

A Temporally Validated Machine Learning Framework for Predicting Anti-Tuberculosis Drug Stockouts: A Synthetic-Data Evaluation

Ahmad Asyhadi^{1*}, Rudolf Sinaga^{2*}, Widyadhana Candraningtias^{3**}

* Sistem Informasi, Universitas Dinamika Bangsa

** Manajemen, Universitas Dinamika Bangsa

aasyhadi26@gmail.com¹, rudolf@unama.ac.id², widyadhn@unama.ac.id³

Article Info

Article history:

Received 2026-09-04

Revised 2026-09-25

Accepted 2026-10-06

Keyword:

*Stockout Prediction;
Machine Learning;
Temporal Validation;
Pharmaceutical Inventory;
Anti-Tuberculosis Drugs.*

ABSTRACT

Continuous availability of anti-tuberculosis drugs is essential for maintaining tuberculosis treatment services, yet future stockouts arise from interacting demand, inventory, procurement, and supply conditions that are not fully represented by static stock thresholds. This study evaluates a temporally validated machine-learning framework for future anti-tuberculosis drug stockout prediction using synthetic longitudinal data as a methodological test bed prior to evaluation with operational hospital data. Drug-day predictors were derived from four domains: demand and consumption, inventory, procurement and supply, and historical stockouts. A 14-day horizon was used as the primary prediction target, with 7- and 30-day horizons considered for sensitivity analysis. Logistic Regression, Random Forest, XGBoost, and LightGBM were evaluated using temporally separated training, validation, and locked out-of-time test periods and compared with reorder-point and coverage-based inventory rules. Model assessment included recall, precision, specificity, F1-score, ROC-AUC, PR-AUC, Brier score, calibration, alert burden, temporal robustness, feature ablation, and predictive explanation. Logistic Regression was selected during validation but its recall decreased to 0.556 in the locked test period, whereas LightGBM descriptively achieved a recall of 0.889 and XGBoost a PR-AUC of 0.960. The results demonstrate the importance of locked temporal evaluation, multidomain predictors, threshold selection, and operationally oriented assessment. These findings validate the analytical pipeline rather than establish real-world effectiveness; external evaluation using SIMRS data remains required.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

Tuberculosis (TB) remains an infectious disease that requires continuous treatment over a relatively long period [1,2]. Treatment success depends not only on diagnosis and the treatment regimen, but also on the continuous availability of anti-tuberculosis drugs [3,4]. Therefore, the management of anti-tuberculosis drug supplies is a critical component of the TB care system, particularly in health facilities responsible for ensuring that medications are available when patients need them [5–7].

Drug shortages and stockouts are multidimensional problems [8–11], with causes categorized into supply, demand, and regulatory factors, including raw material

constraints, manufacturing and logistical challenges, changes in demand, and uncertainty regarding needs [8]. Data analytics studies in the pharmaceutical supply chain also demonstrate the potential of advanced analytics for demand forecasting, inventory optimization, and shortage management; however, the utilization of operational data sources remains suboptimal [12]. Thus, the risk of stockouts at the hospital level should be viewed as the result of the interaction between stock levels, consumption patterns, demand uncertainty, lead times, supply reliability, and inventory control policies [12–15].

OAT reported stockouts of first-line anti-TB medicines at health facilities and identified supply delays, shipment volumes lower than demand, and stockouts at supplier

institutions as key determinants [6,16]. These findings indicate that stockouts are not only related to internal inventory levels but also to the replenishment process and supply reliability.

Indicators such as current stock, safety stock, reorder point (ROP), and days of supply (DOS) remain operationally important. However, simple thresholds do not necessarily capture the nonlinear relationship between consumption, stock levels, demand variability, lead time, supplier performance, and stockout history. Inventory management, therefore, goes beyond simply identifying low stock levels; it also involves the ability to issue warnings before stockouts occur [14,15].

Machine-learning approaches have increasingly been applied to pharmaceutical forecasting and shortage prediction [17–22]. One of the studies, for example, demonstrated that historical pharmacy information can support advance prediction of drug shortages [19], while the other studies [21] developed a hospital drug-shortage risk model based on logistic regression and evaluated both discrimination and calibration. More recent studies have extended predictive modeling to broader shortage and regulatory settings [20,22]. Collectively, these studies establish the feasibility of machine-learning-based shortage prediction. However, their prediction targets, units of analysis, validation designs, and operational objectives differ from forecasting future anti-tuberculosis drug stockouts at the hospital inventory level.

Despite these developments, an important methodological gap remains. Existing studies have primarily demonstrated that machine learning can predict pharmaceutical shortages, but comparatively less attention has been given to a unified hospital-level framework that treats stockout as a future event, constructs predictors strictly from information available at the prediction origin, separates model development from a locked future test period, and evaluates performance from both statistical and operational perspectives. For anti-tuberculosis drug inventory, this distinction is particularly important because stockout risk reflects the joint behavior of demand, stock coverage, procurement, supplier performance, and previous shortage history [6,13–16]. Prediction systems intended for operational use therefore require more than discrimination alone: they also require leakage-safe temporal validation, probability calibration, threshold selection, alert-burden assessment, and interpretable risk estimates.

Accordingly, the scientific contribution of this study is not the comparison of machine-learning algorithms in isolation. Rather, it is the integration of several components into a single temporally structured stockout-prediction framework: (1) multidomain predictors covering demand, inventory, procurement, and previous stockout history; (2) leakage-aware feature construction at the drug-day prediction origin; (3) model development separated from a locked out-of-time evaluation; (4) comparison against transparent operational inventory rules; and (5) joint evaluation of discrimination, calibration, threshold-dependent alert burden, temporal robustness, feature-domain contribution, and predictive

explanations. A synthetic longitudinal dataset is used at this stage as a controlled methodological test bed for evaluating the complete analytical pipeline before external validation with real SIMRS data.

This study therefore aims to evaluate the feasibility and temporal behavior of the proposed stockout-prediction framework by comparing candidate machine-learning models and operational inventory rules, examining the contribution of predictor domains, assessing validation-to-test performance stability, and evaluating discrimination, calibration, threshold-dependent alert burden, and predictive explanations under a locked temporal evaluation design.

II. METHOD

A. Study Design

The study employed a longitudinal design with a time-series-based predictive evaluation to model OAT stockouts over a 14-day forecast horizon. Each observation represents the status of a single OAT at a single prediction time. The primary design used end-of-day predictions so that information up to the end of day t could be utilized, while the outcome began on the following day.

B. Synthetic Data Generation

The methodological experiment used a synthetic longitudinal drug-inventory dataset organized at the drug-day level. One observation represents one anti-tuberculosis drug at one prediction date. The dataset contains variables from four operational domains: demand/consumption, inventory status, procurement/supply, and previous stockout history. The synthetic design was used to test the end-to-end prediction pipeline under controlled temporal conditions before external validation with routinely collected SIMRS data [23]. Consequently, the synthetic observations should not be interpreted as a statistical sample of a hospital population or as empirical estimates of actual stockout prevalence.

