PRISM-TAT: Real-Time Turnaround Time Prediction And Dynamic Sample Routing In Pathology Workflows Using Machine Learning

Mavchi Shilesh Mahesh[1], Gholap Poonam Sunil[2], Prof. Priyanka S. Aher[3]
[1,2,3]*Department of Computer Science and Applications, K. T. H. M. College, Nashik, Maharashtra, India*
doi.org/10.64643/IJIRT211716

***Abstract*—Turnaround time (TAT) in clinical laboratories de-pends on service demand, queue congestion, analyzer compatibility, and urgent-specimen priorities. This paper evaluates PRISM-TAT as a reproducible, entirely synthetic test bed for stage-adaptive prediction and online sample assignment. Extra Trees models estimate individual service times and arrival-to-report TAT; residual quantiles inform a constrained routing heuristic. The service estimator achieved 2.22-min mean absolute error (MAE) and 89.42% empirical coverage of nominal 90% intervals on 5,000 independent synthetic cases. On 3,600 held-out workflow cases, a same-checkpoint ablation found that observed queue-state features reduced post-preanalytical TAT MAE from 24.23 to 6.40 min. In 30 paired nominal replications, PRISM-risk reduced mean TAT versus priority round-robin (37.79 versus 49.64 min), but did not outperform predicted-finish routing (37.39 min). A separately trained stage-TAT-based dispatch variant also showed no reliable advantage over that comparator. After a 28% service-time shift, static service-interval coverage fell to 64.38% in an additional sequential test; completion-ordered rolling residual recalibration attained 89.60%, with wider intervals. Multi-analyzer and risk-weight sensitivity experiments further delimit the method's behavior. The contribution is a traceable evaluation of predictive information, adaptive uncertainty diagnostics, and routing baselines, not demonstrated clinical superiority.**

***Index Terms*—clinical laboratory, turnaround time, machine learning, discrete-event simulation, uncertainty-aware routing**

I. INTRODUCTION

TIMELY laboratory reporting supports clinical decision-making, yet turnaround time (TAT) is not a single universal measurement. It may start at order entry, collection, receipt, or registration and end at analytical completion, verification, or report availability [1], [2], [3]. This distinction matters when comparing interventions: reductions in intra-laboratory waiting cannot automatically be interpreted as equivalent reductions in patient-facing time. Moreover, the reported mean can conceal lengthy queues for individual specimens and delays in urgent work. Multi-site benchmarking and investigations of delay causes have highlighted workload, non-analytical steps, and reporting variability [3], [4], [5].
Machine learning (ML) has previously been used to predict laboratory TAT. Tsai *et al.* trained regression models on 90,543 clinical-chemistry samples, reporting 2.42-min MAE for an Extra Trees estimator in a particular automated laboratory [6]. Park *et al.* combined laboratory information system (LIS)-linked middleware, end-time prediction, monitoring, and urgent requests in an outpatient chemistry service [7]. Digital-shadow monitoring with process improvements has also been evaluated using real laboratory timestamps [8]. These established contributions rule out describing PRISM-TAT simply as the first ML-based laboratory TAT predictor or the first combination of prediction and action.

There is similarly extensive work on operational improvement. Novak *et al.* developed stochastic online centrifuge-batching decisions under uncertain sample transport, evaluated through a discrete-event simulation and university-hospital case study [9]. Biomedical specimen transportation has been optimized using matheuristics [10], while a 2026 IEEE conference contribution proposes a multi-agent reinforcement-learning and digital-twin framework for laboratory workflows [11].
The present work therefore addresses a narrower, falsifiable question: *in a declared synthetic laboratory,*

do stage-updated estimates and calibrated service-time uncertainty support safe dispatch, and is a risk penalty better than transparent deterministic routing alternatives?

Here *real-time* denotes event-triggered updates within a simulation, and *dynamic routing* denotes assignment of newly ready specimens using the currently observed system state; the study does not represent a live laboratory deployment. Our contribution is a reproducible comparison of predictive information and operational dispatch rules. The simulator represents distinct specimen arrivals, pre-analysis, compatible analyzer dispatch, non-preemptive processing, and report release. We evaluate *total* report-TAT prediction at two decision checkpoints on the *same* independent specimens without exposing downstream outcomes to the predictor. We further implement a groupwise residual-calibrated service-time model and a constraint-filtered, risk-aware routing rule with STAT-first queued service. Finally, we report negative comparisons, workload surges, QC-stop scenarios, and a service-distribution shift and perform same-checkpoint feature ablation, causal rolling-recalibration analysis, and three-/five-analyzer replication. None of these assessments uses real clinical data.

This is an engineering simulation and reproducibility study, not a clinical trial, an analysis of patient records, or a deployment claim. All times, deadlines, and specimen classes in the experiments are synthetic assumptions; prior-study performance figures are cited only to characterize previous work and are not comparable to our synthetic scores.

II. RELATED WORK AND RESEARCH POSITIONING

TABLE I POSITION OF THE PRESENT STUDY RELATIVE TO SELECTED PRIOR WORK

Work	Established contribution	Difference in scope here
Tsai *et al.* [6]	ML prediction of chemistry TAT on hospital samples	Synthetic stage updates and dispatch evaluation
Park *et al.* [7]	LIS-linked prediction and workflow support	Paired policy counterfactuals
Novak *et al.* [9]	Stochastic centrifuge batching	Analyzer assignment with compatibility masks
Cai *et al.* [8]	Digital-shadow TAT visibility	Explicit ML calibration stress test
Salaheldin [11]	MARL and laboratory digital twin	Small transparent, non-RL test bed

A. TAT definitions and interventions

Hawkins reviewed laboratory TAT definitions, expectations, measurement, and approaches to improvement [1]. Goswami *et al.* described TAT as a laboratory performance benchmark [2]. An investigation across 738 laboratories documented variation in monitoring and stage-specific indicators [3]. More recent studies have examined pre-analytical bottlenecks, test rejection, and local process redesign, all relevant to choosing feasible intervention points [12], [13], [5]. A systematic review of Lean applications identified laboratory-process changes across heterogeneous clinical settings rather than endorsing a single universal dispatch policy [14]. Lee *et al.* assessed operational strategies to shorten outpatient laboratory delays [15]; Compeau *et al.* specifically studied removal of a clotting stage for urgent biochemical testing [16]. Such studies demonstrate that changes to preparation and reporting may rival changes to analyzer assignment.

B. Prediction, monitoring, and optimization

The predictive features in Tsai *et al.* included priority and arrival-time workload indicators, and their reported accuracy was based on one clinical setting [6].

