INTERNATIONAL JOURNAL OF
MEDICAL BIOCHEMISTRY

DOI: 10.14744/ijmb.2026.79106
Int J Med Biochem 2026;9(4):251–263

Research Article

# Specimen type as the dominant predictor of pre-analytical rejection in the emergency department: An explainable machine-learning (SHAP) analysis

**Yunus Goren**[1], **Hakan Ozerol**[2]

[1]Department of Medical Biochemistry, Gaziantep City Hospital, Gaziantep, Türkiye
[2]Department of Emergency Medicine, Gaziantep City Hospital, Gaziantep, Türkiye

## Abstract

**Objectives:** Preanalytical specimen rejection is a leading source of laboratory errors in emergency departments, but most evidence comes from small studies lacking a true denominator. We examined factors associated with specimen rejection in a complete 34-month cohort and assessed the extent to which apparent predictability was attributable to specimen type.

**Methods:** All tubes collected in the emergency department of a tertiary hospital between October 2023 and August 2026 were included. The outcome was tube rejection. Predictors were restricted to information available at the time of request: patient characteristics, timing (including hours since shift handover), workload, collection context, specimen type, and panel size. Four algorithms were compared using temporal validation on tubes collected between January and August 2026 and evaluated for discrimination, calibration, net benefit, subgroup performance, and Shapley additive explanations. Adjusted odds ratios (ORs) were calculated using standard errors clustered by presentation.

**Results:** Of 1,190,944 tubes collected during 404,247 presentations, 37,633 (3.16%) were rejected, with hemolysis accounting for 67.6% of rejections. LightGBM achieved an out-of-time area under the receiver operating characteristic curve (AUROC) of 0.780 (95% confidence interval [CI], 0.775–0.785), with a calibration slope of 0.983. Biochemistry tubes had an adjusted OR of 10.48 (95% CI, 10.05–10.93). Specimen type alone achieved an AUROC of 0.709, whereas all patient characteristics, timing, and workload variables combined achieved an AUROC of 0.601. The odds of rejection were highest during the final 4 hours of the 24-hour shift (OR, 1.15; 95% CI, 1.09–1.21). The lower unadjusted odds observed in pediatric and obstetric units were no longer evident after adjustment.

**Conclusion:** Specimen type was the strongest predictor of rejection. Models developed using laboratory information system data reflect differences in specimen mix and should be benchmarked against specimen type.

**Keywords:** Hemolysis, hospital emergency service, machine learning, pre-analytical phase, quality indicators, specimen handling

**How to cite this article:** Goren Y, Ozerol H. Specimen type as the dominant predictor of pre-analytical rejection in the emergency department: An explainable machine-learning (SHAP) analysis. Int J Med Biochem 2026;9(4):251–263.

Preanalytical errors account for most errors in the total testing process, and specimen rejection is their most visible manifestation [1–3]. Emergency departments are the highest-risk settings for specimen collection: sampling occurs under time pressure, frequently through intravenous catheters, and is performed by staff whose primary responsibility is not phlebotomy. Reported rejection and hemolysis rates range from less than 1% of specimens in laboratory-wide series to more than 10% in emergency departments, with hemolysis being the leading cause in nearly every published series [4–7].

Two limitations recur in this literature. First, most studies count rejected specimens without a reliable denominator that

**Address for correspondence:** Yunus Goren, MD. Department of Medical Biochemistry, Gaziantep City Hospital, Gaziantep, Türkiye
**Phone:** +90 506 420 91 28 **E-mail:** ysgoren@gmail.com **ORCID:** 0000-0003-0344-2456

**Submitted:** September 23, 2026 **Revised:** October 07, 2026 **Accepted:** October 09, 2026 **Available Online:** October 15, 2026

**OPEN ACCESS** This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

includes accepted specimens, so the reported rate depends on how the denominator was reconstructed. Second, cohorts are typically small, ranging from thousands to a few tens of thousands of specimens, leaving them underpowered to test the operational hypotheses they most frequently advance: that rejection rates increase at night, with hours on duty, and during periods of high departmental workload. When such associations have been reported, they have rarely been adjusted for the composition of the specimens collected, even though hemolysis affects serum biochemistry tubes far more frequently than hematology or blood gas specimens [4].

Machine learning has recently been applied to preanalytical error detection [8–12], with studies often reporting very high discriminatory performance. However, the variables responsible for this performance are seldom explicitly identified. If a model includes variables encoding specimen type, much of its apparent predictive ability may reflect the near-deterministic relationship between tube type and hemolysis rather than any patient- or process-level insights.

We assembled the complete emergency department laboratory records of a tertiary hospital over 34 consecutive months, including every accepted and rejected tube, and addressed three questions. Which factors are associated with specimen rejection after adjusting for specimen type? Are circadian, shift, and workload factors associated with rejection, and is the cohort sufficiently large to exclude associations of practical relevance? How much of an explainable machine learning model's discriminatory performance reflects specimen mix rather than modifiable process factors?

## Materials and Methods

### Setting and design

This retrospective cohort study included all specimen tubes collected in the emergency department of a tertiary care hospital between 6 October 2023 and 10 August 2026. The department comprises adult, obstetric, and four pediatric units. Patients are allocated to red, yellow, or green triage areas, and nursing staff work 24-hour shifts with a single handover at 08:00. Specimens are sent to a co-located core laboratory and screened upon receipt.

The study is reported in accordance with TRIPOD+AI [13]. This study was approved by the Gaziantep City Hospital Institutional Ethics Committee (Date: 22 July 2026, Decision No: 624/2026), which waived the requirement for informed consent for the secondary use of anonymized routine data. Approval details are provided on the title page.

The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki.

### Data source and dataset construction

Records were extracted from the laboratory information system (LIS) at the individual test-request level, comprising 10,283,214 rows. Each row contained an anonymized presentation identifier, collection barcode, patient age and sex, emergency department unit, triage area, admission status, requested test, request timestamp, and, where applicable, rejection reason and laboratory timestamps.

Because rejection applies to a tube rather than an individual test, rows were aggregated by barcode, yielding 1,190,944 tubes. For each tube, we recorded the set of distinct tests requested, from which specimen type was derived and classified into one of six mutually exclusive categories: biochemistry (serum), complete blood count (CBC), blood gas, urine, coagulation, and other. Panel size was defined as the number of distinct tests requested for each tube.

### Outcome

The primary outcome was tube rejection by the laboratory, defined as the presence of any rejection reason in any constituent test row. This definition was chosen because it corresponds to the operational event relevant to the emergency department, namely specimen loss and the need for repeat collection, as well as the rejection-rate quality indicator reported to laboratory management and harmonized internationally [14]. Rejection reasons were retained verbatim and grouped post hoc into specimen-quality reasons (hemolysis, clotting, insufficient volume, lipemia, icterus, inappropriate container or fill level, unseparated serum, temperature deviation, and breakage) and administrative reasons (absent blood-draw record, barcode and data-entry errors, duplicate entry, requested repeat sampling, and tests not performed in the emergency laboratory). Because these two groups reflect different processes, two secondary outcomes were analyzed: specimen-quality rejection, for which tubes rejected solely for administrative reasons were treated as non-events, and hemolysis-specific rejection.

### Predictors

Predictors were restricted to information available at the time the request was entered. Patient variables included age, sex, emergency department unit, triage area, and admission status. Age was exported from the LIS in completed years; therefore, a value of 0 denoted an infant younger than 1 year and was analyzed as a valid value rather than as a missing or default entry. Timing variables included hour of day (sine and cosine terms), hours since the 08:00 handover, day of week, weekend, season, and calendar time. Workload variables included the number of tubes drawn in the same unit during the same hour and across the department during the same hour and day. Collection-context variables included the number of tubes drawn from the patient during the same minute and the interval since the first specimen collected during the presentation. Specimen variables included specimen type and panel size.

Triage area was missing for 1,278 tubes (0.107%); these tubes were retained in all analyses, and no imputation was performed. All other predictors were complete.

## Screening for information leakage

Three candidate variables that encoded the outcome were excluded before model fitting.

The number of result rows per tube was excluded because accepted tubes accumulate repeat and reflex measurement rows, whereas rejected tubes do not (among CBC tubes, 97.4% of accepted specimens but only 4.2% of rejected specimens contained more than one row per test). This variable was replaced by the count of distinct requested tests.

Presentation-level totals (tubes, tests, and duration per presentation) were excluded because rejection triggers a repeat draw and therefore increases these totals, making them consequences of the outcome.

The rank of a tube within a presentation was excluded. For the 85% of tubes recorded in the same minute as another tube from the same patient, this rank was derived from the row order in the data export rather than from an observed sequence, and it encoded specimen type: 49% of tubes ranked first were biochemistry tubes, compared with 3.5% of those ranked fifth. Retaining this variable increased the apparent area under the receiver operating characteristic curve (AUROC) by 0.075 while reducing within-specimen-type discrimination below chance levels.

LIS timestamps were used only to describe turnaround times and were never included as predictors in any model.

## Statistical analysis

The cohort was divided temporally. Tubes collected from October 2023 through December 2025 formed the development set (n=910,776), and tubes collected in 2026 formed the test set (n=280,168). Within the development set, the final 10% of observations by time were held out for early stopping.

Four algorithms were fitted: LightGBM [15], XGBoost [16], random forest, and L2-penalized logistic regression with a restricted cubic spline for age. No class weighting was applied. Gradient-boosting hyperparameters were explored across twelve configurations (15–255 leaves; learning rates of 0.015–0.05), and the number of boosting rounds was determined by early stopping using the time-ordered holdout set described above. Discrimination was summarized using AUROC and the area under the precision–recall curve (AUPRC), and calibration was assessed using the calibration slope, calibration-in-the-large (CITL), and scaled Brier score [17]. AUROC confidence intervals were calculated using bootstrap resampling clustered by presentation (400 replicates). Paired differences between algorithms were tested using the DeLong method [18]. Clinical usefulness was assessed using decision-curve analysis [19] and the proportion of rejections captured when a specified fraction of the highest-risk tubes was flagged.