C. Data Sources and Population

The dataset used in the experiment includes variables on consumption, inventory, procurement, and OAT stockout history at the drug-day unit of analysis. The data are organized in a longitudinal format to allow for the creation of predictors at time t and outcomes in the period following the prediction time. The unit of analysis used is the drug-day, meaning one observation for each drug at each observation time.

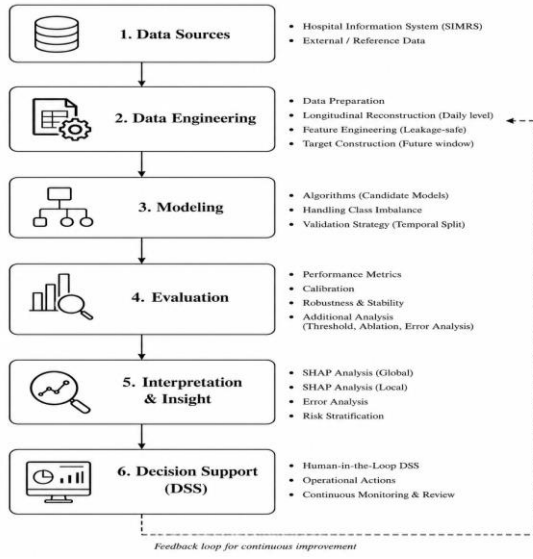


Figure 1. Research Methodology Workflow.

D. Outcome and Prediction Horizons

For the primary 14-day analysis, observations were ordered chronologically and partitioned without random shuffling. The training period covered January 31–March 20, 2025, the validation period March 21–April 5, 2025, and the locked out-of-time test period April 6–20, 2025. Model fitting and preprocessing parameters were estimated using training data; hyperparameter and operating-threshold decisions were made using development/validation data only. The final test period remained inaccessible to model selection, feature selection, imputation fitting, resampling, calibration fitting, and threshold optimization.

For the synthetic experiment, a stockout was operationally defined as a drug-day on which usable stock on hand reached zero [24]:

$$Y_{i,t}(H) = \mathbb{I}(\exists k \in \{1, \dots, H\}: Stock_{i,t+k} = 0) \quad (2)$$

where $H \in \{7, 14, 30\}$ days. Thus, the information on day t is used to predict stockouts from $t+1$ through $t+H$, ensuring that there is no overlap between the prediction time and the future outcome window.

E. Predictor Domains

The predictor is built from four domains: demand/consumption, inventory, procurement/supply, and historical stockouts. Candidate features include lagged consumption, 7/14/30-day moving averages, demand variability/trend, stock on hand, DOS, safety-stock gap, ROP gap, lead time, lead-time variability, ordered/received quantity, supplier fill rate, delayed/partial delivery history, pending POs, previous stockout count, and days since the last stockout.

Days of Supply (DOS):

$$DOS_{i,t} = \frac{StockOnHand_{i,t}}{AverageDailyConsumption_{i,t}} \quad (3)$$

Reorder Point Gap:

$$ROPGap_{i,t} = StockOnHand_{i,t} - ReorderPoint_{i,t} \quad (4)$$

Coverage–Lead-Time Gap:

$$CoverageLeadTimeGap_{i,t} = DOS_{i,t} - ExpectedLeadTime_{i,t} \quad (5)$$

F. Leakage-Safe Feature Engineering

Every predictor was timestamped relative to the prediction origin t . Stock on hand and procurement status represented information available at the close of day t ; lagged consumption and stockout-history variables were calculated only from dates $\leq t$; moving averages used trailing rather than centered windows; and supplier fill-rate or lead-time summaries were calculated only from completed deliveries known by t . Realized quantities, delivery dates, lead times of still-open orders, consumption after t , future stock levels, and any variable derived from the target window were excluded [25]. All predictors must satisfy:

$$Timestamp(X_j) \leq t \quad (6)$$

Future consumption, future stock, future receipt, realized lead time for uncompleted orders, future stockout information, and label-derived risk scores are not used. Trailing windows are used; centered moving averages are prohibited. Figure 2 illustrates the temporal separation between predictor information available at prediction time t and the subsequent future-outcome window used to define the stockout target.

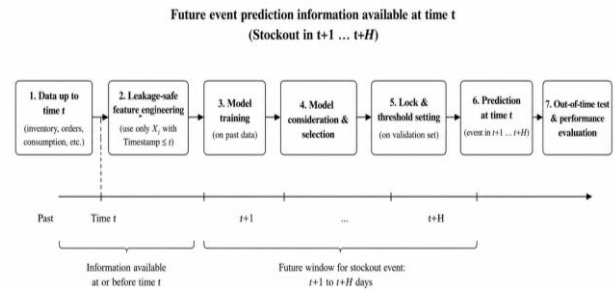


Figure 2. Temporal prediction/leakage design

G. Missing Data

Structural zeros are distinguished from missing information. Imputation is performed only on the training data. The median may be used for appropriate continuous features; categorical missing values may be assigned a separate category. A missingness indicator is considered if it is informative.

H. Temporal Partitioning

A random train-test split is not used as the primary strategy. The dataset is divided by time into training, validation, and locked-out-of-time test sets. The test period is not used for feature selection, imputation fitting, hyperparameter tuning, resampling, threshold selection, or calibration fitting.

Expanding-window validation is used if the event count allows it.

I. Class Imbalance

Stockout prediction was treated as an imbalanced-event problem because future stockout observations were less frequent than non-stockout observations. The natural event prevalence was preserved in the validation and locked out-of-time test sets so that model performance reflected the expected class distribution during prospective evaluation. Model development incorporated imbalance handling through validation-based operating-threshold selection, with recall, precision, specificity, and alert burden jointly considered rather than accuracy alone. Any class weighting or resampling procedure was restricted to the training data and was not applied to the validation or locked test periods. PR-AUC was emphasized because it reflects performance on the minority stockout class and was interpreted relative to the corresponding stockout prevalence.

J. Prediction Models

Four models were compared using the same set of predictors and temporal framework: Logistic Regression as an interpretable statistical baseline; Random Forest; XGBoost; and LightGBM. Complex models were not assumed to be superior prior to the out-of-time evaluation.