In Park *et al.*, end-time prediction was embedded in a broader middleware-assisted workflow, establishing that operational action based on expected completion is prior art [7]. Cai *et al.* used real-time digital-shadow visibility and Lean Six Sigma to expose stage-specific delays [8].

Novak *et al.* incorporated uncertain transport into stochastic batching; their decisions concern centrifuge initiation and should not be confused with our non-preemptive analyzer assignment problem [9]. Mazzanti *et al.* studied transport routing across external collection sites, a different spatial optimization problem [10]. Salaheldin's conference framework shows that adaptive laboratory scheduling and digital twins were being considered by 2026 [11].

C. Uncertainty and evidence standards

Prediction intervals should not be treated as guarantees under arbitrary operational changes. Split conformal prediction calibrates residuals on data separate from model fitting; standard coverage arguments depend on an appropriate exchangeability structure. Oliveira et al. analyze split conformal prediction beyond the simplest exchangeable setting and underscore possible coverage penalties under temporal dependence [17]. Gibbs and Candès previously investigated adaptive conformal inference under distribution shift [18]; our rolling-quantile diagnostic is simpler and offers no corresponding online coverage guarantee. Grouped residual quantiles are used as heuristic risk signals; they do not calibrate the aggregate duration of future queue workloads. Reporting recommendations from TRIPOD+AI inform clarity about data sources, evaluation, and limitations, although our synthetic operational model is not a clinical diagnostic/prognostic model and does not claim formal TRIPOD+AI compliance [19].

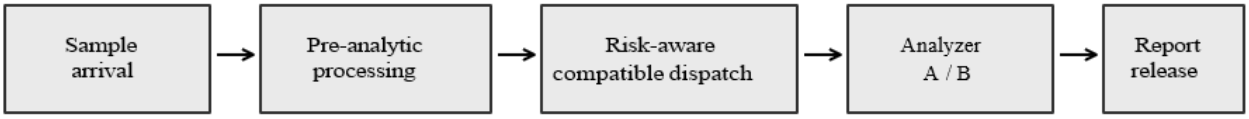


Fig. 1. Simulation sequence. Only the current time, observed specimen descriptors, queue state and published QC status are accessible to a dispatch decision.

III. PROBLEM FORMULATION AND DECISION PROTOCOL

A. Events, endpoints, and feasibility

Each synthetic specimen i has registration/arrival a_i, pre-analytical completion r_i, analysis start b_i, analysis completion c_i, and report availability f_i. Our primary endpoint is

$$T_i = f_i - a_i = (r_i - a_i) + (b_i - r_i) + (c_i - b_i) + (f_i - c_i), \quad (1)$$

measured in minutes. This is *arrival-to-report* laboratory TAT, not collection-to-clinician or total patient TAT. Urgent (STAT) and routine specimens have illustrative 45- and 95-min TAT thresholds, respectively. These thresholds are not presented as universal clinical standards. Primary policy outcomes include the mean of T_i, 95th-percentile T_i, STAT 95th percentile, class-specific deadline breach fractions, and waiting-time indicators.

Let $\mathbf{M}_i$ be the eligible analyzer set defined by a simulated assay-compatibility matrix. At dispatch time t, machine m may be assigned only when $m \in \mathbf{M}_i$ and its QC-stop status $o_m(t) = 0$. If no analyzer is eligible and online, the specimen waits in a holding lobby. A running analysis can finish during a subsequently initiated QC stop, but no *new* analysis may start within that window. All policies use non-preemptive STAT-first dequeue ordering on each analyzer. This is a deliberately limited priority safeguard: it does not provide global preemption, starvation prevention, clinical specimen stability checks, or proof of deadline compliance.

B. Learning observable service demand

For each eligible pair (i, m), x_{im} contains only the specimen's synthetic class, panel count, specimen-matrix flag, and candidate analyzer identity. The unobserved realized service duration is s_{im}. An Extra Trees regressor $\hat{s}(x_{im})$ is trained on disjoint synthetic training specimens. Actual service durations are generated in advance as potential outcomes for comparable, paired-seed policy runs, but are hidden from routing decisions. Thus, the simulator knows s_{im} to advance its event clock while the policy only knows $\hat{s}(x_{im})$ and observed timestamps.

A held-out calibration set provides residuals $e_j = |s_j - \hat{s}(x_j)|$. For each compatible specimen-class/analyzer group g, our finite-sample residual threshold is

$$q_g = e_{\lceil (n_g+1)(1-\alpha) \rceil}, \quad \alpha = 0.10 \quad (2)$$

where the parenthesized subscript denotes the sorted residual and the index is capped at n_g. An interval for one service instance is $[\max(0, \hat{s} - q_g), \hat{s} + q_g]$. Coverage is assessed on an independent test set, both overall and by group. This per-service interval is not automatically a calibrated interval for total laboratory

TAT, a sum of multiple dependent queue contributions, or a sequentially adaptive control system.

C. Dispatch scores and safeguards

At analyzer m, $A_m(t)$ denotes an in-service specimen (if any), with observed start b_{A_m}. The estimated residual service is

$$\hat{R}_m(t) = max\{0,\ \hat{s}_{A_m,m} - (t - b_{A_m})\} \quad (3)$$

Let $Q_m^i(t)$ contain jobs already queued that would precede i under STAT-first non-preemptive service; queued routines are omitted when i is STAT. A projected completion increment is

$$L_{im} = \hat{R}_m(t) + \sum_{j \in Q_m^i(t)} \hat{s}_{jm} + \hat{s}_{im} \quad (4)$$

The predicted-finish baseline chooses the online compatible analyzer with minimum L_{im}. Our PRISM-risk policy instead minimizes

$$J_{im} = L_{im} + \lambda_i U_{im} \quad (5)$$

$$U_{im} = \sqrt{q_{g(i,m)}^2 + \sum_{j \in B_{im}(t)} q_{g(j,m)}^2}$$

where $B_{im}(t) = Q_m^i(t) \cup \{A_m(t)\}$, excluding $A_m(t)$ when the analyzer is idle. The design-fixed coefficients are $\lambda_i = 0.75$ for STAT and 0.32 for routine. Equation (5) is a risk-weighted score, not a probabilistically calibrated completion-time upper bound. No simulated oracle is used for future service time, downstream job completion, or future arrivals.

