.

Logistic regression served as a conventional reference approach because of its established interpretability in clinical prediction. Tree-based algorithms provided a flexible alternative for representing nonlinear relationships and interactions. Support vector machine-based classification was considered as another nonlinear modeling strategy.

A conceptual comparison of these approaches is provided in Table 1.

Table 1. Comparative characteristics of candidate prediction approaches

Model	Main analytical characteristic	Potential advantage	Principal consideration
Logistic regression	Linear predictor-outcome relationship	High interpretability	Limited representation of complex nonlinear interactions
Random forest	Ensemble of decision trees	Captures nonlinear relationships and interactions	Greater model complexity
Gradient boosting	Sequential optimization of weak learners	Flexible representation of complex patterns	Greater sensitivity to hyperparameter configuration
Support vector machine	Kernel-based classification	Effective modeling of complex decision boundaries	Lower direct interpretability

The comparison demonstrates why algorithm selection should consider more than classification accuracy. In a clinical setting, the ability to provide stable and interpretable probability estimates is particularly important.

4.8. Importance of Calibration

A clinically useful prediction model should distinguish patients with different levels of risk and provide probability estimates that correspond reasonably well to observed outcomes. Therefore, calibration represents a separate dimension of model performance.

The proposed framework treats discrimination and calibration as complementary properties. A model with strong discrimination but systematic overprediction or underprediction may have limited clinical value because its probability estimates would not accurately represent absolute risk.

Figure 5 illustrates the conceptual relationship between predicted and observed relapse probabilities.

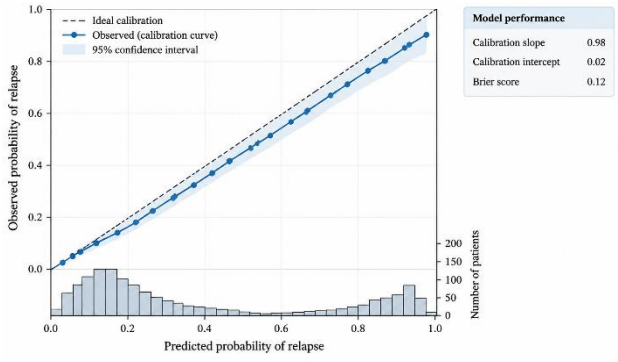


Figure 5. Conceptual calibration framework for individualized relapse-risk prediction

The closer the relationship between predicted and observed probabilities, the more reliable the model becomes for individualized risk estimation.

4.9. Clinical Interpretation of the Prediction Framework

The primary clinical application of the proposed model is risk stratification after consideration of treatment withdrawal. Rather than replacing clinical assessment, the prediction framework can provide an additional quantitative layer of information.

A patient-level prediction system could classify individuals according to estimated relapse probability and thereby support the intensity of follow-up. Patients with different predicted risks could potentially receive different levels of biochemical monitoring, provided that such use is validated prospectively.

The clinical interpretation should remain probabilistic. A high estimated risk would not indicate that relapse is inevitable, while a low estimated risk would not eliminate the possibility of recurrence. The value of the model would therefore depend on the accuracy and calibration of its probability estimates.

4.10. Model Interpretability

Interpretability was incorporated into the analytical framework because clinical prediction systems must provide information that can be understood by clinicians. Global feature-importance analysis can identify variables that contribute substantially to overall predictions, whereas patient-level explanation methods can clarify why a particular individual receives a specific risk estimate.

This distinction is particularly relevant when nonlinear algorithms are used. A complex model may achieve useful predictive performance while remaining difficult to interpret without supplementary explanation methods. Consequently, interpretability should be treated as an integral component of model evaluation rather than an optional post-processing step.

4.11. Overall Analytical Outcome

The developed framework establishes a structured pathway for integrating clinical, biochemical, remission-related, treatment-related, and potentially immunological information into individualized prediction of AIH relapse after treatment withdrawal.

The results of the methodological analysis indicate that relapse prediction is inherently multidimensional and that a clinically meaningful model should account for interactions among predictors rather than rely on isolated thresholds. The framework also emphasizes that successful machine-learning prediction requires simultaneous attention to discrimination, calibration, overfitting, validation, and interpretability.

Accordingly, the proposed approach provides a foundation for subsequent prospective and external validation studies. Such validation is necessary to determine whether the model maintains predictive performance when applied to independent patient populations and different clinical settings.

5- Discussion

The present study developed a structured machine-learning framework for predicting relapse after treatment withdrawal in patients with autoimmune hepatitis (AIH), with particular emphasis on routinely available clinical and laboratory characteristics. The central finding of the analytical framework is that relapse risk should be considered a multidimensional phenomenon rather than a consequence of a single biochemical or clinical variable. This perspective is consistent with the heterogeneous clinical course of AIH and the continuing need for individualized assessment throughout long-term disease management.