K. Operational Baselines

Machine learning is compared to simple inventory rules, specifically $\text{Stock} < \text{ROP}$ and $\text{DOS} < \text{expected lead time}$. This comparison evaluates the incremental predictive value of ML relative to transparent and easy-to-apply rules.

$$\begin{aligned} \text{StockOnHand}_{i,t} &< \text{ROP}_{i,t} \\ \text{DOS}_{i,t} &< \text{ExpectedLeadTime}_{i,t} \end{aligned} \quad (7)$$

where $\text{Stock}_{i,t}$ denotes current usable inventory, ROP_t the reorder point, $\text{DOS}_{i,t}$ days of supply, and $\text{ELT}_{i,t}$ expected replenishment lead time. These rules represent operationally interpretable benchmarks against which the incremental predictive value and alert behavior of the machine-learning models were evaluated.

L. Hyperparameter and Threshold Optimization

Hyperparameter tuning is performed only on the development data. Accuracy is not the primary tuning objective. The 0.50 threshold is used only for comparison; the operating threshold is selected from the validation data by considering the minimum acceptable recall, precision, specificity, and alert burden, and is then locked before the out-of-time test.

M. Model Evaluation

Key metrics include recall/sensitivity for the stockout class, precision, specificity, F1-score, ROC-AUC, PR-AUC, confusion matrix, and alert burden. Accuracy is only supplementary. PR-AUC is emphasized for minority-event prediction. For correctly detected stockout events, operational warning lead time was defined as the number of days between the first model-generated alert and the first subsequent stockout occurring within the prediction horizon.

Since false negatives represent stockouts that failed to receive an early warning, recall is one of the key operational metrics [26]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

where TP is the number of stockout events successfully detected and is the number of stockout events not detected by the model. Then there are Precision, Specificity, and F1-score.

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \quad | \quad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad | \quad \text{F1} \\ &= 2 \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \end{aligned} \quad (9)$$

In addition to conventional classification metrics, operational usefulness was evaluated using the number and proportion of generated alerts, false alerts, and missed future stockout events. Alert burden was defined as the proportion of evaluated observations classified as future stockout risk, $(\text{TP} + \text{FP})/N$. The false-alert proportion was defined as $\text{FP}/(\text{TP} + \text{FP})$, corresponding to the proportion of generated alerts that were not followed by a stockout event within the prediction horizon. The missed-stockout proportion was defined as $\text{FN}/(\text{TP} + \text{FN})$, equivalent to $1 - \text{recall}$. These measures were examined jointly because lowering the operating threshold may increase stockout detection while also increasing the number of alerts requiring operational review. Conversely, a restrictive threshold may reduce unnecessary alerts at the cost of missing future stockouts.

False negatives were considered operationally more consequential than false positives because they represent future stockout events for which no early warning was generated. However, no monetary cost function was assigned to false-positive or false-negative predictions in the present synthetic experiment because empirically validated hospital-specific cost estimates were unavailable. Consequently, the trade-off was evaluated through recall, precision, alert burden, and the absolute numbers of missed stockouts and false alerts rather than through an economic utility function. Because adjacent drug-day observations and forecast windows may be temporally dependent, these uncertainty estimates are interpreted as measures of estimation precision rather than evidence of independent observations. This validation-based threshold selection was particularly important under class imbalance because a fixed 0.50 cutoff could produce an operating point that was inconsistent with

the study's emphasis on detecting future stockout events while limiting unnecessary alerts.

N. Calibration

Probability calibration was evaluated separately from discrimination because a model with good ranking performance may still produce probabilities that do not accurately represent the observed frequency of future stockout events. Calibration was assessed using a reliability plot and the Brier score [27]. The reliability plot compares predicted stockout probabilities with the observed event frequency, while the Brier score quantifies the mean squared error between predicted probabilities and binary outcomes:

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2 \quad (10)$$

Where p_i is the predicted probability of a stockout for the i -th observation, y_i is the actual outcome (0/1), and N is the number of evaluation observations. A lower Brier score indicates a more accurate probabilistic prediction. Calibration intercept and slope were considered only when the number of events was sufficient for stable estimation. Any recalibration procedure, such as Platt scaling or isotonic regression, would be fitted exclusively using development or validation data and never using the locked out-of-time test set.

O. Ablation and Robustness

Feature ablation was performed using four nested predictor sets to assess the incremental predictive contribution of each information domain [28]: (1) demand/consumption only; (2) demand plus inventory; (3) demand, inventory, and procurement/supply; and (4) the full feature set including historical stockout information. Each feature set was evaluated using the same temporal validation framework and performance metrics. The order of addition reflects increasing operational information available for stockout prediction. Because the experiment used a small synthetic validation sample, differences between feature sets were interpreted descriptively rather than as formal evidence of statistical superiority. Additional robustness analyses examined prediction horizons and successive future temporal windows.

P. Explainable AI and Error Analysis

SHAP is applied to the best tree-based models for global and local predictive explanations. False negatives are the primary focus, as they represent missed early warnings; false positives are analyzed to understand unnecessary alerts and the alert burden [29].

Q. Reproducibility, Ethics, and Governance

The analytics pipeline documents random seeds, package versions, feature definitions, search spaces, thresholds, and exclusion rules. The use of operational data during the external validation phase must comply with the institution's

ethical approval and data governance requirements, including the principles of data minimization and de-identification in accordance with applicable policies. The model is designed as an early-warning decision support tool, not an autonomous procurement system.

III. RESULTS AND DISCUSSION

The results in this section are derived from a synthetic dataset and the uploaded experiment notebook. Because the data is synthetic, the following figures should be treated as simulation results or pipeline validation results, rather than as final estimates of clinical and operational performance on the SIMRS. All key performance claims must be re-evaluated on the real SIMRS dataset using the same temporal protocol.

A. Dataset Characteristics

The synthetic dataset consists of 320 drug-day observations for four anti-TB drugs during the period January 31–April 20, 2025. The primary target variable, `stockout_next_14d`, consists of 48 positive observations (15.0%) and 272 negative observations (85.0%). No duplicate drug-day rows or missing values were found in the initial audit. A leakage audit of 38 candidate predictors did not detect any suspected leakage columns.

TABLE I
TEMPORAL PARTITION AND STOCKOUT PREVALENCE

Partition	Periode	N	Stockout events	Prevalence
Training	31 Jan–20 Mar 2025	196	29	14.8%
Validation	21 Mar–5 Apr 2025	64	10	15.6%
Out-of-time test	6–20 Apr 2025	60	9	15.0%

The number of events in each partition is relatively limited, particularly in the validation and out-of-time tests; therefore, differences in metrics across periods should be interpreted with consideration of the size of the available events.

B. Operational Baseline Performance

On the validation set, neither the reorder-point rule nor the coverage rule generated a positive alert. Consequently, both baselines produced a recall of 0, precision of 0, specificity of 1.000, F1-score of 0, and an alert rate of 0. These results indicate that, under the parameterization of the current synthetic experiment, the two threshold-based inventory rules did not provide sufficiently early signals for the 14-day future-stockout target.

