PREPARE ALL: An Artificial Intelligence Tool for Predicting Relapse in Children With Acute Lymphoblastic Leukemia

Subikksha Saravanan, ME[1]; Raghunathan Rengaswamy, PhD[2]; Gaurav Narula, MD[3]; Sameer Bakhshi, MD[4]; Rachna Seth, MD[5]; Nandana Das, PhD[6]; Manash Pratim Gogoi, BHM[6]; Shripad Banavali, MD[3]; Prasanth Srinivasan, DM[7]; Gargi Das, DM[7]; TK Balaji, MD[7]; Shekar Krishnan, PhD, MRCP, FRCPath[6]; Vaskar Saha, MD[6]; Vijayalakshmi Ramshankar, PhD[1]; and Venkatraman Radhakrishnan, MD, MBBS, MSc, DM[7]

DOI https://doi.org/10.1200/CCI-25-00222

ABSTRACT

PURPOSE The Pediatric Relapse Prediction and Risk Evaluation for Acute Lymphoblastic Leukemia (PREPARE-ALL) tool aims to predict relapse in pediatric ALL by integrating clinical expertise with artificial intelligence and machine learning (ML), particularly Extreme Gradient Boosting (XGBoost). PREPARE-ALL demonstrates that multicenter, protocol-driven clinical and laboratory data can be used through ML to generate reproducible relapse predictions with greater sensitivity than individual clinician assessments.

METHODS PREPARE-ALL was developed using data from the ICiCLe ALL-14 pretrial cohort across five centers, incorporating 33 clinical and laboratory features.

RESULTS Among 2,252 patients enrolled in the study, 565 (25.1%) relapsed. Using an 80:20 train-test split, XGBoost achieved a sensitivity of 68.5% (245/447 relapses detected). Additional metrics included a positive predictive value of 31.3%, a negative predictive value of 82.8%, an accuracy of 54.8%, and a specificity of 50.3%. Key predictors of relapse included high hyperdiploidy and BCR-ABL1 fusion positive, positive measurable residual disease status at the end of induction, sex, age, highest presenting WBC, and final risk group. Three clinicians scored the validation data set; the developed model achieved a higher recall (68.5%) compared with clinical judgment (approximately 31%-36%).

CONCLUSION PREPARE-ALL identifies twice as many relapses as clinicians and serves as a practical decision-support tool for early relapse triage and treatment planning, enabling timely therapeutic adjustments and improved outcomes in pediatric ALL.

ACCOMPANYING CONTENT

Data Supplement

Accepted December 8, 2025
Published January 21, 2026

JCO Clin Cancer Inform
10:e2500222

INTRODUCTION

ALL is the most common childhood cancer globally as well as in India.[1] Although early detection and prompt initiation of treatment can significantly increase the chances of cure and survival to 90%, ALL remains one of the leading causes of childhood mortality. Studies show that approximately 20%-25% of children with ALL experience relapse, emphasizing the need for better predictive tools and management strategies.[2] Cure rates after relapse are significantly lower, making it critically important to stratify patients for appropriate chemotherapeutic treatments and enable early recognition of outcomes to mitigate poor survival.

Machine learning (ML) has emerged as a powerful tool in health care, particularly in oncology, for identifying patterns that are not discernible through traditional statistical methods.[3,4] ML classifiers can analyze complex data sets, decipher patterns, and derive predictions on the basis of patient characteristics and disease presentation. Algorithms such as gradient boosting, support vector machines (SVM), random forest, k-nearest neighbor (KNN) on the basis of the vote of nearest neighbors, decision trees, and neural networks have previously been used for treatment response prediction.[5-7]

Although measurable residual disease (MRD) remains the strongest prognostic marker for relapse, it fails to capture a subset of biologically aggressive cases, underscoring the need for additional tools to refine risk stratification.

ML models have demonstrated significant potential in predicting outcomes and improving risk stratification in various cancers, including breast, lung, and pancreatic cancers, achieving notable accuracy improvements.[8-10] With

CONTEXT

Key Objective

Can machine learning–based clinical decision tools improve early relapse prediction in children with ALL beyond conventional risk stratification and individual clinician judgment?

Knowledge Generated

Using data from 2,252 patients with pediatric ALL treated under a uniform ICiCLe protocol, the PREPARE-ALL XGBoost model identified 68.5% of relapse cases, nearly twice the recall of clinicians (approximately 31%-36%) using the same baseline variables.

Relevance *(D.S. Bitterman)*

A scalable, interpretable relapse prediction model for pediatric ALL may improve prognostication and, thereby, enhance clinical decision making.*

*Relevance section written by *JCO Clinical Cancer Informatics* Associate Editor Danielle S. Bitterman, MD.

available gene expression data and digitized phenotypic data, ML has been successfully applied to predict outcomes in different types of cancers, including leukemia.[11]

In this study, we evaluated ML classification models to predict relapse in pediatric ALL using data from the Indian Collaborative Childhood Leukemia group (ICiCLe) ALL-2014 pretrial cohort.