We compare four policies: priority round-robin (RR) distributes among currently compatible and online analyzers; shortest queue minimizes assigned work-item counts; predicted finish minimizes (4); and PRISM-risk minimizes (5). All use exactly the same priority and compatibility implementation, so apparent improvements cannot be attributed merely to adding STAT priority to one policy. Assigned samples are not subsequently migrated between analyzers; "dynamic routing" here means online assignment at the ready checkpoint, not continuous rerouting of already assigned specimens.

TABLE II SIMULATION, LEARNING, AND EVALUATION PROTOCOL

Parameter	Fixed experimental setting
Training / calibration / test	22,000 / 5,000 / 5,000 independent service examples
Service predictor	Extra Trees, 90 trees, max depth 16, min leaf 3
TAT update models	Extra Trees, 75 trees, max depth 16, min leaf 3 each
Stage train / test	11,520 / 3,600 specimens; disjoint simulation seeds
Dispatch replications	30 seeds per scenario, 420 specimens per seed and policy
Scenarios	Nominal; 1.40× arrival rate; two QC stops; 1.28× service durations
Policy thresholds	STAT 45 min; routine 95 min (illustrative)
Inference constraints	No realized service durations or future event times in policy features
Supplemental studies	Same-checkpoint ablation; paired 30-seed routing; 12-seed risk and warm-up sensitivity; 3/5-analyzer variants; 5,000-case online recalibration

IV. SYNTHETIC DATA AND REPRODUCIBLE EXPERIMENT

A. Declared simulator assumptions

The event engine uses a priority heap and deterministic random streams. Three abstract test classes are used: BASIC (60% of arrivals) can use either analyzer, EXTENDED (25%) requires analyzer A, and SPECIAL (15%) requires analyzer B. These are logical processing classes, not actual chemistry, hematology, or microbiology cross-compatibilities. Arrivals follow exponential interarrival times at 0.134 specimens/min in the nominal environment. STAT prevalence is 15%. Panel count is drawn uniformly from one through six; a binary specimen-matrix indicator has probability 0.30.

Pre-analysis takes max (0.65, Lognormal (ln 3.7, 0.38)) minutes and post-analysis verification takes 1 + Gamma (2, 0.42) minutes in the Python parameterization (shape, scale). For compatible analyzers, service times follow

$$s_{im} = \beta_{k_i}(0.68 + 0.115\, p_i + 0.09\, z_i)\rho_{k_i,m}\, exp(\varepsilon_{im}) \quad (6)$$

with β = (8.5, 12.5, 16.0) minutes, class/analyzer multipliers $\rho_{BASIC,A} = 1.00$, $\rho_{BASIC,B} = 1.17$, $\rho_{EXTENDED,A} = 0.98$, $\rho_{SPECIAL,B} = 1.02$, and $\varepsilon_{im} \sim N(0, \sigma^2)$ with $\sigma = 0.22$ (log-scale standard deviation). Here k_i is the class

of specimen i, p_i its panel count, and z_i its specimen-matrix flag. The same potential service outcomes, arrivals and specimen covariates are presented to all policies in a replication. The figures are plausible test-bed assumptions only; they were not fitted to a laboratory information system, analyzer specification, or patient population.

B. Two-stage report-TAT prediction

To evaluate stage updates rather than only service prediction, separate Extra Trees models predict the same outcome $T_i = f_i - a_i$. The arrival-only model uses observed specimen class, priority, panel count, and matrix flag. Its counterpart is evaluated after pre-analysis and adds *then-observable* elapsed pre-analysis time, selected analyzer, active-job indicator, queued-job count, predicted backlog, service-point estimate, and calibrated service-residual quantile. The later model does not use actual analytical service, analysis start or finish, or report timestamp as a feature. Both are evaluated on the *same 3,600 held-out specimens* generated from seeds 2000–2009; training uses 11,520 specimens from seeds 1000–1031, simulated under a common priority-RR dispatcher. The sample-level endpoint is identical, but the later model enjoys additional real-time information. These reported errors are therefore a demonstration of the value of updates, not a same-information comparison or causal gain due exclusively to model architecture. Stage-prediction runs use 360 specimens per seed; policy-dispatch runs use 420 specimens per seed. These are separate simulated evaluation cohorts. The original risk-based routing score uses the service-demand estimator and its residual quantiles, *not* the stage-level TAT forecasts. A separate exploratory control evaluates candidate dispatch using a fitted stage-TAT score; results from this extension are reported distinctly and cannot be taken as proof of improved integrated control.

C. Paired dispatch experiments and tests

For each of four scenarios, 30 independent fixed seeds (4000–4029) generate 420 specimens. Each seed is replayed across all four policies with identical exogenous arrivals, candidate service outcomes, and simulated class/priority labels. The *surge* increases arrival rate by 40%. The *QC-outage* case blocks new starts on analyzer A for 105 min and on B for 90 min at two distinct intervals relative to that replication's workload horizon. The *distribution-shift* case increases every potential service duration by 28% while keeping the trained model and calibration thresholds fixed. No parameters are tuned after examining these evaluation seeds.

The experimental unit for inference is the simulation *seed*, not the many correlated specimens within a run. For each selected metric u, we report the per-seed paired difference $d_r = u_r(\text{PRISM}) - u_r(\text{comparator})$, its mean, and a per-centile interval from 3,000 bootstrap resamples of the 30 seeds.

Exploratory paired t-test values are retained in the machine-readable artifact but are *unadjusted* for multiple comparisons and are not treated as confirmatory evidence. Deterministic replay, completion, monotone event timestamps, instrument compatibility and no start inside scheduled QC windows are asserted by executed invariant tests.

D. Event causality, trace construction, and reproducibility

The event scheduler maintains a time-ordered heap of specimen-ready, analyzer-completion, QC-stop-on and QC-stop-off events. The simulator advances its internal clock to the next event, updates the machine state, and invokes assignment only for specimens currently ready for analytical processing. A specimen routed to an idle and online analyzer can begin immediately; a specimen routed to a busy analyzer enters that machine's queue. At service completion, the active state is released and the next eligible queued specimen starts using the common STAT-first rule. A specimen blocked from every compatible analyzer is retained in a holding lobby and reconsidered when an analyzer comes online. No simulated specimen is deliberately discarded during an outage.

There are two deliberately separate information layers. The *simulation oracle* stores latent potential service outcomes to schedule completion events, while the *decision layer* sees only fitted service estimates, previous observable start times, queue members, compatibility and QC status. The predicted active-job residual uses the elapsed time since its observed start, rather than the oracle's future completion time. Similarly, a routing score can sum estimated processing times of previously assigned tasks but cannot use future arrival counts. This distinction prevents a common evaluation error in which the policy inadvertently benefits from a simulator-only

lookahead. Because each simulated specimen has a stable identifier, the trace includes exactly one arrival, ready checkpoint, analysis start, analysis completion and report release for every completed specimen.