To determine whether discrimination was limited by modeling or by available information, the same twelve gradient-boosting configurations were applied to the 23-variable set and an expanded 85-variable set that included panel-group indicators. Three additional blocks were then added sequentially: indicators for the 48 most frequently requested tests, rejection rates in the same unit over the preceding 7 and 30 days, and an indicator of prior rejection during the same presentation that had already been recorded at the time of the request. For this analysis, the highest test-set AUROC across configurations is reported. This represents an optimistic upper bound and is therefore conservative with respect to the conclusion that discrimination is limited by available information.

Adjusted associations were estimated using logistic regression with standard errors clustered by presentation. Categorical predictors were entered as indicator variables, with the reference categories specified in Table 1. Age was modeled using linear and quadratic terms, and workload was entered as a standardized variable. Predictor blocks were added sequentially and evaluated in the same test set to quantify the contribution of each block.

Model behavior was interpreted using Shapley additive explanations (SHAP), [20, 21] calculated with the TreeSHAP algorithm implemented in the shap library on 20,000 randomly selected test-set tubes. The tree-path-dependent formulation was used, in which the contributions of absent predictors were estimated from the training-data coverage of each tree path; therefore, no separate background dataset was required. Values are expressed on the model's log-odds scale. SHAP values describe how the fitted model uses each predictor rather than its causal or biological importance [22], and correlated predictors, such as specimen type and panel size, share attribution according to the structure learned by the trees [23].

For this revision, the complete pipeline was rerun from the raw LIS export using a fixed random seed (2026), and all settings were recorded. In the rerun, specimen type was assigned based on the defining test for each tube. As a result, 38,822 tubes (3.3%) originally classified as other specimens were reclassified into the blood-gas and coagulation categories. The rerun was used to assess the robustness of the main findings and to perform the specimen-quality analysis. All other estimates were derived from the original analysis, and the software versions listed below correspond to those used in the original analysis.

All analyses were performed in Python 3.11.15 (Python Software Foundation, Wilmington, Delaware, USA) with the open-source libraries scikit-learn 1.8.0, statsmodels 0.14.6, XGBoost 3.2.0, shap 0.51.0 and LightGBM 4.7.0 (Microsoft Corp., Redmond, Washington, USA). A two-sided $p<0.05$ was considered statistically significant.

## Sample size

No sample size was specified in advance because the study included every eligible tube collected during the extraction period. Sample size adequacy was assessed retrospectively according to the criteria of Riley et al. [24] for prediction models with a binary outcome. Assuming a development-set prevalence of 3.02%, a Cox–Snell $R^2$ of 0.034

**Table 1. Adjusted odds ratios for specimen rejection (n=1,190,944 tubes), from logistic regression with standard errors clustered on presentation**

| Variable | Adjusted OR | 95% CI | p |
|---|---|---|---|
| Specimen type: biochemistry (serum) | 10.48 | 10.05–10.93 | <0.001 |
| Specimen type: coagulation | 2.41 | 2.24–2.61 | <0.001 |
| Specimen type: urine | 1.89 | 1.78–2.02 | <0.001 |
| Triage area: red | 1.75 | 1.66–1.86 | <0.001 |
| Specimen type: complete blood count | 1.34 | 1.28–1.41 | <0.001 |
| Age, per 10 years | 1.17 | 1.14–1.21 | <0.001 |
| Triage area: yellow | 1.14 | 1.08–1.19 | <0.001 |
| Inpatient (vs outpatient) | 1.12 | 1.08–1.16 | <0.001 |
| Male sex | 1.12 | 1.09–1.15 | <0.001 |
| Specimen type: blood gas | 1.11 | 1.03–1.18 | 0.003 |
| Last 4 h of the 24-h nursing shift (binary indicator) | 1.07 | 1.01–1.14 | 0.03 |
| Tubes drawn simultaneously, per tube | 1.07 | 1.06–1.08 | <0.001 |
| Concurrent workload, per SD | 1.04 | 1.02–1.05 | <0.001 |
| Hours since handover, per SD | 1.02 | 0.99–1.06 | 0.15 |
| Pediatric unit | 1.02 | 0.95–1.09 | 0.62 |
| Weekend | 1.00 | 0.97–1.02 | 0.88 |
| Age, quadratic (plateau) term | 0.993 | 0.990–0.996 | <0.001 |
| Obstetric unit | 0.99 | 0.92–1.06 | 0.71 |
| Night, 00:00–08:00 | 0.97 | 0.89–1.05 | 0.46 |
| Evening, 16:00–24:00 | 0.95 | 0.91–1.00 | 0.05 |

Reference categories: other specimen type; green triage area; female sex; outpatient; adult unit; weekday; daytime (08:00–16:00). Before adjustment for specimen type the pediatric unit OR was 0.82 and the obstetric unit OR 0.64. The last-4-h term is a binary indicator entered in this model; Appendix 4 reports the same exposure as six 4-h blocks in a separate model, in which the final block gave an OR of 1.15 (1.09–1.21). CI: Confidence interval; OR: Odds ratio; SD: Standard deviation.

(corresponding to the observed C-statistic of 0.78), and up to 150 candidate predictor parameters, at least 39,168 development tubes (1,182 events) would be required to maintain expected shrinkage at or above 0.90 and minimize optimism in apparent model fit. The development set contained 910,776 tubes and 27,489 events, approximately 23 times the required sample size. The 2026 test set contained 10,144 events, substantially exceeding the recommended minimum of 100–200 events for validation [25].

## Results

### Cohort

Of the 1,190,944 tubes collected during 404,247 presentations, 37,633 (3.16%) were rejected (Table 2). Rejected and accepted tubes differed most in specimen type (standardized mean difference [SMD] for biochemistry, +1.08), age (+0.33), unit (adult, +0.31), triage area (red, +0.23), and concurrent workload (+0.19), whereas they were essentially identical in hour of request (−0.01), shift (≤0.02), and weekend status (0.00).

A black-area code, which was not part of the triage scheme used in this setting, appeared on 2,324 tubes (0.195%), almost all from the adult unit and distributed evenly throughout the study period, consistent with a data-entry origin. These tubes were analyzed as recorded. Of the 40,644 tubes recorded for patients aged 0 years, 99.6% came from pediatric units, consistent with actual infant age rather than a default entry.

### Causes of rejection and an artifactual trend

Hemolysis accounted for 25,439 rejections (67.6%), clotting for 4,518 (12.0%), insufficient volume for 2,487 (6.6%), absent blood-draw records for 2,130 (5.7%), and lipemia for 720 (1.9%) (Appendix 1a).

The category of absent blood-draw records first appeared in 2025 (n=390) and increased to 1,740 in 2026, accounting for 17.2% of that year's rejections. After excluding administrative categories, the rejection rate remained stable throughout the study period: 2.70%, 3.08%, 2.74%, and 2.90% for 2023–2026, compared with 2.78%, 3.18%, 2.92%, and 3.62% for all rejections (Appendix 1b, c). Administrative reasons therefore contributed 0.08, 0.10, and 0.18 percentage points to the annual rejection rate in 2023–2025, respectively, and 0.72 percentage points in 2026. The apparent 24% relative increase between 2025 and 2026 (chi-square test, $p=1.3\times10^{-62}$) was therefore largely a coding artifact. Over the same period, the proportion of rejections attributed to clotting decreased from 18.8% to 7.0%, whereas the proportion attributed to hemolysis remained between 63% and 70%.

### Model performance

LightGBM achieved an out-of-time AUROC of 0.780 (95% CI, 0.775–0.785), AUPRC 0.146 against a base rate of 3.62%, scaled Brier +5.66%, calibration slope 0.983 and CITL 0.239 (Appendix 2; Fig. 1a–c). Random forest reached 0.776, XGBoost 0.774

**Table 2. Characteristics of 1,190,944 emergency department specimen tubes by rejection status, with standardized mean differences**