METHODS

Ethical Committee Approval

The study was approved by the Ethics Committee of Cancer Institute (WIA), Chennai (CIWIA-IEC/2022/July 03). Written informed consent was obtained from each patient before study entry, in accordance with the Declaration of Helsinki. The ICiCLe ALL-2014 pretrial cohort was registered with the Clinical Trials Registry of India (registration number: CTRI/2015/12/006434).

Population and Outcome

Clinical data were retrieved for a cohort of 2,331 patients newly diagnosed with pediatric ALL enrolled in the ICiCLe ALL-2014 pretrial cohort between 2014 and 2017 at the Cancer Institute (WIA) in Chennai, Tata Medical Center in Kolkata, Tata Memorial Hospital in Mumbai, and All India Institute of Medical Sciences in New Delhi. The ICiCLe ALL-2014 pretrial cohort patients were treated according to a nonrandomized, risk-stratified protocol.[12] Patients with B-precursor ALL were risk-stratified as standard-risk (SR), intermediate-risk (IR), and high-risk (HR) on the basis of age, baseline WBC count, response to prednisolone prophase, cytogenetics, and MRD status at the end of induction (EOI). Patients with T-cell ALL were treated as HR.[13]

Comprehensive baseline data were used to develop and validate the ML model for relapse prediction. Patients were retrospectively analyzed and categorized on the basis of their treatment response and relapse status, focusing on identifying key predictive features associated with relapse.

ML Strategy

The strategy was based on building a reliable and interpretable classification framework by combining clinical insights with supervised ML learning techniques. The overall methodology was organized into the following core stages: preparing the data to reduce noise and bias, evaluating a diverse set of algorithms suited for structured clinical data, identifying the most informative features associated with relapse, and validating the model performance through repeated sampling. This structured framework ensured the model maintained a balance between statistical robustness and clinical relevance. The implementation pipeline, highlighting the ML development steps and the complete workflow from data input to mobile application deployment, is illustrated (Data Supplement, Fig S1).

Data Preprocessing

Relapse in the study was defined as any recurrence of disease after achieving morphological remission at the EOI and occurring during or after consolidation therapy. Induction failures and deaths during induction were not considered for relapse analysis, as these patients were never at risk of postremission relapse. Similarly, remission status at the EOI was excluded from the feature set to prevent label leakage, although it was retained in outcome definitions. Patients who did not achieve remission at EOI (n = 79) were classified as refractory and excluded, since induction failure reflects treatment refractoriness instead of relapse biology and

retaining them would distort relapse-specific modeling, resulting in a final evaluable cohort of 2,252 patients. They were not modeled as separate events because refractory disease represents primary treatment resistance rather than postremission recurrence, and combining these distinct clinical trajectories would confound prediction targets.

The data set demonstrated minimal missingness (1.16% of all values across given variables). No variable exceeded the predefined 20% threshold for missing data, and no feature showed zero variance; hence, all were retained. Missing values were handled using median imputation for continuous features and mode imputation for categorical features. The final data set was randomly split into training (80%) and test (20%) subsets, and a stratified 5-fold cross-validation (Data Supplement, Appendix A) was applied to ensure stable model performance under class imbalance.

Feature Selection

To minimize bias and prevent information leakage, variables directly linked to study outcomes or patient identifiers (eg, unique IDs, event dates, and follow-up outcomes) were excluded (Data Supplement, Table S1). All selected features were static values recorded at presentation (age, sex, lineage, bulky disease, cytogenetic groups, and presenting WBC), at day 8 (CNS status and prednisolone response), or at EOI (MRD status and final risk group). Longitudinal follow-up or post-therapy variables were excluded to prevent information leakage. Selection was guided by previous pediatric leukemia evidence and clinical expert consultation, with emphasis on interpretability, biological plausibility, and reproducibility. Pairwise Pearson correlation analysis (Data Supplement, Fig S2) confirmed that none of the included features were highly redundant.

Model Development

Ten ML classifiers were benchmarked for relapse prediction, which included linear methods (logistic regression,[14] SVM,[15] and Gaussian naive Bayes[16]), tree-based approaches (decision trees[17] and random forest[18]), boosting algorithms (XGBoost,[19] CatBoost,[20] and AdaBoost[21]), and a distance-based method (KNN[22]). All models were implemented in Python using scikit-learn, with dedicated libraries. Preprocessing pipelines ensured reproducibility: numeric features were median-imputed and standardized where necessary; categorical features were label-encoded, while detailed cytogenetics was one-hot encoded because it contained multiple unique subcategories without a defined hierarchy of severity. For algorithms sensitive to feature scale (logistic regression, SVM, and KNN), z-score normalization was applied, while tree-based models used raw values.

Training and validation followed a stratified 80:20 split, with stratified 5-fold cross-validation inside the training set for model selection. To address class imbalance (Data Supplement, Appendix B1), class weighting was applied consistently rather than oversampling (Data Supplement, Table S2), as this approach avoids synthetic data distortion and preserves the true event distribution in a clinical cohort.