We use these fields to compute queue waiting $b_i - r_i$ and report TAT $f_i - a_i$. The event-level trace also stores the assignment's estimated queue backlog and service-

time prediction, making later leakage audits possible. The implementation is standard-library event simulation rather than a hospital-validated digital twin; using an event heap does not imply detailed fidelity to instrument firmware or LIS message semantics.

E. What the statistical intervals do and do not measure

The statistical interval around an individual service prediction and the bootstrap interval around a difference between routing policies refer to different targets. In (2), the residual distribution is estimated from a designated calibration population and coverage is evaluated at the observation level. In the paired-policy analysis, the bootstrap samples entire synthetic *replications*, preserving the correlated specimens and time-dependent routing interactions inside each replicate. Neither method independently validates the generative assumptions in (6).

The relative safety implication is also restricted. A routing algorithm can satisfy the two tested hard checks—compatibility and absence of new starts during a QC stop—yet still be operationally inappropriate because of an unmodeled requirement for sample stability, test-specific critical values, reagent quality, staff availability, or verification authority. The current tests should be interpreted as engineering smoke and invariant tests of explicitly represented constraints, not as certification of a clinical laboratory decision-support system.

F. Prespecified supplemental experimental protocols

To test the dependence on queue information, we trained three after-preanalytical Extra Trees models on the same 11,520 RR-generated specimens and evaluated them on the same 3,600 disjoint-seed cases: without queue features, with queue features, and with queue features but without service-prediction features. The first two models observe the same checkpoint and differ only in candidate queue workload indicators. Their comparison is not confounded by later observation time.

A second 30-seed experiment (seeds 5000–5029, 420 specimens per seed) added (i) assignment to the analyzer with the lowest estimated isolated service duration, and (ii) a candidate-specific stage-TAT routing score. The latter uses an 80-iteration histogram gradient boosting model trained on the 11,520 RR traces, including observed elapsed pre-analysis, candidate analyzer type, queue backlog, service estimate, and calibrated service-residual quantile. The dispatch decision predicts from candidate features available at readiness, not oracle service outcomes. This different learned predictor is a *candidate-routing extension*, not the same 75-tree Extra Trees estimator used for the principal prediction result. Scores learned from RR traces may be out of distribution under another dispatcher.

A risk-weight sensitivity experiment multiplied both predeclared coefficients by {0, 0.25, 0.5, 1, 2, 4} across 12 paired seeds per scenario (5300–5311) under nominal and shifted service. A 12-seed initial-condition test (5400–5411) simulated 840 arrivals per run and compared the first 420 with the following 420 specimen-index cohort; this is a burn-in sensitivity test, not proof of stationary operation. For three- and five-analyzer extensions, physical instruments share abstract trained A/B processing types, while distinct synthetic machine-specific potential service durations are drawn in advance. We used arrangements ABA and ABABA, nominal arrivals proportional to the number of machines, 630 or 1050 specimens per seed, and 30 independent seeds (5500–5529). These are hypothetical capacity models, not calibrated instrument configurations.

Finally, to examine shift adaptation, 5,000 sequential independent service examples were evaluated under nominal and 28%-longer service durations (seeds 6100 and 6101). At each specimen-ready time, a groupwise residual quantile was recomputed from at most 120 previously completed specimens (at least 30 required); until that minimum, the original calibration quantile was retained. The current job's observed duration entered the rolling history only after simulated completion. Each instance is a standalone service-time task, deemed completed at its ready time plus realized service duration; the separate analyzer-queue simulator is not coupled to this recalibration test. This prevents label lookahead within the

standalone task stream but does not model hospital-specific result-availability delays. Reported coverage and full widths are descriptive properties of these particular streams; no online conformal validity theorem is claimed.

V. RESULTS

A. Service-demand estimation and reliability of intervals

On 5,000 independent service examples the Extra Trees point predictor achieved MAE 2.223 min and root mean squared error (RMSE) 2.971 min. Groupwise residual calibration targeted 90% interval coverage, obtaining 89.42% marginal empirical coverage and 9.00 min mean full interval width. Coverage differed by eligible class–machine group: 89.60% for BASIC/A, 89.77% for BASIC/B, 90.19% for EXTENDED/A and 87.08%

for SPECIAL/B. Thus, even without deliberate shift, a subgroup can experience poorer coverage than a nominal percentage suggests. These are intervals for individual service durations, *not* independently calibrated TAT forecasts.

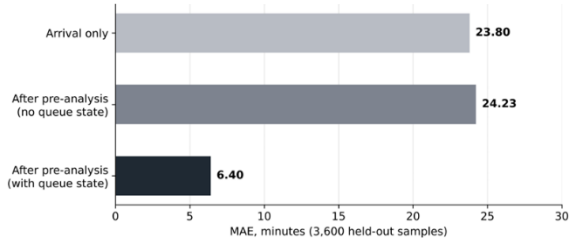


Fig. 2. Arrival-only and after-preanalysis TAT forecasts on 3,600 held-out synthetic specimens. The two after-preanalysis bars compare models with and without queue-state features at the same checkpoint.

Under the 28% service-time increase imposed on a distinct 5,000-case stress set, the unchanged service model's MAE rose to 3.953 min and nominal-90% interval coverage dropped to 64.26%. This failure is operationally important: a scheduler that interprets stale residual bounds as trustworthy risk limits may become overconfident after reagent, workload, instrument or case-mix changes. In our synthetic setting the chosen risk penalty remains a heuristic rather than a validated clinical reliability mechanism. Recalibration or out-of-distribution monitoring would be a necessary future safeguard.

B. Stage-updated arrival-to-report predictions

As shown in Fig. 2, using information available at pre-analytical completion yielded an MAE of 6.401 min for total arrival-to-report TAT, compared with 23.800 min for the arrival-only predictor evaluated on identical 3,600 held-out cases. RMSE likewise fell from 32.021 to 9.813 min. The apparent 73.1% MAE reduction is *conditional on waiting for a later prediction checkpoint*; it does not mean that a better initial forecast was achieved, nor does it establish that an earlier clinical intervention could have used those later data. It instead quantifies the predictive value of updated queue and elapsed-stage observations under a fixed simulated dispatch policy.