This result should not be interpreted as evidence that reorder-point or stock-coverage rules are ineffective for inventory management. Rather, these rules were designed primarily to trigger action when current inventory conditions cross predefined operational limits, whereas the prediction task in this study asks whether a stockout will occur within a future time window. The comparison therefore provides a

deliberately transparent benchmark for assessing whether models that combine demand, inventory, procurement, and historical stockout information can identify future risk earlier than contemporaneous threshold rules [13–15]. Because the present experiment uses synthetic data, the relative performance of these baselines and the machine-learning models must be reassessed using operational SIMRS data before conclusions about practical superiority are made.

C. Validation Model Performance

At the model's initial threshold, all models achieved a recall of 1.00 on the validation set. Logistic Regression yielded an ROC-AUC of 0.983 and a PR-AUC of 0.929; XGBoost had an ROC-AUC of 0.972 and a PR-AUC of 0.847; LightGBM had an ROC-AUC of 0.956 and a PR-AUC of 0.732; and Random Forest had an ROC-AUC of 0.939 and a PR-AUC of 0.725. Precision ranged from 0.417 to 0.476. Based on the notebook procedure, Logistic Regression was selected during the validation stage, and its threshold was set to 0.991328. Logistic Regression was selected based on the validation results, while performance on the out-of-time test was used as a separate evaluation and was not used to reselect the model.

TABLE II
VALIDATION PERFORMANCE OF CANDIDATE PREDICTION MODELS

Model	Precision	Recall	Specificity	F1	ROC-AUC	PR-AUC	Brier
LR	0.417	1.000	0.741	0.588	0.983	0.929	0.197
RF	0.435	1.000	0.759	0.606	0.939	0.725	0.118
XGBoost	0.476	1.000	0.796	0.645	0.972	0.847	0.125
LightGBM	0.435	1.000	0.759	0.606	0.956	0.732	0.188

Unlike the two operational baselines, which generated no validation alerts, all four machine-learning models identified positive future-stockout observations at their evaluated operating points. This comparison suggests that combining multiple predictor domains may provide additional early-warning information beyond the two contemporaneous threshold rules in the present synthetic experiment.

D. Threshold Optimization

The threshold was selected on the validation data based on a trade-off recall, precision, and alert rate. Logistic Regression used a threshold of 0.9913 (recall 0.80; precision 0.889; alert rate 0.141). Random Forest used 0.6920 (recall 0.90; precision 0.562; alert rate 0.250). XGBoost used 0.9903 (recall 0.80; precision 0.727; alert rate 0.172), while LightGBM used 0.9996 (recall 0.80; precision 0.667; alert rate 0.188).

TABLE III
VALIDATION-BASED OPERATING THRESHOLD SELECTION

Model	Selected threshold	Recall	Precision	Alert rate
Logistic Regression	0.9913	0.800	0.889	0.141
Random Forest	0.6920	0.900	0.562	0.250
XGBoost	0.9903	0.800	0.727	0.172
LightGBM	0.9996	0.800	0.667	0.188

E. Locked Out-of-Time Test

In the out-of-time test ($n=60$; 9 stockout events), LightGBM achieved the highest recall of 0.889 with a precision 0.800, specificity 0.961, F1 0.842, ROC-AUC 0.983, PR-AUC 0.889, Brier score 0.055, and alert rate 0.167. XGBoost achieved an ROC-AUC 0.991 and PR-AUC 0.960 with a precision 1.00 but a recall 0.556 at a locked threshold. Random Forest achieved recall 0.556, precision 0.833, PR-AUC 0.909, and Brier 0.053. Logistic Regression—the model selected during the validation stage yielded a recall 0.556, precision 1.00, specificity 1.00, PR-AUC 0.903, and Brier 0.117. The difference in performance between the validation and out-of-time tests indicates that a model selected based on validation results may not necessarily maintain the same performance characteristics in subsequent periods. The results from the out-of-time test are used to evaluate temporal generalization and not as a basis for selecting a new model.

TABLE IV
LOCKED OUT-OF-TIME PERFORMANCE OF CANDIDATE PREDICTION MODELS

Model	Threshold	Precision	Recall (95% CI)	Specificity	F1	ROC-AUC	PR-AUC	Brier	Alert rate
LR	0.9913	1.000	0.556 (0.212–0.863)	1.000	0.714	0.980	0.903	0.117	0.083
RF	0.6920	0.833	0.556 (0.212–0.863)	0.980	0.667	0.985	0.909	0.053	0.100
XGBoost	0.9903	1.000	0.556 (0.212–0.863)	1.000	0.714	0.991	0.960	0.054	0.083
LightGBM	0.9996	0.800	0.889 (0.518–0.997)	0.961	0.842	0.983	0.889	0.055	0.167

The uncertainty estimates further demonstrate the effect of the small number of positive events in the locked test period. Logistic Regression, Random Forest, and XGBoost each detected five of nine stockout events, corresponding to a recall of 0.556 with an exact 95% confidence interval of 0.212–0.863. LightGBM detected eight of nine events, corresponding to a recall of 0.889 with a 95% confidence interval of 0.518–0.997. Although the point estimate for LightGBM was substantially higher, the wide confidence intervals indicate considerable uncertainty arising from the limited number of stockout events. Consequently, the test-set

differences should be interpreted descriptively rather than as evidence of statistically established model superiority.

Previous studies have demonstrated the feasibility of predicting pharmaceutical shortages using historical pharmacy or shortage data [19,21,22]. The present findings complement this evidence by showing that favorable model performance during development does not necessarily remain unchanged in a subsequent temporal period. This distinction is particularly relevant for operational stockout prediction, where prospective predictions must be generated using only information available before the future outcome occurs.

F. Confusion Matrix and Operational Alert Characteristics

For Logistic Regression with a fixed threshold, the confusion matrix consists of TN=51, FP=0, FN=4, and TP=5. Thus, the model did not generate any false-positive alerts during the test period, but missed four out of nine future stockout events. The accuracy of 0.933 is not used as the primary basis for interpretation because the target class is imbalanced. However, the recall value during the test period must be interpreted with caution because the number of positive events available in the out-of-time test is relatively small.

TABLE V

CONFUSION MATRIX OF THE VALIDATION-SELECTED LOGISTIC REGRESSION

	Predicted non-stockout	Predicted stockout
Actual non-stockout	51	0
Actual stockout	4	5

For the validation-selected Logistic Regression, the locked out-of-time confusion matrix consisted of 51 true negatives, no false positives, four false negatives, and five true positives. Thus, five alerts were generated among 60 test observations, corresponding to an alert burden of 8.3%. None of these alerts were false, but four of the nine future stockout events were missed. Consequently, the model detected 55.6% of future stockout events while failing to provide an early warning for 44.4%. Although the resulting accuracy was 0.933, accuracy was not treated as the primary operational measure because 85.0% of the locked test observations represented the non-stockout class. In this setting, a high overall accuracy may coexist with a clinically and operationally important number of missed stockout events.