Model Assessment and Validation

Model performance was assessed using five complementary metrics suitable for binary classification under imbalance: accuracy, precision, recall (sensitivity), F1 score, and the area under the receiver operating characteristic curve (AUROC; Data Supplement, Appendix B2).[23] Accuracy reflected overall correctness, precision quantified the proportion of true relapses among predicted relapses, recall measured sensitivity for detecting relapses, and F1 captured the balance between recall and precision. AUROC assessed discriminatory power across thresholds. Model selection emphasized recall and F1 as primary objectives, given the clinical need to minimize missed relapse cases, with AUROC reported for global discrimination. All metrics were first estimated on validation folds during cross-validation and subsequently confirmed on the independent 20% test set. To ensure stability, the split-train-tune procedure was repeated across multiple random seeds, and test-set confidence intervals were calculated using bootstrapping. Agreement between cross-validated and test results was used as a measure of generalizability, and calibration was verified to check whether predicted probabilities aligned with observed relapse frequencies.

Model Selection

A heterogeneous set of supervised ML classifiers was intentionally selected to capture different decision boundaries and learning mechanisms relevant to structured clinical data. Linear models (logistic regression and naive Bayes) were included for their interpretability and strong performance on linearly separable patterns. Tree-based classifiers (decision trees and random forest) were incorporated to model nonlinear interactions and heterogeneous feature effects. Distance-based learning (KNN) was tested as a nonparametric comparator. Boosting algorithms (AdaBoost, XGBoost, and CatBoost) were selected because of their documented ability to handle tabular clinical data, account for class imbalance through weighting, and improve performance over weak learners. This allowed for systematic benchmarking and ensured that the final selection was based on performance rather than algorithmic bias. Priority was placed on recall and F1 score during evaluation, given the clinical imperative to minimize false negatives in relapse prediction. On the basis of these selection criteria and comparative performance (Table 1), XGBoost was finalized as the primary model for deployment.

Clinician Versus XG Boost Performance

Clinician evaluation was conducted on the validation data set (n = 447; 20% of the cohort) to assess concordance with model predictions. Three clinicians (G.D., P.S., and T.K.B.)

TABLE 1. Model Performance of the Different Machine Learning Models

Model	Accuracy TP + TN/(TP + TN + FP + FN) %	ROC AUC SCORE %	Recall TP/(TP + FN) %	F1 Score 2 × (PREC × REC)/(PREC + REC) %	Precision TP/(TP + FP) %
XGBoost	54.81	63.21	68.47	42.94	31.28
Random forest	63.31	61.47	41.41	35.94	31.72
Gradient boosting	59.51	61.65	45.05	35.59	29.41
Logistic regression	50.12	59.53	68.47	40.53	28.79
CatBoost	62.86	62.71	41.41	35.66	31.29
AdaBoost	62.42	64.75	57.66	43.24	34.60
Decision tree	57.78	59.61	47.75	35.93	28.80
SVM	58.39	50.95	55.86	40.46	31.16
KNN	51.91	56.70	56.76	36.96	27.39
Naive Bayes	47.42	55.68	67.57	38.96	27.37

Abbreviation: FN, false negative; FP, false positive; KNN, K-nearest neighbor; PREC, precision; REC, recall; ROC AUC, area under the receiver operating characteristic curve; SVM, support vector machine; TN, true negative; TP, true positive.

independently evaluated the masked test data set, viewing only baseline variables available to the model at prediction time, without access to outcomes, model outputs, or each other's assessments. They provided binary (yes/no) relapse predictions on the basis of clinical judgment. These assessments were compared against model outputs to evaluate agreement and identify patterns of concordance and divergence between clinician judgment and model predictions.

Sensitivity Analysis: Feature Reduction and Selection

A reduced set of seven clinically established predictors (MRD at EOI, highest presenting WBC, sex, age, prednisolone response, final risk group, and cytogenetic groups) was evaluated against the full 14-feature model to assess robustness and interpretability. Although the reduced model was designed to test whether a smaller core set could preserve performance, the 14-feature configuration was retained for the final Pediatric Relapse Prediction and Risk Evaluation for Acute Lymphoblastic Leukemia (PREPARE-ALL) implementation. Inclusion of the additional variables improved stability across risk subgroups, reduced variability in cross-validation, and ensured that all clinically relevant information, particularly detailed cytogenetic categories, was incorporated. Retaining the full feature set also supports future external validation, where additional predictors may carry greater weight, and enhances clinician confidence by aligning with the breadth of routinely collected data. This approach balances parsimony with comprehensiveness, ensuring both robustness and practical applicability of the final tool.

PREPARE-ALL Web Application Development

The PREPARE-ALL web application was developed as a pilot decision support tool hosted on a secure institutional server using Python Flask framework with an HTML-CSS-JavaScript front end. The application accepts individual or batch patient data inputs in CSV format, processes them through the XGBoost model backend, and provides relapse risk outputs with a five-point confidence meter. Data privacy is ensured via restricted access login, and no patient identifiers are stored in the hosted application to maintain confidentiality. Future deployments of the application will integrate real-time database connectivity with encrypted transfer protocols for seamless clinical use.

RESULTS

Patient Characteristics

Of 2,252 patients, 565 experienced relapses (25.08%) as shown in Table 2, with males showing a higher relapse rate (27.8%, 403/1,450) than females (20.2%, 162/802; $P < .001$). The median age at diagnosis was 5.0 years (range, 1.0-18.0 years), with significant variation in highest presenting WBC counts. MRD status was strongly associated with relapse ($P < .001$). The median follow-up was 47.9 months (range, 0.07-111.2 months).