C. Same-checkpoint queue-information ablation

At the *same* after-preanalysis checkpoint and on identical 3,600 held-out specimens, the Extra Trees predictor without queue-workload indicators obtained 24.229-min MAE (RMSE 32.738 min). Adding candidate backlog, queued-job count, and active-analyzer state reduced MAE to 6.401 min (RMSE 9.813 min). A queue-aware variant that omitted the service estimate and residual quantile obtained 6.438-min MAE. Fig. 2 contrasts the arrival-only result and these same-checkpoint models. This ablation supports the value of observed queue state in the specified workload generator, rather than attributing the improvement simply to observing at a later time.

TABLE III NOMINAL-SCENARIO RESULTS: MEAN OF 30 REPLICATIONS

Policy	Mean TAT (min)	p95 TAT (min)	STAT p95 (min)	STAT miss (%)	Routine miss (%)
Priority RR	49.64	126.94	39.71	2.61	13.43
Shortest queue	37.92	83.30	39.14	2.25	4.83
Predicted finish	37.39	83.46	37.27	1.66	4.72
PRISM-risk	37.79	83.97	37.39	1.61	5.16

Results are simulation-only. p95 is computed per seed and then averaged, not pooled across seeds.

TABLE IV NOMINAL PAIRED DIFFERENCES, PRISM MINUS COMPARATOR

Comparator	Outcome	Mean Δ	95% bootstrap CI
Priority RR	Mean TAT (min)	−11.85	[−15.14, −9.35]
Priority RR	p95 TAT (min)	−42.97	[−57.98, −31.38]
Shortest queue	Mean TAT (min)	−0.13	[−0.65, 0.38]
Shortest queue	p95 TAT (min)	+0.68	[−1.36, 2.60]
Predicted finish	Mean TAT (min)	+0.40	[−0.01, 0.93]
Predicted finish	p95 TAT (min)	+0.51	[−0.68, 1.84]

D. Nominal dispatch performance

Table III shows means of the per-replication outcomes across 30 nominal replications (12,600 specimens for each policy). Priority RR yielded 49.64-min mean TAT and 126.94-min mean per-replication p95 TAT. PRISM-risk produced 37.79 and 83.97 min, improvements of 11.85 and 42.97 min against RR (23.9% and 33.9%, respectively). These percentage changes are computed from averages across replications and are not hospital-effect estimates. The average STAT deadline-breach fraction decreased from 2.61% to 1.61%, while routine breaches decreased from 13.43% to 5.16%.

The strong comparator changes the conclusion. Against *predicted finish*, PRISM-risk was 0.403 min slower in mean TAT and 0.511 min slower in p95 TAT in the nominal scenario; both confidence intervals include zero (Table IV). The shortest-queue approach also obtained competitive tail TAT. Consequently, the evidence does *not* justify claiming that the particular uncertainty penalty in (5) improves throughput beyond good deterministic policies. The experiment does, however, show that a predictor and conservative resource-selection interface can be implemented without violating the tested compatibility or QC-stop invariants.

E. The role of a strong baseline

The nominal policy ranking is informative because all algorithms retain the same two safety gates and priority queue. Had the comparison stopped at priority round-robin, a reduction from 126.94 to 83.97 min in per-seed p95 TAT might misleadingly be attributed to uncertainty handling. The predicted-finish policy instead uses the *same learned point estimate* and the same observed backlog, but no residual penalty. It achieved 83.46 min, a slightly better observed value. A shortest-queue rule that requires no fitted model achieved 83.30 min. Thus, the data favor a limited claim: moving beyond alternating assignment is beneficial in this environment, whereas an incremental advantage for uncertainty-aware scoring was not established.

TABLE V STRESS RESULTS: PRISM-RISK AND PREDICTED-FINISH

Scenario	Policy	Mean TAT (min)	p95 TAT (min)	Routine miss (%)
QC outage	Pred.-finish	46.30	110.29	12.50
QC outage	PRISM	46.70	111.48	12.85
Surge	Pred.-finish	240.79	465.88	85.52
Surge	PRISM	239.91	463.80	85.28
Service shift	Pred.-finish	194.62	381.14	78.53
Service shift	PRISM	194.11	381.07	78.70

A possible secondary benefit of residual bounds is communicating uncertainty to laboratory personnel, but this paper did not measure human responses, alert fatigue, or decision quality. We do not substitute that untested possibility for the missing throughput gain. Similarly, a nominally lower mean at a shifted workload is not proof of robustness when the paired intervals, calibration failure and severe deadline breaches are considered together. A policy should only be described as better for a named outcome, test bed and comparator when the evidence supports that exact wording.

F. Stress scenarios and priority outcomes

Fig. 3 and Table V report stress outcomes without replacing individual-scenario results by a selectively

favorable overall average. QC stops raised all policies' waiting times and TAT tails. Under QC outages, PRISM-risk gave mean TAT 46.70 min and p95 111.48 min; predicted finish achieved 46.30 and 110.29 min, respectively. A 40% arrival-rate surge made nominal service capacity insufficient for many specimens; PRISM-risk mean TAT rose to 239.91 min, and the average routine deadline-breach proportion reached 85.28%. Such severe degradation cannot be corrected by choosing a slightly better dispatch score when incoming work persistently exceeds available capacity.

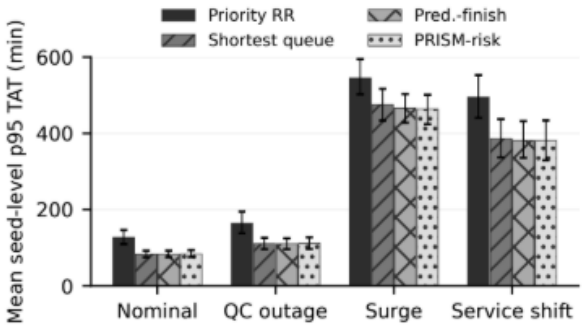


Fig. 3. Across-scenario mean of the per-seed 95th-percentile arrival-to-report TAT (30 seeds per scenario). The common vertical scale reveals much higher tail delays under overload; relative policy ranking is not constant.

In the 28% service-duration shift, PRISM-risk mean TAT rose to 194.11 min with p95 381.07 min. The predicted-finish comparator obtained 194.62 and 381.14 min. These differences are small compared with scenario-level congestion and do not establish robustness. STAT p95 remained much lower than routine p95 across scenarios mainly because all policies prioritize STAT specimens in each analyzer queue; the burden on routine service is visible in routine deadline breaches. This is a tradeoff to report, not a safe operating recommendation for a real service.