The operational trade-off differed across the candidate models. Logistic Regression and XGBoost each generated five alerts among 60 observations (8.3% alert burden), with no false alerts, but each missed four of the nine stockout events. Random Forest generated six alerts (10.0%), including one false alert, and also missed four stockout events. LightGBM generated ten alerts (16.7%), of which eight were true stockout alerts and two were false alerts, while missing only one of the nine future stockout events. Therefore, the higher recall of LightGBM was accompanied by approximately twice the alert burden of Logistic Regression and XGBoost. These results illustrate that model

usefulness cannot be assessed from discrimination metrics alone: an operating model must balance the number of missed stockouts against the number of alerts requiring operational review.

G. Calibration

Calibration was evaluated together with discrimination to determine whether predicted stockout probabilities were sufficiently reliable for risk prioritization. In the locked out-of-time test, Logistic Regression produced a Brier score of 0.117, whereas Random Forest, XGBoost, and LightGBM produced lower Brier scores of 0.053, 0.054, and 0.055, respectively. Thus, although Logistic Regression was selected during the validation procedure, its probability predictions showed greater overall probabilistic error in the subsequent test period than those of the three tree-based models. The calibration plot for Logistic Regression is presented in Figure 6 and shows departures from the line of perfect calibration, indicating that predicted probabilities were not consistently matched by the observed event frequencies across the available probability ranges. The calibration behavior of the validation-selected Logistic Regression on the locked test period is illustrated in Figure 6.

These findings should be interpreted cautiously because the locked test period contained only nine positive stockout events. With such a small event count, reliability estimates within individual probability ranges are unstable and may change substantially with only a few observations. Therefore, the calibration results are treated as descriptive evidence of probabilistic behavior rather than as a precise estimate of deployment calibration. No recalibration was performed using the locked test data because doing so would compromise the independence of the final temporal evaluation.

From an operational perspective, calibration is relevant because predicted probabilities may be used to rank or prioritize inventory interventions. A probability reported as high risk should ideally correspond to a comparably high observed frequency of future stockout events. Therefore, discrimination and calibration serve different purposes: ROC-AUC and PR-AUC describe the ability to rank higher-risk observations above lower-risk observations, whereas calibration evaluates whether the magnitude of the predicted risks is credible. The present results demonstrate that strong discrimination alone should not be interpreted as sufficient evidence that model probabilities are ready for operational decision thresholds. For context, a constant-probability forecast equal to the locked-test stockout prevalence of 0.150 would yield a Brier score of approximately 0.128. All four models produced lower Brier scores than this prevalence-based reference, although the improvement was substantially greater for Random Forest, XGBoost, and LightGBM than for Logistic Regression. This comparison provides context for probabilistic prediction error but does not replace formal calibration assessment.

H. Feature Ablation and Robustness

Feature ablation was used to examine whether progressively adding inventory, procurement, and historical stockout information improved predictive performance beyond demand features alone. Table VI summarizes the results for the four nested feature sets.

The strong improvement following the addition of inventory and procurement features is also operationally plausible because hospital stockout risk is influenced not only by consumption but also by stock coverage, replenishment timing, lead-time uncertainty, and supply reliability [13–16]. The present ablation results should nevertheless be interpreted as evidence from the simulated data-generating process rather than as proof that the same relative feature contributions will be observed in real SIMRS data.

TABLE VI
FEATURE ABLATION PERFORMANCE

Feature set	Precision	Recall	Specificity	F1	ROC-AUC	PR-AUC	Brier	Alert rate
Demand only	0.213	1.000	0.315	0.351	0.557	0.175	0.276	0.734
Demand + Inventory	0.571	0.800	0.889	0.667	0.915	0.591	0.129	0.219
+ Procurement	0.667	0.800	0.926	0.727	0.965	0.804	0.113	0.188
Full + Historical stockout	0.889	0.800	0.981	0.842	0.974	0.861	0.116	0.141

The largest performance gain occurred after inventory information was added to the demand-only feature set: PR-AUC increased from 0.175 to 0.591 and ROC-AUC from 0.557 to 0.915. Adding procurement information further increased PR-AUC to 0.804 and ROC-AUC to 0.965, while adding historical stockout information produced a smaller additional increase to 0.861 and 0.974, respectively. The same progression reduced alert burden from 0.734 for demand-only predictors to 0.141 for the full feature set, while precision increased from 0.213 to 0.889. These results suggest that inventory and procurement information contributed substantial incremental predictive information beyond consumption history in the synthetic experiment. However, the differences are descriptive and should not be interpreted as statistically established superiority because the ablation analysis was conducted on a limited synthetic validation sample.

I. Temporal Robustness

Expanding the temporal windows revealed variations in performance. In the first window ($n=32$; 3 events), recall was 0.333 and PR-AUC was 0.494. In the second window ($n=32$; 12 events), recall increased to 1.00 and PR-AUC was 0.905.

In the third window ($n=32$; 4 events), recall remained at 1.00 with a PR-AUC of 0.888. These variations confirm that aggregate performance alone is insufficient to assess temporal robustness, especially with a small number of events.

TABLE VII
TEMPORAL ROBUSTNESS ACROSS SUCCESSIVE FUTURE WINDOWS

Window	n	Events	Precision	Recall	Specificity	F1	ROC-AUC	PR-AUC	Brier
1	32	3	0.250	0.333	0.897	0.286	0.816	0.494	0.105
2	32	12	0.632	1.000	0.650	0.774	0.929	0.905	0.198
3	32	4	0.308	1.000	0.679	0.471	0.982	0.888	0.215

Performance varied substantially across the successive temporal windows, but this variation should not be interpreted as formal evidence of temporal drift. The three windows contained only 3, 12, and 4 positive events, respectively, so changes in recall and PR-AUC may partly reflect the instability of point estimates under small event counts. The analysis therefore demonstrates the importance of evaluating model behavior across successive future periods rather than relying solely on aggregate performance, while larger operational datasets are required to distinguish true temporal distribution shifts from sampling variation.

J. Explainability and Error Analysis

The notebook generates global and local SHAP explanations for tree-based models. The case with the highest risk according to the numeric-only SHAP model is Ethambutol on April 13, 2025, with a predicted risk of 0.9974. In the locked Logistic Regression test predictions, the error distribution consists of 51 true negatives, 5 true positives, and 4 false negatives; there are no false positives. The four false negatives had predicted probabilities ranging from 0.9386 to 0.9882, but remained below the locked threshold of 0.9913. This indicates that a very high threshold can suppress false positives while simultaneously increasing the number of missed warnings.