XGBoost Showed Superior Performance in the Prediction of Relapse

Among the models evaluated (Table 1), XGBoost emerged as the most balanced performer, with an accuracy of 54.81%, a recall of 68.47%, an F1 score of 42.94%, and an ROC AUC of 63.21%. Although overall accuracy was modest (54.8%), recall was prioritized because the clinical objective is to minimize missed relapse cases, where failure to identify a true relapse carries far greater consequences than a false positive. AdaBoost and CatBoost achieved comparable accuracies (approximately 62%-63%), both trailed in recall (≤58%), which is critical for relapse detection. Gradient boosting and random forest showed moderate performance (accuracy approximately 59%-63%, recall approximately 41%-45%) but lower F1 scores, limiting sensitivity. Simpler

TABLE 2. Patient Characteristics Used for the Application

Clinical Parameter	Total Patients (N = 2,252)	No Relapse, No. (%)	Relapsed, No. (%)
Age, years			
0-10	1,768	1,346 (76.1)	422 (23.9)
10-18	484	341 (70.5)	143 (29.5)
Sex			
Female	802	640 (79.8)	162 (20.2)
Male	1,450	1,047 (72.2)	403 (27.8)
WBC at presentation, cells/μL			
0-50,000	1,780	1,367 (76.8)	413 (23.2)
>50,000	469	318 (67.8)	151 (32.2)
Missing	3	2 (66.7)	1 (33.3)
NCI risk			
High	879	617 (70.2)	262 (29.8)
Standard	1,372	1,069 (77.9)	303 (22.1)
Missing	1	2 (66 · 7)	1 (33 · 3)
Lineage			
B	2,087	1,558 (74.7)	529 (25.3)
T	165	129 (78.2)	36 (21.8)
Prednisolone response			
Good	1,803	1,350 (74.9)	453 (25.1)
Poor	346	251 (72.5)	95 (27.5)
Missing	103	86 (83.5)	17 (16.5)
Bulky disease			
No	1,448	1,089 (75.2)	359 (24.8)
Yes	743	546 (73.5)	197 (26.5)
Missing	61	52 (85.2)	9 (14.8)
Cytogenetic group			
High-risk	242	153 (63.2)	89 (36.8)
Non–high-risk	1,777	1,352 (76.1)	425 (23.9)
Not required	165	129 (78.2)	36 (21.8)
Missing	68	53 (77.9)	15 (22.1)
CNS involvement			
No	2,079	1,548 (74.5)	531 (25.5)
Yes	76	60 (78.9)	16 (21.1)
Missing	97	79 (81.4)	18 (18.6)
Provisional risk			
High	705	497 (70.5)	208 (29.5)
Intermediate	729	549 (75.3)	180 (24.7)
Standard	653	512 (78.4)	141 (21.6)
T	165	129 (78.2)	36 (21.8)
Final risk			
High	967	671 (69.4)	296 (30.6)
Intermediate	558	423 (75.8)	135 (24.2)
Standard	459	361 (78.6)	98 (21.4)
T	150	114 (76.0)	36 (24.0)
Missing	118	118 (100.0)	0 (0.0)

Abbreviation: NCI, National Cancer Institute.

linear models such as logistic regression and naive Bayes achieved high recall (approximately 68%) but poor precision (<30%), resulting in modest F1 scores. KNN, decision tree, and SVM demonstrated intermediate metrics without a clear advantage.

ROC curve analysis (Fig 1) confirmed that XGBoost and AdaBoost provided stronger discriminative ability than logistic regression or naive Bayes, while remaining competitive with other methods. Taken together, XGBoost was selected for PREPARE-ALL because of its superior recall, balanced F1 score, and robust AUC, aligning with the clinical priority of minimizing missed relapse cases.

Shapley Additive Explanations Analysis With the Top Features

The Shapley additive explanations (SHAP; Data Supplement, Appendix C) beeswarm plot (Figs 2A and 2B) quantified both linear and nonlinear effects, providing interpretable insights into how clinical, cytogenetic, and MRD-based variables shaped relapse predictions. High hyperdiploidy, MRD status at EOI, and BCR-ABL1 emerged as the strongest predictors, followed by highest presenting WBC, age, sex, and final risk classification. Directionality analysis showed that high hyperdiploidy contributed negatively to relapse prediction, consistent with its known favorable prognosis, whereas BCR-ABL1 positivity, poor MRD response, and hyperleukocytosis (>50,000/μL) contributed positively, increasing relapse risk. Additional features including ETV6-RUNX1, CNS status, provisional risk, and lineage contributed moderately but were retained to preserve completeness, as no two variables were highly correlated ($r > 0.8$). Clinically, these results confirmed established associations, with poor MRD response and elevated WBC driving higher relapse risk, while protective effects of hyperdiploidy were also captured by the model.