| Characteristic | All tubes | Accepted | Rejected | SMD |
|---|---|---|---|---|
| Tubes, n | 1,190,944 | 1,153,311 | 37,633 | |
| ED presentations, n | 404,247 | 403,822 | 34,373 | |
| Age, mean (SD) | 39.2 (25.7) | 39.0 (25.7) | 47.3 (24.8) | +0.331 |
| Age, median (IQR) | 38 (18–61) | 38 (17–60) | 49 (27–68) | |
| 0–1 years, n (%) | 70,636 (5.9) | 69,164 (6.0) | 1,472 (3.9) | -0.096 |
| 2–17, n (%) | 224,036 (18.8) | 220,086 (19.1) | 3,950 (10.5) | -0.244 |
| 18–44, n (%) | 388,807 (32.6) | 377,602 (32.7) | 11,205 (29.8) | -0.064 |
| 45–64, n (%) | 259,011 (21.7) | 249,316 (21.6) | 9,695 (25.8) | +0.098 |
| 65–79, n (%) | 175,310 (14.7) | 167,495 (14.5) | 7,815 (20.8) | +0.164 |
| ≥80, n (%) | 73,144 (6.1) | 69,648 (6.0) | 3,496 (9.3) | +0.122 |
| Sex, n (%) | | | | |
| Female | 659,954 (55.4) | 640,344 (55.5) | 19,610 (52.1) | -0.069 |
| Male | 530,990 (44.6) | 512,967 (44.5) | 18,023 (47.9) | +0.069 |
| Emergency department unit, n (%) | | | | |
| Adult | 827,370 (69.5) | 796,506 (69.1) | 30,864 (82.0) | +0.305 |
| Pediatric 1 | 226,363 (19.0) | 222,217 (19.3) | 4,146 (11.0) | -0.232 |
| Obstetric | 76,301 (6.4) | 74,728 (6.5) | 1,573 (4.2) | -0.103 |
| Pediatric 2 | 28,655 (2.4) | 28,174 (2.4) | 481 (1.3) | -0.086 |
| Pediatric 3 | 26,448 (2.2) | 25,945 (2.2) | 503 (1.3) | -0.069 |
| Pediatric 4 | 5,807 (0.5) | 5,741 (0.5) | 66 (0.2) | -0.056 |
| Triage area, n (%) | | | | |
| Red area | 190,473 (16.0) | 181,184 (15.7) | 9,289 (24.7) | +0.225 |
| Yellow area | 880,019 (73.9) | 853,897 (74.0) | 26,122 (69.4) | -0.103 |
| Green area | 116,850 (9.8) | 114,757 (10.0) | 2,093 (5.6) | -0.165 |
| Black area | 2,324 (0.2) | 2,195 (0.2) | 129 (0.3) | +0.030 |
| Missing | 1,278 (0.1) | 1,278 (0.1) | 0 (0.0) | -0.047 |
| Specimen type, n (%) | | | | |
| Complete blood count | 335,042 (28.1) | 331,031 (28.7) | 4,011 (10.7) | -0.466 |
| Biochemistry (serum) | 309,048 (25.9) | 281,953 (24.4) | 27,095 (72.0) | +1.082 |
| Other | 279,155 (23.4) | 276,549 (24.0) | 2,606 (6.9) | -0.486 |
| Blood gas | 118,946 (10.0) | 117,574 (10.2) | 1,372 (3.6) | -0.260 |
| Urine | 111,092 (9.3) | 109,494 (9.5) | 1,598 (4.2) | -0.209 |
| Coagulation | 37,661 (3.2) | 36,710 (3.2) | 951 (2.5) | -0.039 |
| Inpatient, n (%) | 203,758 (17.1) | 195,522 (17.0) | 8,236 (21.9) | +0.125 |
| Weekend, n (%) | 338,601 (28.4) | 327,975 (28.4) | 10,626 (28.2) | -0.004 |
| Timing and workload | | | | |
| Hour of request, median (IQR) | 15 (10–19) | 15 (10–19) | 14 (10–19) | -0.014 |
| Hours since the 08:00 handover, median (IQR) | 10 (5–14) | 10 (5–14) | 10 (5–14) | -0.007 |
| Workload: tubes in unit that hour, median (IQR) | 34 (17–49) | 34 (17–49) | 38 (23–52) | +0.187 |
| Tubes drawn simultaneously, median (IQR) | 3 (2–4) | 3 (2–4) | 3 (3–4) | +0.170 |
| Shift 08–16, n (%) | 442,919 (37.2) | 428,566 (37.2) | 14,353 (38.1) | +0.020 |
| Shift 16–24, n (%) | 542,126 (45.5) | 525,428 (45.6) | 16,698 (44.4) | -0.024 |
| Shift 24–08, n (%) | 205,899 (17.3) | 199,317 (17.3) | 6,582 (17.5) | +0.005 |
| Study period, n (%) | | | | |
| Development (Oct 2023–Dec 2025) | 910,776 (76.5) | 883,287 (76.6) | 27,489 (73.0) | -0.082 |
| Test (Jan–Aug 2026) | 280,168 (23.5) | 270,024 (23.4) | 10,144 (27.0) | +0.082 |
| Missing data | | | | |
| Triage area missing, n (%) | 1,278 (0.107) | | | |
| Age/sex/unit missing, n | 0 | | | |

Values are N (%) unless stated otherwise. Indented rows are categories of the preceding variable. Hours are relative to the 08:00 nursing handover. The black-area triage code is not part of the triage scheme in use in this department and is considered a data-entry artefact; tubes were analyzed as recorded. IQR: Interquartile range; SD: Standard deviation; SMD: Standardized mean difference.

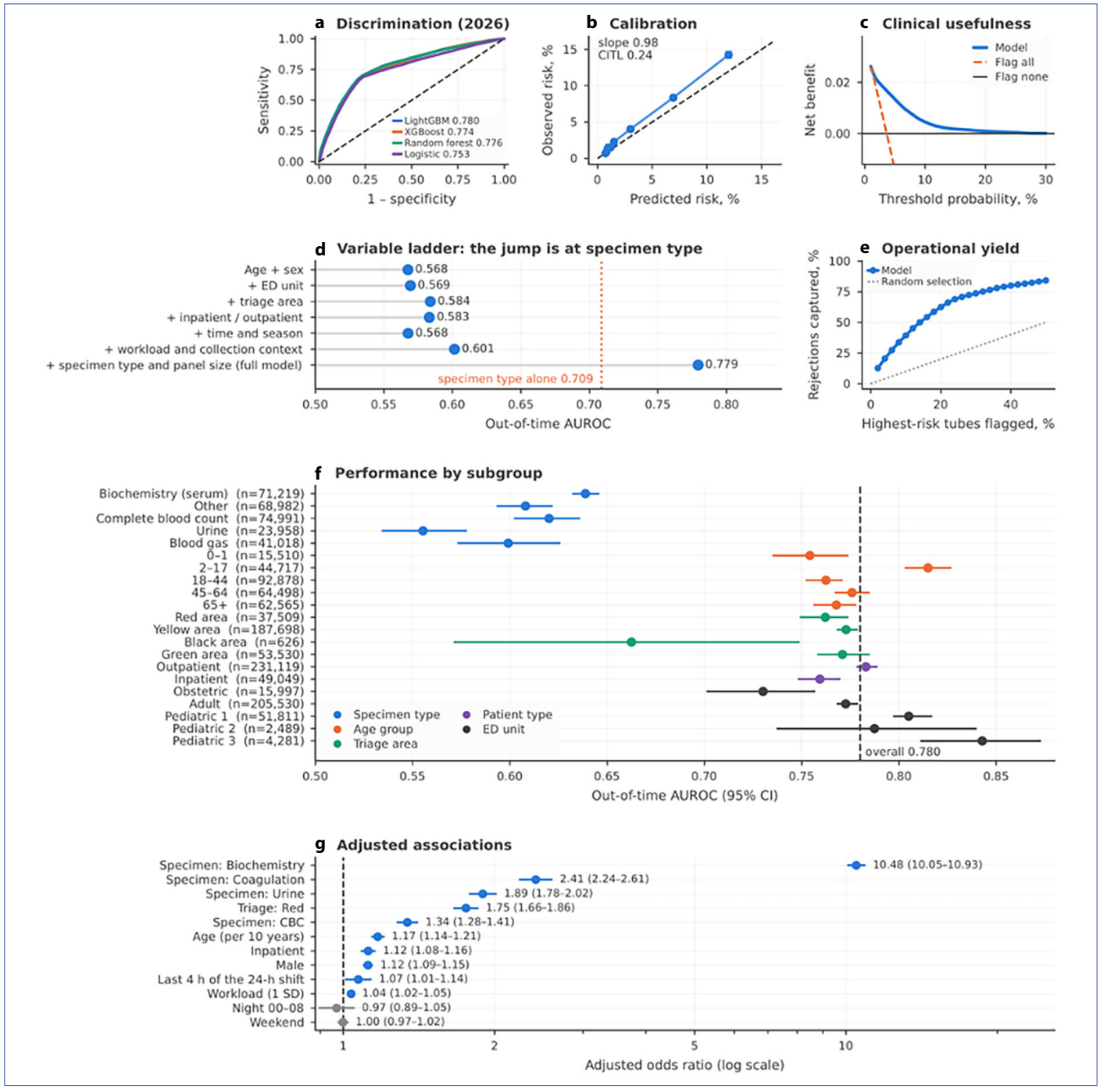



**Figure 1.** Model performance and predictors of specimen rejection. (a) ROC curves for four algorithms in the 2026 out-of-time test set. (b) Calibration of the LightGBM model by decile of predicted risk, with Wilson CIs; the dashed line indicates perfect calibration. (c) Decision-curve analysis comparing the model with flag-all and flag-none strategies. (d) Out-of-time AUROC as predictor blocks are added successively; the dotted vertical line marks the AUROC of a model containing specimen type alone. (e) Proportion of rejections captured as a function of the proportion of highest-risk tubes flagged; the diagonal represents random selection. (f) Out-of-time AUROC with 95% CIs within subgroups defined by specimen type, age group, triage area, patient type and emergency department unit; the dashed vertical line indicates overall performance. (g) Adjusted ORs with 95% CIs; gray markers indicate intervals crossing unity.

CI: Confidence interval; OR: Odds ratio; SD: Standard deviation.

and penalized logistic regression 0.753. Paired DeLong differences favored LightGBM over all three comparators (all $p<0.001$), but differences between tree-based algorithms were small (≤0.007) and their ranking was not stable: in the re-run described below, random forest reached 0.787 and XGBoost 0.783, against 0.783 for LightGBM.

Calibration was achieved without class weighting. With class weighting, discrimination was unchanged but the calibration slope rose to 1.27 and CITL fell to −3.05.

## Where the discrimination originates

Adding predictor blocks successively (Fig. 1d) gave 0.568 for age and sex, 0.569 after unit, 0.584 after triage area and 0.583 after admission status; the timing block did not improve discrimination (0.568), workload and collection context raised it to 0.601, and specimen type with panel size raised it to 0.779. Specimen type alone, a single six-level variable, achieved 0.709.