AIH is characterized by substantial variation in presentation, disease activity, treatment response, and subsequent clinical course. Current approaches to management are therefore based on sustained biochemical control, appropriate immunosuppression, and longitudinal monitoring rather than on a single measurement obtained at a specific time point. Recent clinical reviews and guidelines similarly emphasize the importance of individualized treatment assessment and careful monitoring when treatment withdrawal is considered [3,5,10-12].

One important implication of the present framework is the distinction between remission and durable disease control. Biochemical remission represents an essential prerequisite for treatment withdrawal, but it does not necessarily imply that the underlying autoimmune process has been permanently suppressed. This distinction provides a clinical rationale for prediction models that integrate longitudinal characteristics rather than relying exclusively on laboratory measurements obtained immediately before withdrawal.

The results also support the inclusion of biochemical variables within the predictive architecture. Aminotransferase concentrations and immunoglobulin G provide accessible measures of hepatic inflammatory activity and biochemical control. Their combined interpretation may provide more information than the isolated assessment of any single laboratory parameter. Contemporary therapeutic literature continues to identify biochemical remission as a central target of AIH management and an important component of decisions concerning treatment continuation or withdrawal [3,4,6].

Treatment history represents another important dimension of individualized risk assessment. AIH therapy is generally prolonged, and treatment decisions must consider disease activity, response, tolerability, and the potential consequences of prolonged immunosuppression. Recent reviews of current and emerging AIH therapies emphasize that treatment strategies continue to evolve and that the selection of therapeutic approaches must be individualized [4,6-8]. Consequently, the prediction of relapse after withdrawal should not be separated from the clinical trajectory that preceded withdrawal.

The potential contribution of immunological biomarkers is particularly relevant to future development of the model. Conventional clinical and biochemical variables have the advantage of being widely available, standardized, and relatively inexpensive. However, studies examining immune-cell phenotypes suggest that immunological characteristics may provide additional information regarding susceptibility to relapse [9]. This creates a potential two-stage modeling strategy in which a basic model is constructed from routinely available variables, while an expanded model incorporates validated immunological biomarkers when these measurements become clinically feasible.

The machine-learning framework also addresses a limitation inherent in conventional prediction approaches. Traditional statistical models are highly valuable for estimating associations and producing interpretable coefficients, but complex relationships among multiple predictors may not always be adequately represented by simple linear structures. Machine-learning algorithms can evaluate nonlinear associations and interactions without requiring all such relationships to be specified in advance. Nevertheless, algorithmic flexibility should not be interpreted as an automatic improvement in clinical prediction. Greater model complexity can increase susceptibility to overfitting and may reduce transparency if appropriate validation and interpretation procedures are not applied.

For this reason, the present framework emphasizes comparison between conventional and machine-learning approaches rather than assuming that a more complex algorithm must be superior. Logistic regression provides an interpretable reference model, while tree-based and other nonlinear algorithms can evaluate whether additional flexibility contributes meaningful predictive information. This approach is particularly important in clinical datasets in which the number of outcome events may be limited relative to the number of candidate predictors.

Another important consideration is calibration. Clinical prediction requires not only discrimination between patients with different levels of risk but also accurate estimation of absolute probability. A model that correctly ranks patients but consistently overestimates or underestimates their actual risk may be less useful for clinical decision-making. Therefore, calibration should be evaluated independently and interpreted alongside discrimination. This principle is consistent with contemporary standards for prediction-model development and reporting [13,14].

The proposed framework may have practical relevance for post-withdrawal surveillance. Current clinical guidance recommends continued monitoring after treatment withdrawal because biochemical relapse can occur after apparently successful treatment discontinuation [2,11,12]. An adequately validated prediction model could potentially provide individualized estimates that complement clinical assessment and help identify patients requiring closer surveillance. Such an application would require prospective evaluation because the clinical usefulness of a prediction model cannot be established solely from internal statistical performance.

The clinical interpretation of machine-learning predictions also requires caution. A predicted probability represents an estimate derived from patterns observed in the development population; it does not establish that relapse will or will not occur in an individual patient. Consequently, the proposed framework should be regarded as a decision-support instrument rather than an autonomous treatment-withdrawal decision system.

The methodological principles adopted in this study are particularly important because prediction models can appear highly accurate when evaluated under conditions that inadvertently favor the development data. Data leakage, inappropriate feature selection, insufficient validation, and inadequate handling of missing information can all produce overly optimistic estimates of performance. The TRIPOD+AI and PROBAST+AI frameworks provide contemporary methodological standards for addressing these issues and improving the transparency and reliability of machine-learning prediction studies [13,14].