Sensitivity Analysis and Feature Reduction

A feature reduction analysis evaluated the impact of limiting the model to seven clinically selected predictors compared with the original 14-feature set. Performance on the held-out test set remained largely unchanged, with recall stable (68.4% *v* 64.8%) and only marginal differences in other metrics (Data Supplement, Table S3), confirming that the reduced set retained the key predictive signals.

Performance Across Clinical Risk Groups and Lineages

Subgroup analyses across ICiCLe-defined risk categories (Data Supplement, Table S4). In the SR group (n = 118), the

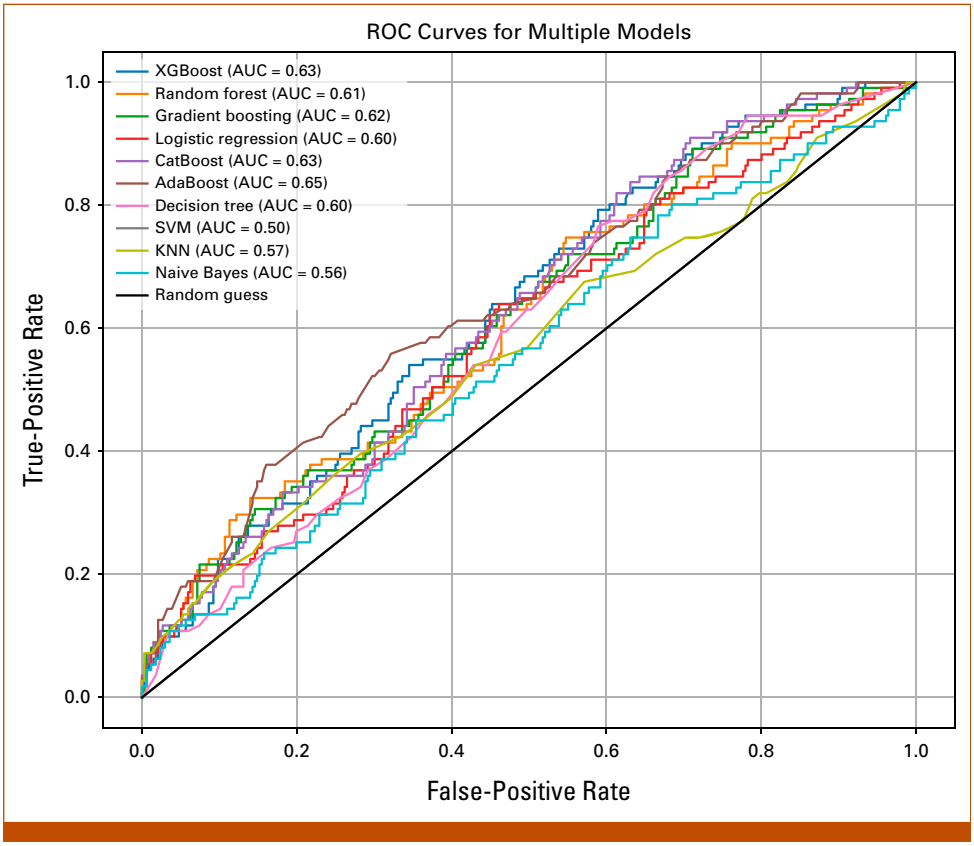

FIG 1. ROC curve of the XGB ML model for predicting relapse in the testing set. KNN, K-nearest neighbor; ML, machine learning; ROC, receiver operating characteristic curve; SVM, support vector machine.