## Is 0.78 a modeling limit or an information limit?

In the original analysis, no configuration and no feature set exceeded an out-of-time AUROC of 0.781 (Appendix 3). The 85-variable model reached 0.781 and the 23-variable model 0.780, and every added feature block reduced discrimination (to 0.775–0.777). The ceiling therefore reflects the information content of the recorded variables, not the choice of algorithm or hyperparameters.

## Adjusted associations

Specimen type was the strongest predictor (Table 1; Fig. 1g). Relative to the "other" specimen category, biochemistry tubes carried an adjusted OR of 10.48 (10.05–10.93), coagulation tubes 2.41 (2.24–2.61), urine 1.89 (1.78–2.02), CBC 1.34 (1.28–1.41) and blood gas 1.11 (1.03–1.18).

Among patient variables, red-area triage carried 1.75 (1.66–1.86) and yellow-area 1.14 (1.08–1.19) relative to green; age 1.17 (1.14–1.21) per 10 years, with a plateau term of 0.993 (0.990–0.996); male sex and admission 1.12 each.

Two associations observed in unadjusted analysis disappeared after adjustment for specimen type. The pediatric unit OR moved from 0.82 to 1.02 (0.95–1.09; p=0.62) and the obstetric unit OR from 0.64 to 0.99 (0.92–1.06; p=0.71). The lower crude rejection rates in pediatric and obstetric units were therefore largely accounted for by the different mix of specimens requested there, and the adjusted estimates were close to unity.

## Time of day, shift hours and workload

Because nursing shifts are 24 h long with a single 08:00 handover, hours on duty and clock time are the same variable and cannot be separated in this design.

With hours since handover entered in 4-h blocks and the first as reference (Appendix 4, 5a), the OR was at or near unity through hours 4–16, rose to 1.05 (1.01–1.10) for hours 16–20 and reached 1.15 (1.09–1.21) in the final 4 h (04:00–08:00). Restricted to biochemistry tubes with hemolysis as the cause, the final-block OR was 1.14 (1.07–1.22). The crude hemolysis rate in biochemistry tubes rose from 7.4% at handover to 9.2%–9.7% in the last three hours of the shift.

Weekend status was not associated with rejection (OR, 1.00; 0.97–1.02; p=0.88). Concurrent workload reached statistical significance but not clinical relevance (OR, 1.04 per SD; 1.02–1.05).

## Interpretation

SHAP attribution ranked panel size first (mean |SHAP|, 0.571), followed by specimen type (0.274), calendar time (0.190), age (0.136) and triage area (0.118); all remaining predictors contributed ≤0.057 (Fig. 2a, b). These values describe how much each variable moved the model's predictions, not biological importance. Because large panels are ordered almost exclusively on serum biochemistry tubes, panel size and specimen type share credit, and the high rank of panel size indicates that it carries much of the specimen-related information used by the model, not that ordering more tests causes rejection.

Dependence plots (Fig. 2c) showed a monotonic age relationship, with markedly lower predicted risk below 5 years, a steep rise to age 18 and a plateau after 60; a step-like panel-size pattern, with lower predicted risk for single-test tubes and higher predicted risk for panels of 15 or more tests (up to +1.2 log-odds); and an ordering of triage areas from red to green.

## Subgroup performance

Discrimination was preserved across every patient stratum (Appendix 6; Fig. 1f), from 0.730 (0.701–0.757) in the obstetric unit to 0.843 (0.811–0.873) in one pediatric unit, with 0.754–0.815 across age groups and 0.762–0.773 across the red, yellow and green triage areas; the small black-area stratum (N = 626) gave 0.663 (0.571–0.749).

Within strata defined by specimen type, discrimination fell as expected when the dominant predictor is held constant: 0.639 for biochemistry, 0.620 for CBC, 0.608 for other specimens, 0.599 for blood gas and 0.555 for urine. All remained above chance, although calibration slopes varied between strata (0.84–1.42).

## Restricting the outcome to hemolysis

Hemolysis was almost exclusively a biochemistry-tube phenomenon: 24,904 of 25,439 hemolysis rejections (97.9%) occurred in biochemistry tubes, where hemolysis represented 91.9% of all rejections. Hemolysis was recorded on 181 coagulation tubes and, implausibly, on 21 specimens in which serum hemolysis is not assessed; these were treated as data-entry errors (Fig. 3a).

Restricting the outcome to hemolysis while retaining all tubes raised the out-of-time AUROC to 0.907 (0.906–0.909). However, specimen type alone predicted this outcome at 0.871, and specimen type with panel size at 0.892 (Table 3; Fig. 3b). Models in this comparison were developed within the stated population and outcome, so their estimates differ slightly from the stratified evaluation of the primary model (Appendix 6). The increment attributable to every patient, timing, workload and collection variable combined was therefore 0.015.

Restricting both the population to biochemistry tubes and the outcome to hemolysis, the clinically coherent design, gave 0.652 (0.644–0.659) across 71,219 test-set tubes and 6,206 events, against 0.633 for any rejection in the same population. Within biochemistry tubes, age alone gave 0.585 and all patient variables together 0.609.

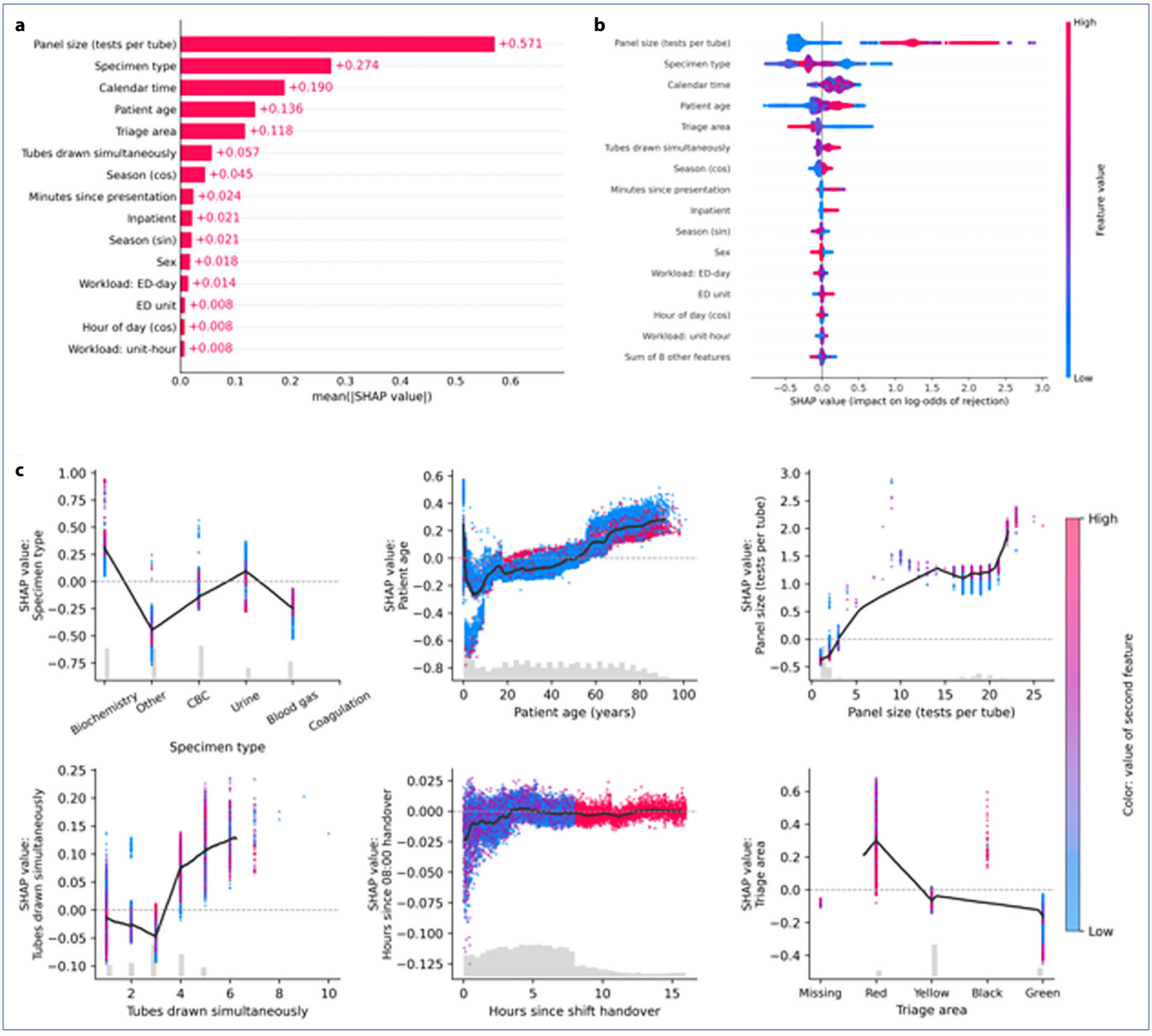



**Figure 2.** SHAP interpretation of the LightGBM model. (a) Mean absolute SHAP value for the 15 highest-ranked predictors. (b) Beeswarm plot of SHAP values for the 16 highest-ranked predictors; color encodes predictor value; positive values indicate higher predicted risk. (c) Dependence plots for specimen type, patient age, panel size, tubes drawn simultaneously, hours since the 08:00 handover and triage area; gray histograms show the marginal distribution and the black line a moving average. SHAP values are on the log-odds scale and describe the model's use of each predictor, not causal or biological importance; correlated predictors, such as panel size and specimen type, share attribution.

SHAP: Shapley additive explanations.