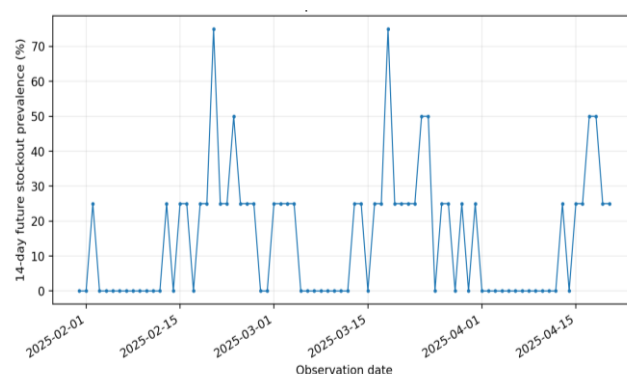


Figure 3. Stockout prevalence over time for the 14-day future-stockout target.

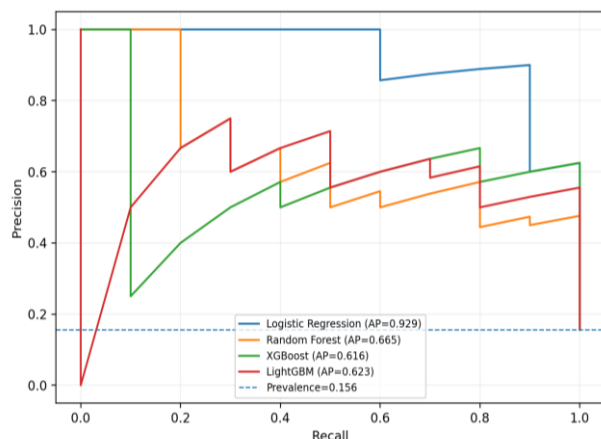


Figure 4. Precision–Recall curves on the validation period.

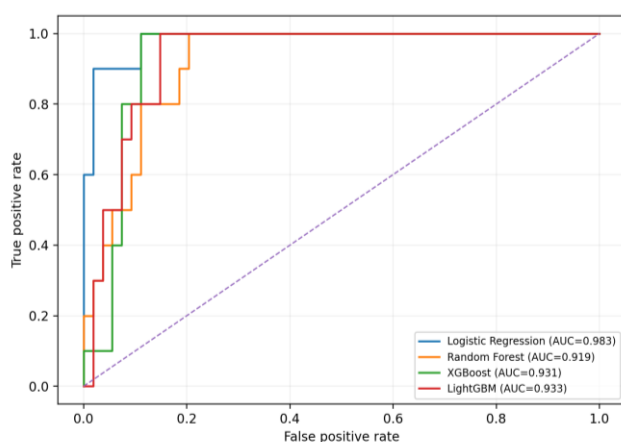


Figure 5. ROC curves on the validation period.

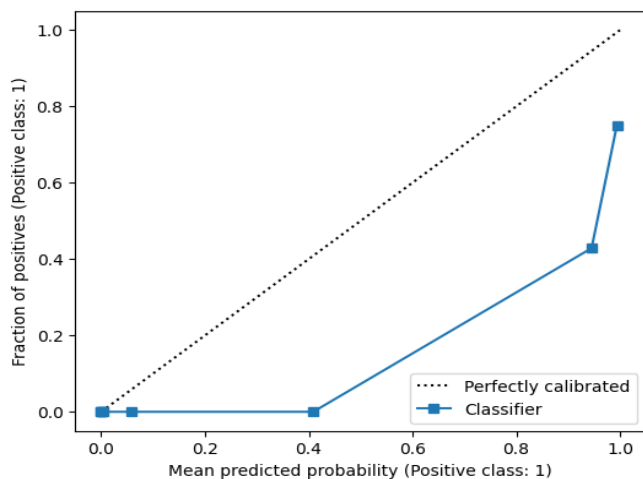


Figure 6. Reliability plot of the validation-selected Logistic Regression evaluated on the locked out-of-time test.

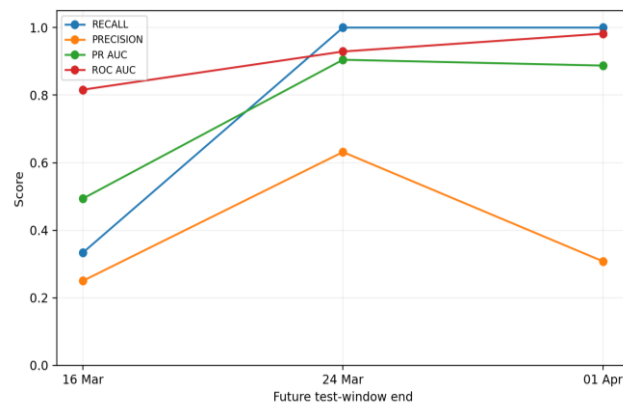


Figure 7. Temporal performance of Logistic Regression across successive future windows.

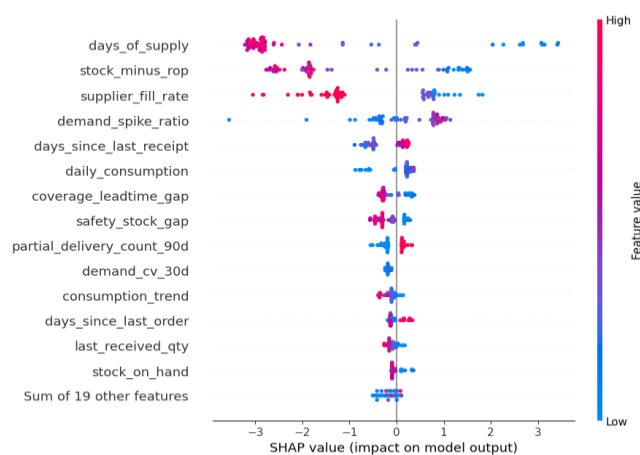


Figure 8. SHAP summary plot from the numeric-only XGBoost explanation model.

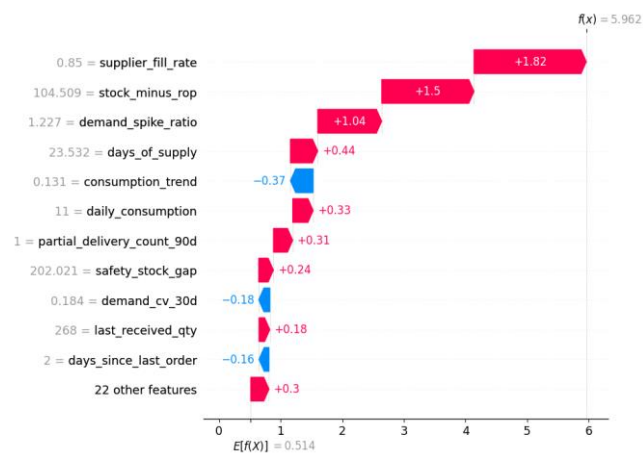


Figure 9. SHAP local explanation for the highest predicted-risk test observation in the numeric-only XGBoost explanation model.