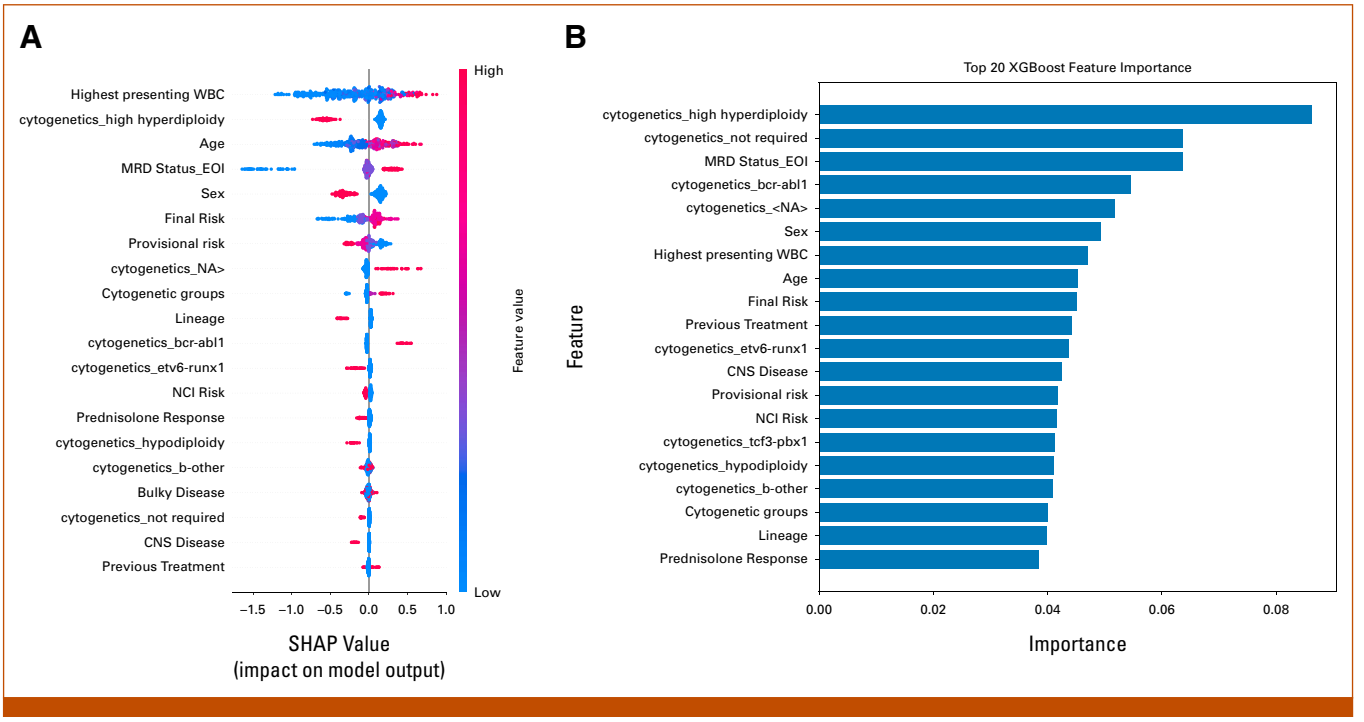

FIG 2. A pair of SHAP summary plots. (A) Beeswarm plot showing the distribution of SHAP values for each feature, indicating individual impact on model predictions. Red indicates positive impact and blue indicates negative impact. (B) Bar chart summarizing mean absolute SHAP values, highlighting features with the greatest overall influence on the model. EOI, end of induction; MRD, measurable residual disease; NA, not available; NCI, National Cancer Institute; SHAP, Shapley additive explanations.

model reached 66.9% accuracy and 52.9% recall, indicating the capacity to detect relapse even in patients typically considered low-risk. The IR group (n = 116) showed weaker performance (accuracy 44.0%, recall 44.8%), consistent with the biological heterogeneity of this intermediate category. In the HR group (n = 181), recall was high at 82.5% (F1 = 52.8%), underscoring clinical utility in minimizing missed relapses, even with modest precision (38.8%).

Overall, our model achieved the strongest sensitivity in HR patients, where relapse risk is most consequential. At the same time, SR detection highlights potential early-warning value, and IR results point to the need for molecular refinement of this gray zone.

Clinician Versus Model Concordance Evaluation

In the patient validation cohort (n = 447), PREPARE-ALL demonstrated high sensitivity for relapse detection compared with clinicians, achieving a recall of 68.5% versus approximately 31%-36% for individual clinicians (Table 3). Although the model exhibited slightly lower overall accuracy (54.8% *v* 58%-62% for clinicians), this was offset by consistently higher F1 scores and precision, reflecting more balanced predictive performance. By contrast, clinician accuracy was primarily driven by better identification of nonrelapse cases. The model also achieved greater discrimination with a higher ROC AUC (63.2% *v* 50%-53% for clinicians). Pairwise concordance analysis showed that approximately half of the predictions overlapped between the model and each clinician (approximately 49%-51%). Cohen's kappa values (Data Supplement, Appendix D) indicated only slight agreement between the model and individual clinicians (κ = 0.02-0.06). Interclinician comparisons revealed moderate agreement between clinicians 1 and 2 (κ = 0.65), whereas agreement with clinician 3 was

TABLE 3. Comparison of the Model and Clinicians (1, 2, and 3) Performance Metrics

Metric	Model	Clinician 1	Clinician 2	Clinician 3
Answered cases	447	447	447	447
Accuracy	54.81%	59.50%	58.20%	61.50%
ROC AUC	63.21%	49.80%	50.80%	52.60%
Recall	68.47%	30.90%	36.30%	35.40%
F1 score	42.94%	27.00%	29.60%	30.80%
Precision	31.28%	23.90%	25.00%	27.20%
Concordant cases	—	220	230	224
Discordant cases	—	227	217	223
Concordance rate, %	—	49.22%	51.45%	50.11%
Discordance rate, %	—	50.78%	48.55%	49.89%

Abbreviation: ROC AUC, area under the receiver operating characteristic curve.

poor to negative ($\kappa = -0.07$ and -0.02). These findings highlight substantial variability in relapse risk assessments across clinicians and demonstrate that PREPARE-ALL provides more consistent and sensitive identification of relapse-prone patients (Data Supplement, Table S5).

DISCUSSION

Pediatric ALL exhibits highly variable treatment responses, underscoring the need for individualized risk stratification.[2,4] This study, one of the first from India to analyze a large pediatric ALL cohort under a uniform protocol, demonstrates the feasibility of ML-based relapse prediction using routinely available clinical and MRD data.

Relapse still affects 15%-20% of children with ALL and remains a leading cause of mortality.[2] Conventional prognostic frameworks, primarily the National Cancer Institute risk classification and MRD status, fail to identify all HR patients. PREPARE-ALL advances beyond these by integrating clinical, laboratory, and cytogenetic features into a single decision-support tool, enabling more comprehensive risk assessment and improving detection of patients who may appear low-risk by MRD alone.