## Re-run and specimen-quality outcome

The re-run from the raw export reproduced the cohort counts exactly (1,190,944 tubes; 37,633 rejections; 10,144 test-set events). LightGBM reached an out-of-time AUROC of 0.783 (0.777–0.787), the twelve configurations 0.776–0.784 and the four algorithms 0.766–0.787; the adjusted OR for the final 4 h of the shift was 1.15 (1.09–1.21) (Appendix 7, 8). With administrative-only rejections treated as non-events (34,428 events; 8,128 in the test set), discrimination was 0.838 (0.834–0.842) against 0.797 for specimen type alone, an increment similar to that for any rejection in the same run (0.783 against 0.737). Specimen type remained the strongest predictor (biochemistry OR, 20.9; 19.6–22.2), the pediatric and obstetric unit ORs moved from 0.80 and 0.58 before adjustment to 1.03 (0.96–1.10) and 0.95 (0.88–1.02) after it, and the final 4 h of the shift carried an OR of 1.11 (1.05–1.18). Concurrent workload was no longer associated with rejection (OR, 0.98 per SD; 0.97–1.00).

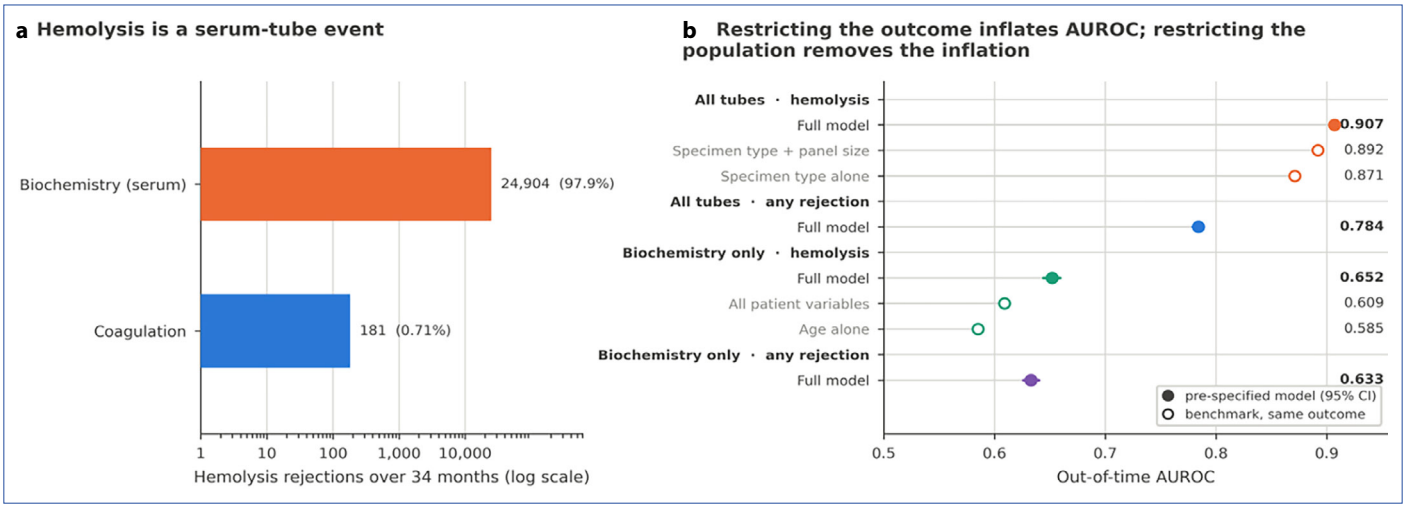



**Figure 3.** The specimen-type trap in hemolysis prediction. (a) Number of hemolysis rejections recorded over the 34-month cohort in the two specimen types in which serum hemolysis is assessed, on a logarithmic scale. A further 21 hemolysis codes recorded on urine, complete blood count and blood gas specimens were treated as data-entry errors and are not shown. (b) Out-of-time AUROC for each combination of analysis population and outcome. Filled markers are the pre-specified model with its 95% CI; open markers are benchmark models fitted to the same outcome. Restricting the outcome to hemolysis while retaining all tubes raises AUROC to 0.907, but specimen type entered alone reaches 0.871 for the same outcome; restricting the population to the specimen type in which the outcome can occur removes the inflation.

AUROC: Area under the receiver operating characteristic curve.

## Clinical usefulness

Flagging the highest-risk 4% of tubes captured 20.5% of rejections at a positive predictive value (PPV) of 18.6% against a base rate of 3.62%; 10% captured 39.4% (PPV, 14.3%) and 20% captured 62.5% (PPV, 11.3%), a 3.1-fold enrichment. Decision-curve analysis showed net benefit above both flag-all and flag-none strategies at threshold probabilities of 1%–30% (Fig. 1c, 1e).

## Turnaround

Median intervals from request were 53 min to result, 61 min to technician approval and 60 min to specialist approval; a rejection was recorded at a median of 62 min (Appendix 1d). For a rejected specimen this interval is lost in full, and a repeat draw with a further cycle follows.

## Discussion

In 1.19 million consecutive emergency department tubes, specimen type was the strongest predictor of laboratory rejection. Biochemistry serum tubes had tenfold higher adjusted odds of rejection than tubes in the "other" specimen category, and specimen type alone discriminated better (AUROC, 0.709) than all patient, timing and workload variables combined (0.601). Because the design is observational, these findings describe associations and sources of predictive discrimina-

**Table 3. Effect of restricting the outcome to hemolysis and the population to biochemistry tubes**

| Population | Outcome | Test-set N | Events | Event rate, % | AUROC | 95% CI |
|---|---|---|---|---|---|---|
| All tubes | Any rejection | 280,168 | 10,144 | 3.62 | 0.784 | 0.779–0.788 |
| All tubes | Hemolysis | 280,168 | 6,463 | 2.31 | 0.907 | 0.906–0.909 |
| All tubes - specimen type alone | Hemolysis | 280,168 | 6,463 | 2.31 | 0.871 | – |
| All tubes - specimen type + panel size | Hemolysis | 280,168 | 6,463 | 2.31 | 0.892 | – |
| Biochemistry (serum) | Any rejection | 71,219 | 7,016 | 9.85 | 0.633 | 0.626–0.640 |
| Biochemistry (serum) | Hemolysis | 71,219 | 6,206 | 8.71 | 0.652 | 0.644–0.659 |
| Biochemistry - age alone | Hemolysis | 71,219 | 6,206 | 8.71 | 0.585 | – |
| Biochemistry - all patient variables | Hemolysis | 71,219 | 6,206 | 8.71 | 0.609 | – |
| Complete blood count | Any rejection | 74,991 | 1,037 | 1.38 | 0.598 | 0.580–0.614 |
| Blood gas | Any rejection | 41,018 | 261 | 0.64 | 0.610 | 0.582–0.643 |
| Urine | Any rejection | 23,958 | 639 | 2.67 | 0.520 | 0.498–0.542 |

Models in this table were developed and evaluated within the stated population and outcome. Across the full 34-month cohort, 24,904 of 25,439 hemolysis rejections (97.9%) occurred in biochemistry tubes and 181 in coagulation tubes; hemolysis was also coded on 21 specimens in which serum hemolysis is not assessed (12 urine, seven complete blood count, two blood gas), and these records were treated as data-entry errors. The all-tubes, any-rejection estimate in this table (0.784) therefore differs marginally from the primary model (0.780; 95% CI, 0.775–0.785; Appendix 2). AUROC: Area under the receiver operating characteristic curve; CI: Confidence interval.

tion rather than causal effects; they nevertheless define what a preanalytical prediction model built on routine laboratory data can be expected to deliver.

The overall rejection rate of 3.16% and the hemolysis share of 67.6% are consistent with the published literature. Hemolysis accounts for 40%–70% of unsuitable specimens in clinical laboratories [4], and a recent analysis of approximately 11 million specimens attributed 69.6% of laboratory errors to hemolysis [3]. Laboratory-wide rejection rates are lower than ours: 0.65% in a Turkish clinical biochemistry laboratory that excluded emergency samples [7], and 0.79% in a recent Turkish series in which the emergency department nevertheless had the highest unit-level rate, 1.92% [26]. Hemolysis is considerably more frequent in emergency care: 6.8%–19.8% in most departments included in a systematic review and above 30% in two [5], and 17.2% of adult emergency encounters and 14.0% of pediatric emergency encounters with intravenous catheter placement in recent US cohorts [27, 28]. Rejection is less frequent than these prevalences because it is applied only when hemolysis threatens the analytes requested, so comparisons between studies must take this difference into account.

Specimen type is not a nuisance variable. It is measurable, known at order entry and biologically coherent: hemolysis interferes selectively with serum biochemistry [4]. Any model trained on LIS data will therefore encode specimen type, whether or not it is named, and reported discrimination should be benchmarked against a specimen-type-only model rather than against chance. In our data that benchmark was 0.709, so the full model's 0.780 represents an increment of 0.071 attributable to everything else recorded.

The hemolysis analysis makes this concrete. Restricting the outcome to hemolysis raises apparent discrimination to 0.907. Yet specimen type alone reaches 0.871 for the same outcome, because 97.9% of hemolysis occurs in biochemistry tubes: a model reported at 0.907 without that benchmark is, in effect, a serum-tube detector. Restricted to biochemistry tubes, where the question is clinically meaningful, discrimination falls to 0.652. Studies of hemolysis prediction should report the specimen-type-only comparator alongside their headline figure and restrict the analysis population to the specimen types in which the outcome can occur.