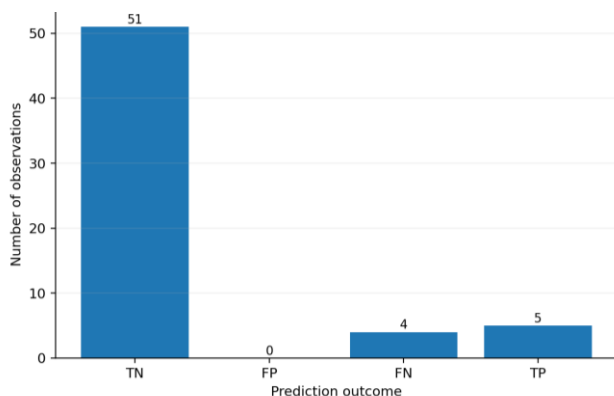


Figure 10. Error-analysis distribution for the validation-selected Logistic Regression on the locked out-of-time test.

The SHAP results in Figures 8 and 9 are interpreted as model-specific predictive explanations rather than evidence of causal effects. Variables such as supplier fill rate, stock relative to the reorder point, demand-related features, and days of supply contributed to the XGBoost predictions for individual observations, but the analysis does not establish that modifying one of these variables would necessarily prevent or induce a stockout. In addition, correlations among inventory and procurement variables may redistribute SHAP attribution across related predictors. The SHAP analysis therefore supports interpretation of model behavior rather than causal inference regarding stockout determinants.

K. Pipeline Sanity Check

The label-permutation sanity check yielded an ROC-AUC of 0.306. Because discrimination is expected to approach chance performance after the predictor–outcome relationship is disrupted, this result indicates that the strong discrimination observed in the original pipeline did not persist under permuted labels. The value below 0.5 is interpreted as finite-sample variation rather than as substantively meaningful inverse prediction. Accordingly, the permutation experiment is used only as a diagnostic check of the analytical pipeline and not as an estimate of predictive performance.

L. Summary of Synthetic-Data Results

Overall, the synthetic experiments show that the pipeline is capable of performing temporal partitioning, leakage audit, model comparison, threshold locking, out-of-time evaluation, calibration, ablation, temporal robustness, SHAP, and error analysis. However, the results also highlight the importance of distinguishing between model selection on the validation set and descriptive comparison on the test set: Logistic Regression was selected and locked based on the validation procedure, while LightGBM descriptively showed the highest recall and F1 scores on the out-of-time test. Since the test set must not be used for model reselection, LightGBM should not be declared the final model based on those test results. All of these figures must be confirmed on the real SIMRS dataset.

IV. LIMITATIONS

1. The experiment uses a synthetic dataset, so the data characteristics do not yet encompass the full range of variation, noise, missing values, and operational dynamics found in hospital inventory systems.
2. Upstream variables such as manufacturing disruptions and national distribution constraints may not be available.
3. Rare-event imbalance can limit precision and increase false-positive alerts.
4. Repeated drug-time observations and overlapping outcome windows introduce dependence.
5. Engineered inventory features may be correlated with one another, thereby affecting the interpretation of SHAP.
6. SHAP and feature importance are predictive, not causal.
7. Temporal validation does not replace external and prospective validation.
8. The study evaluates prediction performance, not intervention effect; predictive ability does not prove that the model reduces stockouts.
9. Risk thresholds are setting-specific and must be tailored to operational capacity.
10. The current experiment evaluates whether a future stockout was detected within a predefined horizon but does not yet quantify the exact number of days gained between the first predictive alert and the realized stockout. This operational lead-time measure should be included in the subsequent SIMRS evaluation

V. CONCLUSION

This study evaluated a temporally structured machine-learning framework for predicting future anti-tuberculosis drug stockouts using synthetic longitudinal data. The framework integrates multidomain predictors, leakage-aware feature construction, operational inventory baselines, validation-based threshold locking, out-of-time testing, calibration, alert-burden assessment, temporal robustness, feature ablation, and predictive explanation. Logistic Regression was selected during validation, but its recall decreased to 0.556 in the locked test period, demonstrating that validation performance did not fully transfer to a subsequent temporal period. LightGBM descriptively achieved the highest test recall of 0.889, while XGBoost achieved the highest PR-AUC of 0.960; however, these test results were not used for post-hoc model reselection. The ablation experiment further indicated that inventory and procurement information contributed substantial predictive information beyond demand features alone. These findings support the methodological feasibility of the proposed analytical pipeline rather than establish its clinical or operational effectiveness. Evaluation using real SIMRS data, uncertainty analysis on larger temporal samples, external validation, and prospective assessment of operational alert usefulness are required before implementation as a hospital stockout early-warning system.