Among the 10 classifiers tested, gradient boosting algorithms consistently outperformed linear and distance-based methods, reflecting their strength in capturing complex nonlinear patterns in clinical data. XGBoost demonstrated the most balanced performance, supporting its selection for the PREPARE-ALL application (Data Supplement, Fig S3). In stratified subgroup analyses, the model achieved the highest recall in HR and T-ALL patient categories, where relapse carries the most serious clinical consequences. It also maintained sound sensitivity in SR patients, suggesting potential value as an early-warning system in lower-risk groups. The model identified approximately 7 of 10 relapses, compared with approximately 3 of 10 by clinicians, which is nearly twice that of clinicians. These findings support its use as an adjunctive triage tool that strengthens, rather than substitutes, current clinical decision-making frameworks.

Despite its strengths, the model missed approximately 30% of relapses, playing the role of a decision-support adjunct rather than a standalone determinant of therapy. Moderate sensitivity (approximately 68%) likely reflects both model constraints and labeling noise from variability in MRD testing, L-asparaginase activity, and diagnostic thresholds across centers. The lack of prospective validation and site-specific assessment limits generalizability, while the use of retrospective anonymized data prevents the evaluation of institutional differences. Interclinician agreement was moderate between two reviewers but poor with the third, underscoring the difficulty of consistently defining relapse risk. In future prospective studies, we plan to implement a standardized binary rubric for clinician scoring, collect rationale notes, and compare the model both against individual raters and a consensus panel. The retrospective design also introduces the risk of overfitting.

Future work will focus on prospective validation within ICiCLe cohorts and real-world clinical deployment with iterative clinician feedback. Integration of additional data streams, such as treatment adherence, drug-level monitoring, and genomic or transcriptomic profiling, may further refine predictions beyond accuracy. This includes calibration curves, decision curve analyses, and cost-effectiveness studies to assess clinical utility and net benefit.

In conclusion, in this multicenter Indian cohort, the PREPARE-ALL study demonstrates that XGBoost-based modeling can improve relapse prediction in pediatric ALL beyond conventional risk stratification and clinician assessment. The web-based tool translates these findings into a practical decision-support system, with SHAP interpretability confirming the clinical relevance of MRD, risk classification, and cytogenetic subgroups. By prioritizing sensitivity, the model addresses the key challenge of missed relapses and serves as a scalable adjunct for risk-adapted therapeutic planning, rather than replacing clinician judgment. Although prospective external validation is required, PREPARE-ALL establishes a foundation for integrating ML into frontline leukemia care, with future expansion to genomic and transcriptomic data expected to further refine precision-guided treatment.

AFFILIATIONS

[1]Department of Cancer Biology and Molecular Diagnostics, Cancer Institute (WIA), Chennai, India
[2]Department of Data Science and Artificial Intelligence, Indian Institute of Technology Madras, Chennai, India
[3]Department of Pediatric Oncology, Tata Memorial Centre, Homi Bhabha National Institute, Mumbai, India
[4]Department of Medical Oncology, Dr BRAIRCH, All India Institute of Medical Sciences, New Delhi, India
[5]Division of Oncology, Department of Pediatrics, All India Institute of Medical Sciences, Ansari Nagar, New Delhi, India
[6]Clinical Research Unit, Tata Translational Cancer Research Centre, Tata Medical Center, Kolkata, India
[7]Department of Medical Oncology, Cancer Institute (WIA), Cancer Institute (WIA), Chennai, India

CORRESPONDING AUTHOR

Venkatraman Radhakrishnan, MD, MBBS, MSc, DM; e-mail: venkymd@gmail.com.

EQUAL CONTRIBUTION

V. Radhakrishnan and V. Ramshankar are joint senior authors.

PRIOR PRESENTATION

Presented at the International Society of Pediatric Oncology (SIOP) 2024 Annual Meeting, Honolulu, HI, October 17-20, 2024.

SUPPORT

Supported by funding from Indian Council of Medical Research (ICMR) vide AI-ADHOC/08/2022-AI CELL; ID No. 2022-16573 granted to V.I.R. The study was funded in part by DBT-Wellcome India Alliance Fellowship (IA/M/12/1/500261), UKIERI-SPARC program (P639), and an institute grant from TCS Foundation.

DATA SHARING STATEMENT

The PREPARE-ALL application is deployed at https://adyarcancergenomics.com, and the full model codebase can be accessed from the GitHub repository at https://github.com/subikkshas/PREPARE-ALL. The participant data underlying the results are available upon reasonable request from the corresponding author.