Published machine-learning applications in the preanalytical phase report much higher discrimination, but they address a different question with features that contain the error signal. Delta-check models for wrong-blood-in-tube errors reached AUROCs of 0.97–0.998 using consecutive results of multiple analytes, with errors created by simulated sample swaps [10, 11]. A model detecting falsely elevated point-of-care potassium in 3,489 emergency encounters reached 0.995 using a panel that included the potassium result itself [9], and image-based classification of serum hemolysis, icterus and lipemia reached 0.983–0.999 in 16,427 sample images [8]. These models detect an error after analysis or imaging, whereas ours estimates risk before collection, a harder task. In routine use, a wrong-blood-in-tube model developed on retrospective data achieved a PPV of only 10.9% [12]. The PPVs of 11%–19% at our operational thresholds, and within-specimen-type AUROCs of 0.56–0.64, should be read in that light.

The main associations reported here, namely the dominance of specimen type, the disappearance of the pediatric and obstetric unit associations after adjustment and the late-shift excess, were estimated by logistic regression (Table 1) and do not depend on the machine-learning models; penalized logistic regression also reached an AUROC of 0.753, close to the 0.780 of LightGBM. This agrees with systematic evidence that machine learning offers no consistent advantage over logistic regression for clinical prediction [29]. The machine-learning framework was used for two narrower purposes: to establish an information ceiling, by showing that no configuration, feature set or algorithm moved discrimination materially beyond 0.78, and to characterize non-linear relationships, such as the age and panel-size profiles, without specifying their form in advance. It was not intended to deliver a deployable prediction tool.

Pediatric and obstetric units had lower unadjusted odds of rejection (ORs, 0.82 and 0.64), but both associations disappeared after adjustment for specimen type, because these units order proportionally fewer serum biochemistry panels. This does not contradict reports of frequent hemolysis in pediatric emergency care. In 10,462 pediatric emergency encounters with intravenous catheter placement, hemolysis affected 14.0% of encounters and 20.1% of those in infants, rates similar to the 17.2% reported in adult emergency encounters [27, 28]. Those studies measured hemolysis per encounter among patients who had blood drawn, whereas our denominator includes every tube ordered, with a specimen mix that differs between units. Unadjusted comparisons between collection areas, a common form of internal quality benchmarking, will therefore misattribute differences in specimen mix to collection technique.

Weekend status was not associated with rejection, and concurrent workload, although statistically significant at this sample size, corresponded to a 4% change in odds per SD. This association disappeared when administrative rejections were excluded. The closest published evidence, on blood-culture contamination, showed a modest and imprecise excess in the busiest quartile of hourly emergency department volume [30]. Odds were close to unity through the first 16 h of the 24-h shift, rose slightly at 16–20 h and were highest in the final 4 h (adjusted OR, 1.15). A recent Turkish series observed the same pattern by clock time: rejection was highest between 04:00 and 08:00 (incidence rate ratio, 1.63) [26]. In industrial studies, risk rises both on night shifts and with hours on duty, and the two have been described as separate components [31]; among hospital nurses, error risk increases with shifts exceeding 12 h [32]. In our single-handover roster the last hours of duty coincide with the early morning, and fatigue was not measured. The excess may therefore reflect accumulated fatigue, circadian factors, early-morning staffing or case mix, or a combination, and these explanations cannot be distinguished with the present data; designs with staggered handovers, or comparisons between rosters, would be required.

An administrative category silently inflated a quality indicator. The institutional rejection rate appeared to rise by 24% between 2025 and 2026. Most of that rise (0.54 of 0.70 percentage points) was attributable to administrative reasons, chiefly a category introduced in 2025 that records a missing blood-draw entry rather than a defective specimen. Once administrative categories were removed, the rate remained between 2.7% and 3.1% in every year. The numerator of this widely used quality indicator [14] is defined by whatever categories the LIS offers, so any institution reporting a trend should confirm that the category set was stable over the reporting window; ours was not, and the effect was large enough to have prompted an unnecessary intervention.

Leakage in routine laboratory data is subtle. Result-row counts and the export-derived position of a tube within a presentation would both be natural choices for an analyst, yet both encoded the outcome or specimen type. A variable that raises overall AUROC while driving within-stratum discrimination below chance, as tube position did, should prompt suspicion.

The ceiling is an information limit (Appendix 5b). Tuning configurations and feature sets extending to individual test flags, unit-level rejection history and prior rejections known at the time of request all converged on an AUROC of about 0.78, and the four algorithms gave 0.766–0.787 in the re-run. The missing variables are those identified as the strongest modifiable risk factors for hemolysis in emergency care: collection through an intravenous catheter rather than by venipuncture, the choice of collection site [5, 33, 34] and the training of the person drawing the sample [35]. None is recorded in the LIS, and no amount of modeling on LIS data will substitute for them.

A calibrated per-tube risk supports targeted rather than universal intervention: flagging the highest-risk fifth of tubes would capture 62.5% of rejections. The variables carrying that risk are, however, largely not modifiable at the bedside. The actionable findings are narrower: the final hours of the shift are a reasonable target for reinforcing collection practice, and specimen-mix differences between units should be treated as an artefact to be adjusted for rather than a performance signal.

This is a single-center study without external validation, and data from other institutions were not available to us. The specimen mix that underlies discrimination, rejection policies, LIS category sets and collection practices all differ between institutions, so the model's coefficients and absolute risks are unlikely to transport without recalibration. Temporal validation on 2026 data addresses reproducibility within the institution but not transportability to a different case mix [36]. The findings most likely to generalize are methodological: the specimen-type benchmark, the reporting of within-stratum performance and the separation of administrative from specimen-quality rejection reasons. External validation in other emergency departments, ideally with harmonized rejection categories and recorded collection-process variables, is required before any use of the model outside this setting [37].

Other limitations apply. Beyond the collection-process variables discussed above, the grade of the collecting staff member, tourniquet time, needle gauge and transport method are not recorded in the LIS. Their absence is the most plausible explanation for the ceiling observed within specimen strata. Serum index values were unavailable, so hemolysis could be studied only through the laboratory's rejection decision, which depends on the analyte profile requested and is therefore not a pure measure of specimen quality. The primary outcome combined specimen-quality and administrative rejections, and the 2026 test set contained the largest share of administrative reasons. The principal findings were unchanged when administrative-only rejections were treated as non-events, but the small workload association was not, suggesting that it partly reflects administrative error. The 2023 data begin on 6 October, limiting the seasonal analysis. Circadian and fatigue effects are not separately identifiable.

In conclusion, across 1.19 million emergency department tubes, specimen rejection was predicted principally by the type of specimen required, secondarily by patient characteristics, and only marginally by the time of collection or departmental workload. Prediction models built on routine laboratory data inherit this structure and should be benchmarked against a specimen-type-only comparator and reported with within-stratum performance. Two findings have immediate operational value: an apparent institutional deterioration in rejection rate was produced by a change in rejection coding, and apparent differences between collection areas were explained by specimen mix rather than collection technique.

## Disclosures

**Ethics Committee Approval:** The study was approved by the Gaziantep City Hospital Ethics Committee (no: 624/2026, date: 22/07/2026).

**Informed Consent:** Informed consent was not obtained. The ethics committee waived the requirement because the study was a retrospective analysis of anonymized routine laboratory data collected in the course of standard clinical care.

**Conflict of Interest Statement:** None declared.

**Funding:** The author declares that this study has received no financial support.

**Use of AI for Writing Assistance:** An AI assistant (Claude, Anthropic) was used to support data processing and analysis code, figure preparation, language editing, formatting, and reference checking. All analyses were performed on the authors' data under their direction; the authors reviewed and verified all code, results, text and references and take full responsibility for the content of the manuscript.

**Authorship Contributions:** Concept – Y.G.; Design – Y.G.; Supervision – Y.G.; Funding – Y.G., H.O.; Materials – H.O.; Data collection and/or processing – Y.G., H.O.; Data analysis and/or interpretation – Y.G.; Literature search – Y.G., H.O.; Writing – Y.G.; Critical review – H.O.

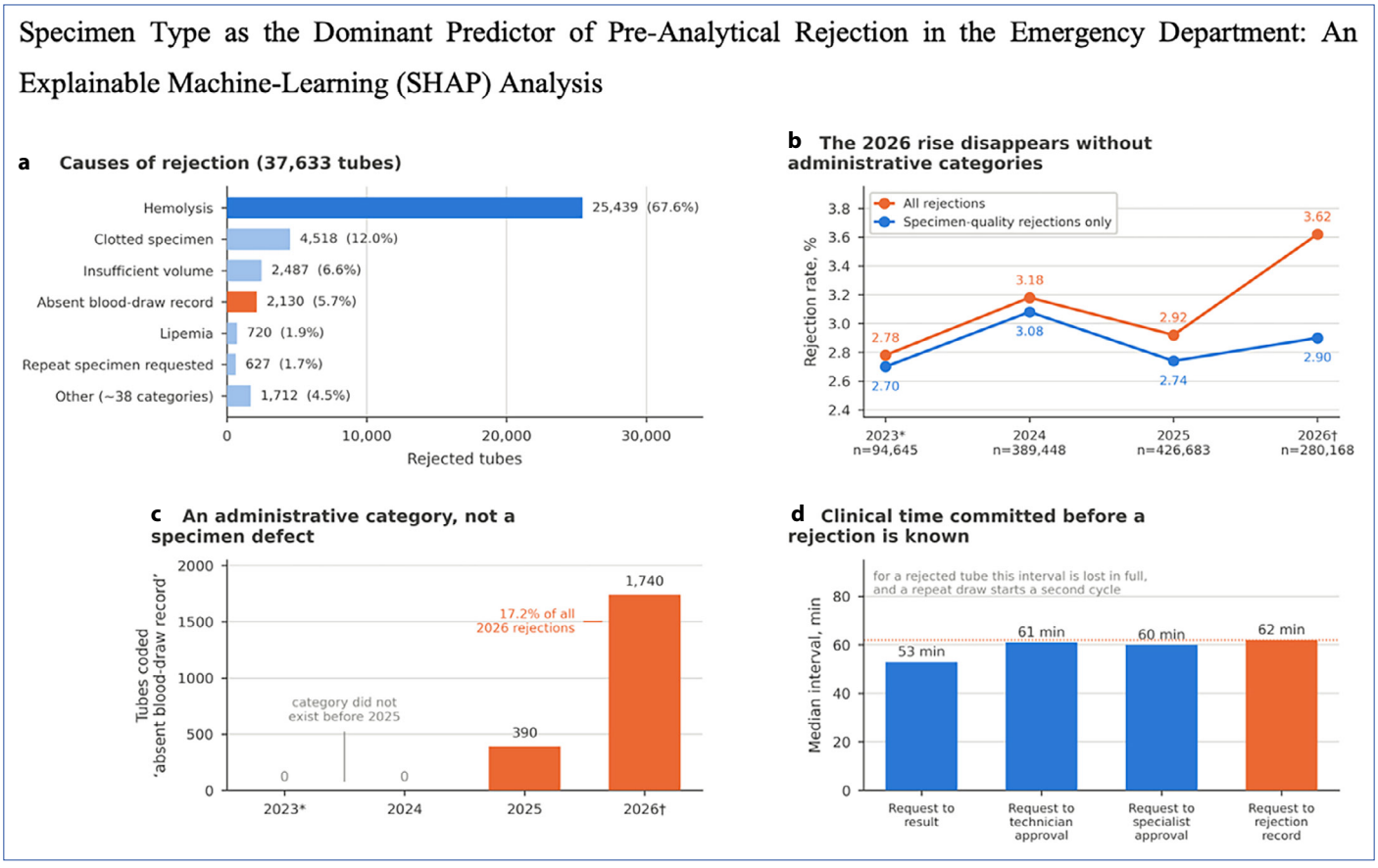
**Acknowledgments:** The authors thank the laboratory information system team of Gaziantep City Hospital for preparing the data extract.