REFERENCES

- [1] Marchelli LP. Recurrent Tuberculosis Due To Lack Of Continuity Of Treatment: Evidence-Based Analysis. *The Brazilian Journal of Infectious Diseases* 2026;30:105767. <https://doi.org/https://doi.org/10.1016/j.bjid.2026.105767>.
- [2] Sazali MF, Rahim SSSA, Mohammad AH, Kadir F, Payus AO, Avoi R, et al. Improving Tuberculosis Medication Adherence: The Potential of Integrating Digital Technology and Health Belief Model. *Tuberc Respir Dis (Seoul)* 2023;86:82–93. <https://doi.org/10.4046/trd.2022.0148>.
- [3] Dartois VA, Rubin EJ. Anti-tuberculosis treatment strategies and drug development: challenges and priorities. *Nat Rev Microbiol* 2022;20:685–701. <https://doi.org/10.1038/s41579-022-00731-y>.
- [4] Kėvelaitienė K, Davidavičienė VE, Danila E. Tuberculosis treatment failure: what are the risk factors? A comprehensive literature review. *Multidiscip Respir Med* 2025;20. <https://doi.org/10.5826/mrm.2025.1030>.
- [5] Tola FB, Anbessa GT, Yikna BB. Anti-Tuberculosis Commodities Management Performance and Factors Affecting It at Public Health Facilities in Dire Dawa City Administration, Ethiopia. *J Multidiscip Healthc* 2020;Volume 13:1677–91. <https://doi.org/10.2147/JMDH.S280253>.
- [6] Sintayehu K, Zeleke ED, Temesgen B, Kifle M, Assefa DG, Zenebe K, et al. Determinants of stock-outs of first line anti-tuberculosis drugs: the case of public health facilities of Addis Ababa city administration health bureau, Addis Ababa, Ethiopia. *BMC Health Serv Res* 2022;22:1047. <https://doi.org/10.1186/s12913-022-08430-3>.
- [7] Yang J, Yu X, Yao Y, Zhao W, Chen Y, Wang Z, et al. Analysis of the impact of national centralized volume-based procurement drug policy on the market dynamics of tuberculosis drugs. *BMC Health Serv Res* 2026;26:756. <https://doi.org/10.1186/s12913-026-14668-y>.
- [8] Shukar S, Zahoor F, Hayat K, Saeed A, Gillani AH, Omer S, et al. Drug Shortage: Causes, Impact, and Mitigation Strategies. *Front Pharmacol* 2021;12. <https://doi.org/10.3389/fphar.2021.693426>.
- [9] Antari NPU, Sujarna IGW, Sari NWRP, Fatoni N, Damayanti NL, Suena NMDS. Roots Cause of the Drug Shortage and Expired Drug Problem at Hospital X. *JMMR (Jurnal Medicoeticolegal Dan Manajemen Rumah Sakit)* 2024;13:67–84. <https://doi.org/10.18196/jmmr.v13i1.46>.
- [10] Cornelissen N, Zielhuis SW, van den Bemt PMLA, van den Bemt BJF. Causes and management of drug shortages: A scoping review. *Research in Social and Administrative Pharmacy* 2026;22:232–53. <https://doi.org/https://doi.org/10.1016/j.sapharm.2025.10.008>.
- [11] Adak S. Impacts of Drug Shortages in the Pharmaceutical Supply Chain. *Universal Journal of Pharmacy and Pharmacology* 2024;3:22. <https://doi.org/10.31586/ujpp.2024.1136>.
- [12] Nguyen A, Lamouri S, Pellerin R, Tamayo S, Lekens B. Data analytics in pharmaceutical supply chains: state of the art, opportunities, and challenges. *Int J Prod Res* 2022;60:6888–907. <https://doi.org/10.1080/00207543.2021.1950937>.
- [13] Anwar AA, Cahyo LM, Wijayanti T. Inventory Control of Medicine with Minimum-Maximum Stock Level at Ibu Fatmawati Soekarno Hospital Surakarta in 2023. *Jurnal Manajemen Dan Pelayanan Farmasi (Jurnal of Management and Pharmacy Practice)* 2025;15:186. <https://doi.org/10.22146/jmpf.98812>.
- [14] Hediarta Widiana Putra P, Ari Kristina S. Efficiency of Drug Inventory Control in Hospitals in Indonesia Using the Minimum-Maximum Stock Level: A Systematic Review. *106600 | JMPF* 2026;16:2026. <https://doi.org/10.22146/jmpf.106600>.
- [15] Oeser G, Romano P. Exploring risk pooling in hospitals to reduce demand and lead time uncertainty. *Operations Management Research* 2021;14:78–94. <https://doi.org/10.1007/s12063-020-00171-y>.
- [16] Shourabizadeh H, Kundakcioglu OE, Bozkir CDC, Tufekci MB, Henry AC. Healthcare inventory management in the presence of supply disruptions and a reliable secondary supplier. *Ann Oper Res* 2023;331:1149–206. <https://doi.org/10.1007/s10479-023-05620-y>.
- [17] Ramadhan W, Noersasongko E, Syukur A, Soeleman MA. Machine Learning Approaches for Pharmaceutical Demand Forecasting: A Bibliometric and Systematic Review of Methods and Research Trends. *Ingenierie Des Systemes d'Information* 2026;31:179–91. <https://doi.org/10.18280/isi.310117>.
- [18] Lam JYJ, Barras M, Scott IA, Masood H, Abdel-Hafez A, Falconer N. Machine learning risk prediction models for medication harm in hospitalised adult patients. *Ther Adv Drug Saf* 2026;17. <https://doi.org/10.1177/20420986251409325>.
- [19] Pall R, Gauthier Y, Auer S, Mowaswes W. Predicting drug shortages using pharmacy data and machine learning. *Health Care Manag Sci* 2023;26:395–411. <https://doi.org/10.1007/s10729-022-09627-y>.
- [20] Jiang W, Wang Y, Shi W, Li C, Li M, Tang Y, et al. Artificial Intelligence in Clinical Pharmacy: From Decision Support to Prescription Execution. *J Clin Pharm Ther* 2026;2026:4675298. <https://doi.org/https://doi.org/10.1155/jcpt/4675298>.
- [21] Dong J, Gao Y, Liu Y, Yang X. Development and validation of the nomogram to predict the risk of hospital drug shortages: A prediction model. *PLoS One* 2023;18:e0284528–e0284528. <https://doi.org/10.1371/journal.pone.0284528>.
- [22] Roe HS, Ko DE, Nam SY, Oh JE, Park SM, Lee SH, et al. Drug shortage in South Korea: machine learning-based prediction models and analysis of duration and causal factors. *Front Pharmacol* 2026;Volume 16-2025.
- [23] Nur A, Harbuwono D. Limitations of Artificial Intelligence Research in Predicting Type 2 Diabetes Macrovascular Complications—A Scoping Review. *Diabetes* 2024;73:1987-LB. <https://doi.org/10.2337/db24-1987-LB>.
- [24] Aronson JK. When I use a word . . . Drug stock-outs. *BMJ* 2025;389. <https://doi.org/10.1136/bmj.r1088>.
- [25] Liu Y, Kalaitzi D, Wang M, Papanagnou C. A machine learning approach to inventory stockout prediction. *Journal of Digital Economy* 2025;4:144–55. <https://doi.org/https://doi.org/10.1016/j.jdec.2025.06.002>.
- [26] Pichardo-Lowden AR, Haidet P, Umpierrez GE, Lehman EB, Quigley FT, Wang L, et al. Clinical Decision Support for Glycemic Management Reduces Hospital Length of Stay. *Diabetes Care* 2022;45:2526–34. <https://doi.org/10.2337/dc21-0829>.
- [27] Yang W, Jiang J, Schnellinger EM, Kimmel SE, Guo W. Modified Brier score for evaluating prediction accuracy for binary outcomes. *Stat Methods Med Res* 2022;31:2287–96. <https://doi.org/10.1177/09622802221122391>.
- [28] Cha K, Lee S, Shin J, Lim JY. Beyond Vital Signs: A Machine Learning Model Using Comprehensive Triage-Time Data to Detect Undertriage in Emergency Department Patients. *AI* 2026;7. <https://doi.org/10.3390/ai7060202>.
- [29] Ganie SM, Dutta Pramanik PK, Zhao Z. Enhanced and Interpretable Prediction of Multiple Cancer Types Using a Stacking Ensemble Approach with SHAP Analysis. *Bioengineering* 2025;12. <https://doi.org/10.3390/bioengineering12050472>.