AUTHOR CONTRIBUTIONS

Conception and design: Raghunathan Rengaswamy, Rachna Seth, Shripad Banavali, Vijayalakshmi Ramshankar, Venkatraman Radhakrishnan
Financial support: Vijayalakshmi Ramshankar, Venkatraman Radhakrishnan
Administrative support: Venkatraman Radhakrishnan
Provision of study materials or patients: Shekar Krishnan, Venkatraman Radhakrishnan
Collection and assembly of data: Gaurav Narula, Sameer Bakhshi, Rachna Seth, Nandana Das, Manash Pratim Gogoi, Shripad Banavali, Shekar Krishnan, Vaskar Saha, Vijayalakshmi Ramshankar, Venkatraman Radhakrishnan
Data analysis and interpretation: Subikksha Saravanan, Gaurav Narula, Sameer Bakhshi, Nandana Das, Prasanth Srinivasan, Gargi Das, TK Balaji, Vijayalakshmi Ramshankar, Venkatraman Radhakrishnan
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.
Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Subikksha Saravanan
Patents, Royalties, Other Intellectual Property: Inventor on a granted patent for the PREPARE-ALL relapse prediction algorithm for pediatric ALL. No royalties received

Raghunathan Rengaswamy
Research Funding: Pfizer (Inst)

Gaurav Narula
Uncompensated Relationships: ImmunoACT

Sameer Bakhshi
Research Funding: Roche India, Hetero, Sellas Life Sciences, AstraZeneca, Intas

Vaskar Saha
Research Funding: Gennova (Inst)

Shekar Krishnan
Employment: DKMS Life Science Lab India, a subsidiary of DKMS Life Science Lab Germany
Research Funding: Gennova Biopharmaceuticals Limited (Inst)

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

The authors thank the patients who volunteered to participate in this study and their families. The authors also thank all physicians, nurses, and other staff members who coordinated the clinical study.

REFERENCES

1. Hunger SP, Mullighan CG: Acute lymphoblastic leukemia in children. N Engl J Med 373:1541-1552, 2015
2. Radhakrishnan V, Gupta S, Ganesan P, et al: Acute lymphoblastic leukemia: A single center experience with Berlin, Frankfurt, and Munster-95 protocol. Indian J Med Paediatr Oncol 36:261-264, 2015
3. deAndrés-Galiana EJ, Fernández-Martínez JL, Luaces O, et al: Analysis of clinical prognostic variables for chronic lymphocytic leukemia decision-making problems. J Biomed Inform 60:342-351, 2016
4. Ghassemi M, Naumann T, Schulam P, et al: A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc 2020:191-200, 2020
5. Jian C, Chen S, Wang Z, et al: Predicting delayed methotrexate elimination in pediatric acute lymphoblastic leukemia patients: An innovative web-based machine learning tool developed through a multicenter, retrospective analysis. BMC Med Inform Decis Mak 23:148, 2023
6. Kelly CJ, Karthikesalingam A, Suleyman M, et al: Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17:195, 2019
7. Khan B, Fatima H, Qureshi A, et al: Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Biomed Mater Devices 1:731-738, 2023
8. Yu KH, Zhang C, Berry GJ, et al: Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 7:12474, 2016
9. Zhu H, Zhou Y, Shen D, et al: An interpretable machine learning model for predicting early liver metastasis after pancreatic cancer surgery. BMC Cancer 25:1117, 2025
10. Delen D, Walker G, Kadam A: Predicting breast cancer survivability: A comparison of three data mining methods. Artif Intell Med 34:113-127, 2005
11. Cruz JA, Wishart DS: Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2:59-77, 2007
12. Das N, Banavali S, Bakhshi S, et al: Protocol for ICiCLe-ALL-14 (InPOG-ALL-15-01): A prospective, risk stratified, randomised, multicentre, open label, controlled therapeutic trial for newly diagnosed childhood acute lymphoblastic leukaemia in India. Trials 23:102, 2022
13. Gogoi MP, Das P, Das N, et al: Risk stratified treatment for childhood acute lymphoblastic leukaemia: A multicentre observational study from India. Lancet Reg Health Southeast Asia 37:100593, 2025
14. Hosmer DW, Lemeshow S. Applied Logistic Regression. ed 1. New York: Wiley; 2000.
15. Cortes C, Vapnik V: Support-vector networks. Mach Learn 20:273-297, 1995
16. Bohra H, Arora A, Gaikwad P, et al: Health prediction and medical diagnosis using naive bayes. Int J Adv Res Comput Commun Eng 6:32-35, 2017
17. Quinlan JR: Induction of decision trees. Mach Learn 1:81-106, 1986
18. Breiman L: Random forests. Machine Learn 45:5-32, 2001
19. Chen T, Guestrin C: XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp 785-794
20. Prokhorenkova L, Gusev G, Vorobev A, et al: CatBoost: Unbiased boosting with categorical features. arXiv. 10.48550/ARXIV.1706.09516
21. Li N, Peng E, Liu F: Prediction of lymph node metastasis in cervical cancer patients using AdaBoost machine learning model: Analysis of risk factors. Am J Cancer Res 15:1158-1173, 2025
22. Liu C, Yang H, Feng Y, et al: A K-nearest neighbor model to predict early recurrence of hepatocellular carcinoma after resection. J Clin Transl Hepatol 10:600-607, 2022

23. Mandrekar JN: Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 5:1315-1316, 2010
24. Ekpa QL, Akahara PC, Anderson AM, et al: A review of acute lymphocytic leukemia (ALL) in the pediatric population: Evaluating current trends and changes in guidelines in the past decade. Cureus 15:e49930, 2023