**Peer-review:** Externally peer-reviewed.

## References


1. Plebani M. Errors in clinical laboratories or errors in laboratory medicine? Clin Chem Lab Med 2006;44(6):750–9. [CrossRef]
2. Carraro P, Plebani M. Errors in a stat laboratory: types and frequencies 10 years later. Clin Chem 2007;53(7):1338–42. [CrossRef]
3. Lin Y, Spies NC, Zohner K, McCoy D, Zaydman MA, Farnsworth CW. Pre-analytical phase errors constitute the vast majority of errors in clinical laboratory testing. Clin Chem Lab Med 2025;63(9):1709–15. [CrossRef]
4. Lippi G, Blanckaert N, Bonini P, Green S, Kitchen S, Palicka V, et al. Haemolysis: an overview of the leading cause of unsuitable specimens in clinical laboratories. Clin Chem Lab Med 2008;46(6):764–72. [CrossRef]
5. Heyer NJ, Derzon JH, Winges L, Shaw C, Mass D, Snyder SR, et al. Effectiveness of practices to reduce blood sample hemolysis in EDs: a laboratory medicine best practices systematic review and meta-analysis. Clin Biochem 2012;45(13-14):1012–32. [CrossRef]
6. Lippi G, Bonelli P, Cervellin G. Prevalence and cost of hemolyzed samples in a large urban emergency department. Int J Lab Hematol 2014;36(1)–6. [CrossRef]
7. Atay A, Demir L, Cuhadar S, Saglam G, Unal H, Aksun S, et al. Clinical biochemistry laboratory rejection rates due to various types of preanalytical errors. Biochem Med (Zagreb) 2014;24(3):376–82. [CrossRef]
8. Yang C, Li D, Sun D, Zhang S, Zhang P, Xiong Y, et al. A deep learning-based system for assessment of serum quality using sample images. Clin Chim Acta 2022;531:254–60. [CrossRef]
9. Benirschke RC, Gniadek TJ. Detection of falsely elevated point-of-care potassium results due to hemolysis using predictive analytics. Am J Clin Pathol 2020;154(2):242–7. [CrossRef]
10. Rosenbaum MW, Baron JM. Using machine learning-based multianalyte delta checks to detect wrong blood in tube errors. Am J Clin Pathol 2018;150(6):555–66. [CrossRef]
11. Mitani T, Doi S, Yokota S, Imai T, Ohe K. Highly accurate and explainable detection of specimen mix-up using a machine learning model. Clin Chem Lab Med 2020;58(3):375–83. [CrossRef]
12. Farrell CJL, Makuni C, Keenan A, Maeder E, Davies G, Giannoutsos J. A machine learning model for the routine detection of "wrong blood in complete blood count tube" errors. Clin Chem 2023;69(9):1031–7. [CrossRef]
13. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385. [CrossRef]
14. Sciacovelli L, Lippi G, Sumarac Z, West J, Garcia Del Pino Castro I, Furtado Vieira K, et al. Quality indicators in laboratory medicine: the status of the progress of IFCC Working Group "Laboratory Errors and Patient Safety" project. Clin Chem Lab Med 2017;55(3):348–57. [CrossRef]
15. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30, 2017:3146–54.
16. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016:785–94. [CrossRef]
17. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17(1):230. [CrossRef]
18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–45. [CrossRef]
19. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565–74. [CrossRef]
20. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30, 2017:4765–74. Available at: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html Accessed October 09, 2026.
21. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56–67. [CrossRef]
22. Janzing D, Minorics L, Blöbaum P. Feature relevance quantification in explainable AI: a causal problem. Proc Mach Learn Res 2020;108:2907–16. Available at: https://proceedings.mlr.press/v108/janzing20a.html
23. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artif Intell 2021;298:103502. [CrossRef]
24. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368. [CrossRef]
25. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med 2016;35(2):214–26. [CrossRef]
26. Dündar B, Özbek B, Bozkurt F, Gök Yurttaş A. Can time determine preanalytical quality? A temporal analysis of specimen rejection rates. J Clin Med 2026;15(12):4752. [CrossRef]
27. Calleja R, Mielke N, Lee R, Johnson S, Bahl A. Hemolyzed laboratory specimens in the emergency department: an underappreciated, but frequent problem. J Emerg Nurs 2023;49(5):744–54. [CrossRef]
28. Mielke N, Lee R, Bahl A. Pediatric hemolysis in emergency departments: prevalence, risk factors, and clinical implications. PLoS One 2024;19(3). [CrossRef]
29. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12–22. [CrossRef]

30. Halverson S, Malani PN, Newton DW, Habicht A, Vander Have K, Younger JG. Impact of hourly emergency department patient volume on blood culture contamination and diagnostic yield. J Clin Microbiol 2013;51(6):1721–6. [CrossRef]
31. Folkard S, Tucker P. Shift work, safety and productivity. Occup Med (Lond) 2003;53(2):95–101. [CrossRef]
32. Rogers AE, Hwang WT, Scott LD, Aiken LH, Dinges DF. The working hours of hospital staff nurses and patient safety. Health Aff (Millwood) 2004;23(4):202–12. [CrossRef]
33. Lippi G, Cervellin G, Mattiuzzi C. Critical review and meta-analysis of spurious hemolysis in blood samples collected from intravenous catheters. Biochem Med (Zagreb) 2013;23(2):193–200. [CrossRef]
34. Clinical and Laboratory Standards Institute. Collection of diagnostic venous blood specimens. 7th ed. CLSI standard GP41. Wayne, PA: Clinical and Laboratory Standards Institute; 2017. Available at: https://clsi.org/standards/products/general-laboratory/documents/gp41/ Accessed October 09, 2026.
35. Cadamuro J, Meyer A, Wiedemann H, Felder TK, Moser F, Kipman U, et al. Hemolysis rates in blood samples: differences between blood collected by clinicians and nurses and the effect of phlebotomy training. Clin Chem Lab Med 2016;54(12):1987–92. [CrossRef]
36. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68(3):279–89. [CrossRef]
37. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353. [CrossRef]




**Appendix 1.** Causes of rejection, an administrative coding artefact, and committed turnaround time. (a) Distribution of recorded rejection reasons across 37,633 rejected tubes; the category shown in orange is administrative rather than a specimen defect. (b) Annual rejection rate for all rejections and for specimen-quality rejections alone; the apparent rise in 2026 is absent once administrative categories are removed. (c) Annual count of tubes coded absent blood-draw record; the category did not exist before 2025. (d) Median intervals from request to each laboratory milestone.

*: 2023 data begin on 6 October; †: 2026 data end on 10 August.

**Appendix 2. Out-of-time performance of four algorithms in the 2026 test set (n=280,168; 10,144 rejections, 3.62%)**

| Algorithm | AUROC (95% CI) | AUPRC | Scaled Brier, % | Calibration slope | CITL | ΔAUROC vs LightGBM (95% CI) | p |
|---|---|---|---|---|---|---|---|
| LightGBM (pre-specified final model) | 0.780 (0.775–0.785) | 0.146 | +5.66 | 0.983 | 0.239 | Reference | - |
| Random forest | 0.776 | 0.139 | +5.50 | 1.089 | 0.059 | 0.0036 (0.0021–0.0051) | <0.001 |
| XGBoost | 0.774 | 0.137 | +5.32 | 0.798 | 0.294 | 0.0062 (0.0047–0.0078) | <0.001 |
| Penalized logistic regression | 0.753 | 0.111 | +4.22 | 0.847 | 0.275 | 0.0272 (0.0241–0.0303) | <0.001 |

The AUROC confidence interval is a cluster bootstrap interval (400 replicates) resampled on presentation; ΔAUROC and its interval are paired DeLong estimates. Class weighting was not applied; with class weighting, discrimination was unchanged, the calibration slope rose to 1.27 and CITL fell to −3.05. AUPRC: Area under the precision–recall curve; AUROC: Area under the receiver operating characteristic curve; CI: Confidence interval; CITL: Calibration-in-the-large.