G. Capacity, deadline stress, and illustrative service targets

At the chosen nominal arrival rate, the two-machine system has enough capacity on average to process most routine work with finite waiting, but dedicated EXTENDED and SPECIAL workloads restrict how much the shared BASIC class can be rebalanced. A 40% increase in arrivals or a 28% increase in service durations can move effective utilization near or beyond sustainable capacity. In those overload settings queueing delay compounds with time: a routing score can redistribute eligible BASIC samples but cannot create analytical capacity for EXTENDED or SPECIAL specimens restricted to one machine. The sharp p95 increases in Fig. 3 are therefore consistent with capacity-constrained queueing, not necessarily with a faulty predictor alone.

The nominal thresholds of 45 min (STAT) and 95 min (routine) were selected to make deadline misses observable and to test a tradeoff; they are not derived from a specific assay, hospital protocol, or professional society requirement. For example, a fixed threshold for an urgent blood gas would not automatically be appropriate for a longer specialty assay. We report class-specific misses rather than a single aggregate miss rate because the STAT-first discipline can protect selected urgent jobs while routine jobs experience substantial aging under overload. No clinical benefit should be inferred from satisfying synthetic thresholds.

H. What is and is not an ablation study

The removal of the uncertainty penalty while retaining the point predictor, compatible dispatch, and priority queue is an explicit *policy ablation*: the predicted-finish comparator corresponds to the point-score term L_{im} alone. This is the strongest direct test of the contribution of λ_i and calibrated residual magnitude. In contrast, comparing an arrival-only TAT predictor with a later stage-updated predictor is *not* a pure feature or architecture ablation because the second estimator is allowed to observe information unavailable at arrival. We separate these claims to avoid conflating the value of new observations with algorithmic superiority.

Additional studies should use a truly same-checkpoint comparison with and without each candidate feature, a strictly held-out site or time block, ablations of the two routing coefficients, and capacity scenarios chosen before evaluation. Such studies might identify settings were uncertainty changes dispatch materially. Our current package does not include that evidence; it is not filled in by assumption or extrapolation.

I. Executed correctness checks

We executed deterministic reruns under every routing policy and asserted full completion of 100 generated

specimens for each, monotone timestamps $a_i \le r_i \le b_i \le c_i \le f_i$, and compatibility of each analyzer choice. A separate 100-specimen QC-outage test asserted no analysis begins inside the specified machine-offline windows. A sixth check confirmed generation of both STAT and routine classes. A seventh zero-residual ablation confirmed that PRISM-risk exactly reproduces predicted-finish dispatch when all uncertainty penalties are set to zero. All seven checks passed. These are software invariants for the *specified* model, not formal proofs of correctness for all possible workflows or a validation of medical-device safety. Full per-seed outcomes, test messages, simulation code, and an illustrative event-level dataset accompany the paper as research artifacts.

J. Supplemental routing benchmarks and model coupling

With 30 new paired seeds, predicted-finish dispatch obtained 38.037-min mean TAT and 85.195-min mean per-seed p95, versus 38.131 and 85.810 min for PRISM-risk. Candidate stage-TAT-based routing obtained 38.070 and 85.279 min. Its paired mean-TAT difference relative to predicted finish was +0.032 min (95% seed-bootstrap CI [−0.234, 0.316]); the p95 difference was +0.084 min ([−1.204, 1.482]). Thus, *actually using* a learned after-preanalytical TAT score for assignment did not demonstrate a dispatch gain. A fastest-isolated-service rule generated very long routine backlogs (369.64-min mean TAT) because it ignored queue congestion. A trial deadline-tardiness scoring formula reduced algebraically to predicted-finish ranking for each fixed-deadline specimen, so it is not reported as an independent scheduling improvement. The model-coupling comparison remains exploratory rather than a validated end-to-end feedback optimizer.

K. Rolling calibration under synthetic service shift

Table VI and Fig. 4 compare unchanged residual bounds with causally updated rolling quantiles on disjoint additional 5,000-case streams. Under nominal service, fixed versus rolling coverage was 89.66% versus 90.12%, and average full interval width rose from 9.03 to 9.38 min. Under the +28% service-duration shift, static coverage was 64.38% versus 89.60% for rolling quantiles, while mean interval width grew from 9.00 to 16.83 min. Over the final 4,000 shifted cases, rolling coverage was 90.18%, versus 64.43% using the unchanged bounds. The tradeoff is materially wider intervals and the need to observe completed service outcomes. The experiment does not establish adaptation to abrupt nonstationary changes, sparse subgroups, delayed labels in real laboratories, or formally calibrated per-job service guarantees. Related adaptive methods already exist [18].

TABLE VI ADDITIONAL SERVICE-INTERVAL COVERAGE AUDIT (N = 5,000 EACH)

Test population	Residual bound	Coverage (%)	Mean width (min)
Nominal	Fixed	89.66	9.03
Nominal	Rolling	90.12	9.38
+28% service	Fixed	64.38	9.00
+28% service	Rolling	89.60	16.83

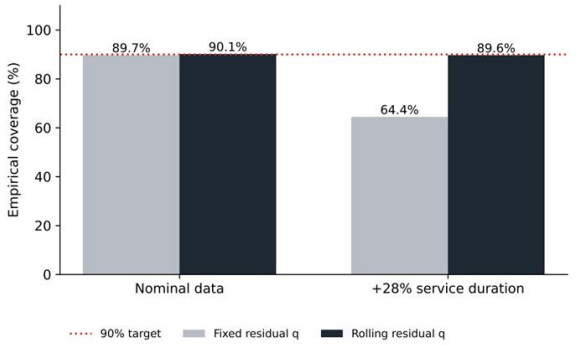


Fig. 4. Descriptive empirical coverage of nominal-90% service-time intervals; rolling quantiles only use residuals observed after completion. Interval widths differ substantially (Table VI).

L. Risk sensitivity, burn-in, and analyzer replication

The 12-seed coefficient sweep found no stable optimum for the risk penalty: nominal mean TAT ranged from 35.31 to 35.47 min across multipliers, while shifted-service mean TAT ranged from 164.58 to 167.42 min. These are small, exploratory samples and should not be used to select a tuned coefficient on the test set. In the 12-seed 840-specimen check, the first 420 specimens had higher mean TAT than the later cohort for predicted finish (45.18 versus 35.91 min) and PRISM-risk (45.07 versus 35.88 min). These differences show *initial-condition sensitivity* but do not prove steady-state convergence; the arrival-

count cohort boundary is not a clock-time warm-up estimator.