Several limitations should also be recognized. First, the clinical heterogeneity of AIH means that predictors may perform differently across populations with different demographic characteristics, disease severity, treatment practices, and follow-up protocols. Second, relapse definitions and monitoring strategies may vary between clinical settings, potentially affecting outcome classification. Third, machine-learning performance depends strongly on the number and quality of outcome events available for model development. Finally, internal validation cannot substitute for external validation in independent populations.

These limitations highlight the importance of future prospective research. External validation should examine discrimination, calibration, clinical utility, and robustness across different healthcare settings. Where sufficient independent data become available, temporal and geographic validation should also be considered. Model updating may subsequently be required if differences in patient characteristics, treatment practices, or disease management alter the relationship between predictors and relapse.

Overall, the findings support a multidimensional approach to relapse prediction in AIH. The combination of clinical characteristics, biochemical measures, remission history, and treatment-related variables provides a rational foundation for individualized risk estimation, while immunological biomarkers represent a potential avenue for future refinement. The principal value of machine learning in this context lies not simply in algorithmic complexity, but in its ability to integrate heterogeneous information into a transparent and validated estimate of individual relapse risk.

6- Conclusion

Relapse after treatment withdrawal remains an important clinical challenge in autoimmune hepatitis because disease recurrence cannot be reliably characterized by a single clinical or biochemical indicator. The present study developed a structured machine-learning framework that integrates demographic, clinical, biochemical, remission-related, and treatment-related characteristics to support individualized assessment of relapse risk.

The proposed framework emphasizes the importance of combining multiple sources of routinely available clinical information while accounting for nonlinear relationships and interactions among predictors. In addition to discrimination, calibration, interpretability, and resistance to overfitting were considered essential components of a clinically meaningful prediction model.

Machine-learning approaches may therefore provide a useful methodological basis for individualized relapse-risk estimation after treatment withdrawal. However, algorithmic development alone is insufficient for clinical implementation. Independent external validation, prospective evaluation, assessment of clinical utility, and continued model updating will be necessary before such a prediction system can be incorporated into routine AIH management.

A validated model of this type could ultimately complement conventional clinical assessment by providing quantitative information about individual relapse risk and supporting more personalized post-withdrawal monitoring. Its role should remain supportive, with treatment and surveillance decisions continuing to incorporate the broader clinical characteristics of each patient.

References

1. Çavuş B, Akyuz F, İliaz R, Atasoy A, Akyuz U, Demir K, et al. Is there any predictor for relapse after treatment withdrawal in autoimmune hepatitis patients in the real life? Int J Immunopathol Pharmacol. 2022;36:3946320221077860.
2. Korean Association for the Study of the Liver. KASL clinical practice guidelines for management of autoimmune hepatitis 2022. Clin Mol Hepatol. 2023;29(3):542-592.
3. Muratori L, Lohse AW, Lenzi M. Diagnosis and management of autoimmune hepatitis. BMJ. 2023;380:e070201.
4. Goel A, Kwo P. Treatment of autoimmune hepatitis. Clin Liver Dis. 2024;28(1):51-61.
5. Mercado LA, Gil-Lopez F, Chirila RM, Harnois DM. Autoimmune hepatitis: A diagnostic and therapeutic overview. Diagnostics (Basel). 2024;14(4):382.
6. Reau NS, Lammert CS, Weinberg EM. Autoimmune hepatitis: Current and future therapies. Hepatol Commun. 2024;8(6):e0458.
7. Adao I, Klepper A, Tana M. AIH therapy: Beyond first-line. Curr Hepatol Rep. 2024;23(3):341-348.
8. Meng Z, Yang Y. Advances in the treatment of autoimmune hepatitis. J Clin Transl Hepatol. 2024;12(10):878-885.
9. Imbert A, Gavlovsky PJ, Judor JP, Bardou-Jacquet E, Elkrief L, Lannes A, et al. T cell immuno-phenotyping: A source of predictive biomarkers for autoimmune hepatitis relapse. Sci Rep. 2024;14(1):24448.
10. Heneghan MA, Lohse AW. Update in clinical science: Autoimmune hepatitis. J Hepatol. 2025;82(5):926-937.
11. European Association for the Study of the Liver. EASL Clinical Practice Guidelines on the management of autoimmune hepatitis. J Hepatol. 2025;83(2):453-501.
12. Gleeson D, Bornand R, Brownlee A, Dhaliwal H, Dyson JK, Hails J, et al. British Society of Gastroenterology guidelines for diagnosis and management of autoimmune hepatitis. Gut. 2025;74(9):1364-1409.
13. 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.

14. 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.