In scaled analyzer configurations, learned service predictions were reused for cloned machine types, not retrained on real instruments. Table VII shows a slightly better PRISM-risk average in both configurations, but paired mean-TAT bootstrap intervals included zero: 3 analyzers, −0.244 min, CI [−0.493, 0.015]; 5 analyzers, −0.023 min, CI [−0.180, 0.141].

The data do not support a general advantage across analyzer counts. All scale-out test trajectories completed, obeyed abstract type compatibility, had monotone timestamps, and passed deterministic replay on a separate 420-specimen validation sample.

TABLE VII EXPLORATORY 3-/5-ANALYZER SCALE-OUT (30 SEEDS PER DESIGN)

Configuration	Policy	Mean TAT (min)	p95 TAT (min)
ABA (3)	Predicted finish	29.15	63.02
ABA (3)	PRISM-risk	28.91	62.19
ABABA (5)	Predicted finish	24.53	46.36
ABABA (5)	PRISM-risk	24.51	46.32

VI. DISCUSSION

A. What the findings establish

The central positive result is methodological: at the same checkpoint, including observed queue state substantially reduced report-TAT prediction error, and a learned service-time model can be used inside a hard compatibility filter and priority-aware dispatcher. The study also illustrates how to compare policies without assigning a different arrival workload to each. Paired random seeds reduce variability attributable to the generated demand stream. A fully specified event trace permits checking whether apparent speedups arise from altered workload, time origin, priority logic, or inadvertent information leakage.

The equally important negative result is that the tested risk penalty does not convincingly improve on estimated-finish dispatch. We retain this result rather than weaken the baseline or tune λi against the holdout. A plausible explanation is that the learned service-time uncertainty is modest under nominal demand, while queue-length and compatibility constraints dominate assignment decisions; adding a residual penalty may redirect work to slower or busier analyzers. This explanation remains a hypothesis. A future study should independently vary uncertainty magnitude, loss asymmetry, confidence levels, dispatch frequency, and capacity, with validation scenarios separated from final reporting seeds.

B. Separation between risk scores and guarantees

Groupwise calibration provided reasonable marginal service-time coverage under nominal simulated data, but the weakest subgroup fell below 90%, and the deliberate shift reduced coverage to 64.26%. The root-sum-of-squares uncertainty term in (5) assumes no exact conditional dependence structure and is not a conformal interval for a queue's total completion time. It must not be marketed as an assurance that a STAT result will be delivered before a clinical deadline. Conformal exchangeability requirements and shift sensitivity are discussed in the statistical literature [17]. Pre-deployment validation would require a defensible site-specific residual process, temporal recalibration, subgroup auditing, clinical quality controls, and governance of unexpected failures.

C. Boundary of the model and threats to validity

Construct validity:

BASIC, EXTENDED, and SPECIAL are abstract workload classes; no real analyzer interoperability is implied. Pre-analysis and reporting are simplified independent delays; real laboratories can include batching, reruns, aliquoting, sample rejection, rework, inter-lab transfers, analyzer modules, reagent stockouts and multiple verification roles. Published clinical settings and test compositions differ markedly [20], [21], [22]. Therefore, the simulation's numerical TAT values do not estimate those environments.

Internal validity:

Although policies share arrival/service potential outcomes within each seed, scheduling changes queue states and hence future decisions. That is the intended intervention mechanism rather than a confound. The arrival-stage model lacks dynamic queue measurements, while the later model has them; the reported accuracy gap is an information-timing

comparison. The two-stage predictors were fitted on RR-generated events; their errors should not be assumed to transport unchanged to another dispatcher. The exploratory candidate-stage routing extension likewise used RR-derived training traces and introduces policy-induced covariate shift. Priority is non-preemptive and analyzer-local; no cross-queue global fairness guarantee is implemented. The QC model prevents starts but permits an in-progress analysis to finish; many real equipment faults would not behave this way.

Statistical validity:

Thirty seeds per scenario are independent Monte Carlo replicates, but comparisons within a seed are dependent by design. Bootstrap intervals summarize this finite synthetic experiment. No adjustment was performed for many exploratory metrics and scenarios. Dependence within a run, rare extreme delays, and arbitrary workload distributions limit inferential reach. Parameter choices were fixed for the study and not identified from measurements. Primary runs began with empty queues; a separate 420-arrival burn-in sensitivity comparison was performed, without formally establishing stationarity. A finite risk-coefficient sensitivity sweeps and additional routing heuristics were examined, but optimized deadline-based or global scheduling remains outside scope.

External and clinical validity:

There is no real patient data, prospective pilot, institutional approval, live LIS connection, calibrated regulatory risk analysis, or clinician evaluation. The findings do not demonstrate reductions in hospital length of stay, patient harm, or staff workload. Reporting guidelines emphasize transparent disclosure of modeling data and validation scope
[19]; no diagnostic or clinical-decision recommendation follows from these results.

D. Replication, data provenance, and future evaluation

The Python implementation specifies the simulation generator, random seeds, model hyperparameters, decision rules, and invariant checks. The reproducibility archive includes 480 primary scenario–policy–seed summaries (4 policies ×4 scenarios × 30 seeds), paired-effect estimates, calibration summaries, an illustrative event-level synthetic trace, and separate supplement tables and scripts for each additional experiment. It is distributed as supplementary research material; no permanent public repository identifier has yet been assigned. The cited studies were selected for topical relevance and should not be mistaken for a systematic literature review.

The next scientifically meaningful validation would be (i) collaboration with an authorized laboratory and ethically approved or appropriately de-identified event logs; (ii) reconstruction of realistic compatibility, QC, transport, staff and rerun logic; (iii) temporal/shift-aware calibration of stage-level remaining-TAT distributions; (iv) robust tests under atypical arrivals and unavailable machines; and (v) preregistered, site-specific comparisons with strong rules already used by laboratory staff.

Any clinical integration must remain advisory until separately validated under local quality management and safety policies.

VII. CONCLUSION

This study evaluated PRISM-TAT as a transparent synthetic test bed for stage-adaptive report-TAT prediction and constraint-filtered online sample dispatch. An independent service-time test yielded 2.22-min MAE with 89.42% marginal nominal-90% interval coverage, and updated total-TAT forecasts were more accurate than arrival-only forecasts once pre-analytical and queue information became available.

The risk-aware dispatcher substantially improved nominal mean and tail TAT versus a priority-RR baseline, but offered no demonstrated advantage over a strong predicted-finish baseline. Severe workload and distribution shifts exposed major capacity and calibration limits. Same-checkpoint ablation identified queue state as a major predictive signal; a completion-ordered rolling residual quantile recovered approximately nominal empirical coverage under service-time shift at a cost in interval width. A separately trained stage-TAT routing score and 3-/5-analyzer tests did not establish general incremental TAT improvements. The contribution is a reproducible and deliberately bounded benchmark of these engineering choices. Real data, time-blocked

validation, and operational governance remain prerequisites for further claims.

RESEARCH ETHICS AND TRANSPARENCY STATEMENT

This computational study used entirely generated data, without patient records or human subjects. Experiments were executed on 23 November 2026. The simulated policies have not been assessed for clinical use.

REFERENCES

[1] R. C. Hawkins, “Laboratory turnaround time,” *Clinical Biochemist Reviews*, vol. 28, no. 4, pp. 179–194, 2007.

[2] B. Goswami, B. Singh, R. Chawla, V. K. Gupta, and V. Mallika, “Turn around time (TAT) as a benchmark of laboratory performance,” *Indian Journal of Clinical Biochemistry*, vol. 25, no. 4, pp. 376–379, 2010, doi: 10.1007/s12291-010-0056-4.

[3] X. Zhang *et al*., “National survey on turnaround time of clinical biochemistry tests in 738 laboratories in China,” *Journal of Clinical Laboratory Analysis*, vol. 32, Art. no. e22251, 2018, doi: 10.1002/jcla.22251.

[4] K. P. Chauhan, A. P. Trivedi, D. Patel, B. Gami, and N. Haridas, “Monitoring and root cause analysis of clinical biochemistry turnaround time at an academic hospital,” *Indian Journal of Clinical Biochemistry*, vol. 29, pp. 505–509, 2014, doi: 10.1007/s12291-013-0397-x.

[5] Naveen K. G., S. Ahmad, P. Kumar, M. Mahto, and S. Bharti, “Turnaround time of laboratory investigations and factors responsible for delays in a tertiary care teaching hospital in Eastern India: A cross-sectional study,” *Cureus*, 2026, doi: 10.7759/cureus.100703.

[6] E. R. Tsai, D. Demirtas, N. Hoogendijk, A. N. Tintu, and R. J. Boucherie, “Turnaround time prediction for clinical chemistry samples using machine learning,” *Clinical Chemistry and Laboratory Medicine*, vol. 60, no. 12, pp. 1902–1910, 2022, doi: 10.1515/cclm-2022-0668.

[7] M. Park *et al*., “Quality improvement of outpatient clinical chemistry tests through a novel middleware-laboratory information system solution,” *Clinical Biochemistry*, vol. 113, pp. 21–28, 2023, doi: 10.1016/j.clinbiochem.2022.12.017.

[8] X. Cai *et al*., “Optimizing clinical laboratory efficiency through digital shadow and lean six sigma integration: A real-time monitoring approach to reduce intra-laboratory turnaround time,” *Digital Health*, vol. 11, 2025, doi: 10.1177/20552076251375939.

[9] A. Novak, A. Gnatowski, and P. Sucha, “Urgent samples in clinical laboratories: Stochastic batching to minimize patient turnaround time,” *Health Care Management Science*, vol. 29, Art. no. 16, 2026, doi: 10.1007/s10729-026-09756-8.

[10] C. Mazzanti, A. M. Anaya-Arenas, V. Bélanger, E. Lanzarone, and A. Ruiz, “Iterative matheuristic for the biomedical sample transportation problem,” *Computers & Industrial Engineering*, vol. 192, Art. no. 110196, 2024, doi: 10.1016/j.cie.2024.110196.

[11] A. M. Salaheldin, “An autonomous clinical laboratory optimization framework using multi-agent reinforcement learning and digital twin modeling,” in *Proc. Int. Conf. AI Innovations and Industry (ICAIII)*, 2026, pp. 1–5, doi: 10.1109/ICAIII69475.2026.11521511.

[12] X. Cheng, H. Yu, L. Zhang, B. Zhang, and Q. Wang, “Evaluation of pre-analytical specimen rejection using Six Sigma metrics: A retrospective single-center study,” *PLOS ONE*, vol. 20, no. 6, Art. no. e0324840, 2025, doi: 10.1371/journal.pone.0324840.

[13] R. R. Guru *et al*., “Streamlining molecular and serology laboratory operations: A Lean Six Sigma approach,” *Cureus*, 2025, doi: 10.7759/cureus.80038.

[14] N. Cherie *et al*., “Improving laboratory turnaround times in clinical settings: A systematic review of the impact of lean methodology application,” *PLOS ONE*, vol. 19, no. 10, Art. no. e0312033, 2024, doi: 10.1371/journal.pone.0312033.

[15] S. Lee, S. Yoon, W. Lee, S. Chun, and W.-K. Min, “Strategies to shorten turnaround time in outpatient laboratory,” *Journal of Clinical Laboratory Analysis*, vol. 36, Art. no. e24665, 2022, doi: 10.1002/jcla.24665.

[16] S. Compeau *et al*., “Does elimination of a laboratory sample clotting stage requirement reduce overall turnaround times for emergency

department stat biochemical testing?” *Cureus*, 2016, doi: 10.7759/cureus.819.

[17] R. I. Oliveira, P. Orenstein, T. Ramos, and J. V. Romano, “Split conformal prediction and non-exchangeable data,” *Journal of Machine Learning Research*, vol. 25, no. 225, pp. 1–38, 2024.

[18] I. Gibbs and E. J. Candès, “Adaptive conformal inference under distribution shift,” in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 1660–1672.

[19] G. S. Collins *et al*., “TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods,” *BMJ*, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378.

[20] W. Liu *et al*., “Blood culture quality and turnaround time of clinical microbiology laboratories in Chinese teaching hospitals: A smulticenter study,” *Journal of Clinical Laboratory Analysis*, vol. 38, Art. no. e25008, 2024, doi: 10.1002/jcla.25008.

[21] M. B. Shiferaw and G. Yismaw, “Magnitude of delayed turnaround time of laboratory results in Amhara Public Health Institute, Bahir Dar, Ethiopia,” *BMC Health Services Research*, vol. 19, Art. no. 240, 2019, doi: 10.1186/s12913-019-4077-2.

[22] P. Prasad, R. Kumar, S. Kumar, and P. Sinha, “Monitoring and root cause analysis of clinical biochemistry turnaround time at a tertiary care institute,” *Cureus*, 2023, doi: 10.7759/cureus.39821.