**Appendix 3. Out-of-time AUROC across hyperparameter configurations and feature sets**

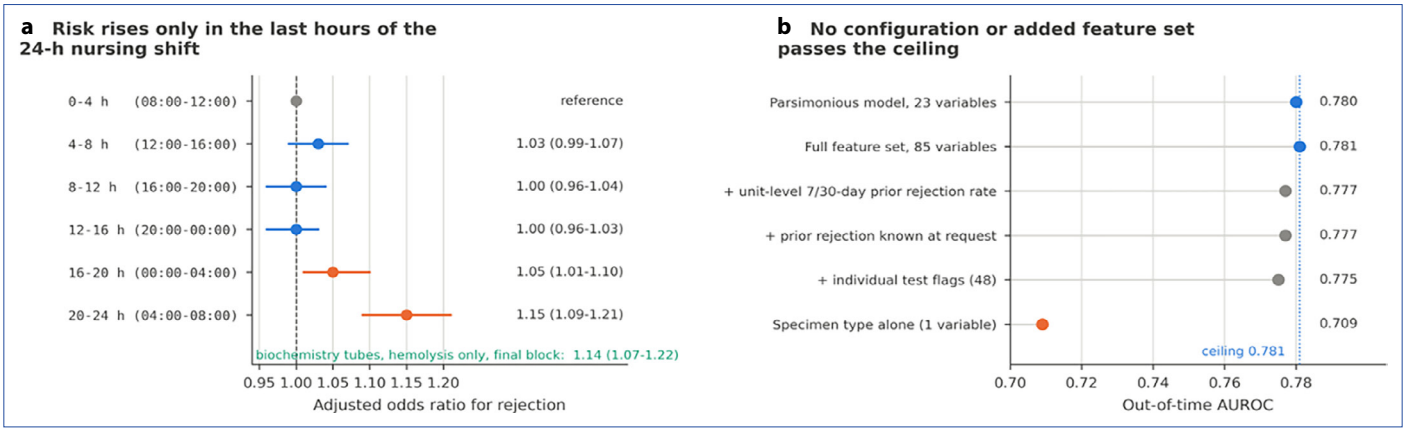
| Model / feature set | Variables | Out-of-time AUROC |
|---|---|---|
| Parsimonious model (pre-specified) | 23 | 0.780 |
| Full feature set | 85 | 0.781 |
| + unit-level rejection rate over the preceding 7 and 30 days | 85 + 2 | 0.777 |
| + indicator of a prior rejection known at the time of request | 85 + 1 | 0.777 |
| + individual test flags | 85 + 48 | 0.775 |
| Specimen type alone | 1 | 0.709 |

Twelve gradient-boosting configurations were evaluated (15–255 leaves; learning rate 0.015–0.05) against each feature set; the table reports the best value obtained for each set. No configuration exceeded 0.781. AUROC: Area under the receiver operating characteristic curve.

**Appendix 4. Adjusted odds ratio for rejection by hours elapsed since the 08:00 shift handover**

| Hours since handover | Clock time | Adjusted OR | 95% CI |
|---|---|---|---|
| 0–4 h | 08:00–12:00 | 1.00 | Reference |
| 4–8 h | 12:00–16:00 | 1.03 | 0.99–1.07 |
| 8–12 h | 16:00–20:00 | 1.00 | 0.96–1.04 |
| 12–16 h | 20:00–00:00 | 1.00 | 0.96–1.03 |
| 16–20 h | 00:00–04:00 | 1.05 | 1.01–1.10 |
| 20–24 h | 04:00–08:00 | 1.15 | 1.09–1.21 |

Adjusted for specimen type, panel size, age, sex, emergency department unit, triage area, admission status, day of week, season and concurrent workload. Restricted to biochemistry tubes with hemolysis as the recorded cause, the final block gave an OR of 1.14 (1.07–1.22). Because nursing shifts are 24 h long with a single 08:00 handover, hours on duty and clock time are the same variable and cannot be separated in this design. CI: Confidence interval; OR: Odds ratio.




**Appendix 5.** Hours on shift and the limit of achievable discrimination. (a) Adjusted OR for rejection with 95% CIs by 4-h block since the 08:00 handover, with the first block as reference; orange denotes intervals excluding unity. Because the roster has a single daily handover, hours on duty and clock time are the same variable. (b) Best out-of-time AUROC obtained for each feature set across twelve gradient-boosting configurations, together with the added feature blocks that failed to improve it; the dotted line marks the ceiling of 0.781.

AUROC: Area under the receiver operating characteristic curve.

**Appendix 6. Out-of-time discrimination within patient and specimen subgroups**

| Stratum | Level | N | Events | Rate, % | AUROC (95% CI) | AUPRC | Scaled brier, % | Slope | CITL |
|---|---|---|---|---|---|---|---|---|---|
| Overall | All tubes | 280,168 | 10,144 | 3.62 | 0.780 (0.775–0.785) | 0.146 | +5.66 | 0.98 | 0.24 |
| Specimen type | Biochemistry (serum) | 71,219 | 7,016 | 9.85 | 0.639 (0.632–0.646) | 0.173 | +2.42 | 1.03 | 0.21 |
| Specimen type | Other | 68,982 | 1,191 | 1.73 | 0.608 (0.593–0.622) | 0.055 | +1.67 | 1.23 | 0.56 |
| Specimen type | Complete blood count | 74,991 | 1,037 | 1.38 | 0.620 (0.602–0.636) | 0.042 | +0.53 | 1.42 | 0.26 |
| Specimen type | Urine | 23,958 | 639 | 2.67 | 0.555 (0.534–0.578) | 0.033 | -0.25 | 0.84 | 0.46 |
| Specimen type | Blood gas | 41,018 | 261 | 0.64 | 0.599 (0.573–0.626) | 0.012 | -0.11 | 1.06 | -0.48 |
| Age group | 0–1 | 15,510 | 462 | 2.98 | 0.754 (0.735–0.774) | 0.118 | +2.39 | 1.45 | 0.36 |
| Age group | 2–17 | 44,717 | 1,128 | 2.52 | 0.815 (0.803–0.827) | 0.129 | +4.96 | 1.27 | 0.27 |
| Age group | 18–44 | 92,878 | 3,042 | 3.28 | 0.762 (0.752–0.771) | 0.126 | +4.61 | 0.94 | 0.20 |
| Age group | 45–64 | 64,498 | 2,602 | 4.03 | 0.776 (0.767–0.785) | 0.148 | +5.91 | 0.95 | 0.23 |
| Age group | 65+ | 62,565 | 2,910 | 4.65 | 0.768 (0.756–0.778) | 0.168 | +6.86 | 0.95 | 0.25 |
| Triage area | Red area | 37,509 | 1,892 | 5.04 | 0.762 (0.749–0.774) | 0.193 | +7.79 | 0.93 | 0.36 |
| Triage area | Yellow area | 187,698 | 7,103 | 3.78 | 0.773 (0.768–0.779) | 0.139 | +5.40 | 0.96 | 0.22 |
| Triage area | Black area | 626 | 29 | 4.63 | 0.663 (0.571–0.749) | 0.087 | +0.26 | 0.53 | 0.43 |
| Triage area | Green area | 53,530 | 1,120 | 2.09 | 0.771 (0.758–0.785) | 0.066 | +2.07 | 1.18 | 0.16 |
| Patient type | Outpatient | 231,119 | 8,039 | 3.48 | 0.783 (0.778–0.789) | 0.142 | +5.49 | 1.02 | 0.24 |
| Patient type | Inpatient | 49,049 | 2,105 | 4.29 | 0.759 (0.748–0.770) | 0.159 | +6.20 | 0.87 | 0.24 |
| ED unit | Obstetric | 15,997 | 315 | 1.97 | 0.730 (0.701–0.757) | 0.056 | +1.42 | 0.78 | -0.19 |
| ED unit | Adult | 205,530 | 8,306 | 4.04 | 0.773 (0.768–0.779) | 0.153 | +5.90 | 0.95 | 0.25 |
| ED unit | Pediatric 1 | 51,811 | 1,334 | 2.57 | 0.805 (0.797–0.817) | 0.122 | +4.20 | 1.32 | 0.29 |
| ED unit | Pediatric 2 | 2,489 | 65 | 2.61 | 0.787 (0.737–0.840) | 0.115 | +3.95 | 1.33 | 0.22 |
| ED unit | Pediatric 3 | 4,281 | 122 | 2.85 | 0.843 (0.811–0.873) | 0.154 | +6.39 | 1.35 | 0.25 |

Evaluation of the single primary LightGBM model within each stratum of the 2026 test set. AUPRC: Area under the precision–recall curve; AUROC: Area under the receiver operating characteristic curve; CI: Confidence interval; CITL: Calibration-in-the-large.

**Appendix 7. Settings of the re-run from the raw LIS export**

| Component | Setting |
|---|---|
| **Data** | |
| Source | Four LIS exports (10,283,214 test rows), aggregated to 1,190,944 tubes by collection barcode |
| Split | Development: request before 1 January 2026 (910,776 tubes); test: 2026 (280,168 tubes); final 10% of development tubes by request time held out for early stopping |
| Specimen type | Defining test of the tube: serum indices (biochemistry), hemogram (CBC), blood gas, completeurinalysis (urine), PT/INR, APTT, fibrinogen or D-dimer (coagulation); otherwise other |
| Missing data | Triage area missing for 1,278 tubes, kept as a separate level; no imputation |
| **LightGBM** | |
| Predictors | 22: age, sex, unit, pediatric unit, triage area, admission, hour (sine, cosine), hours since handover, day of week, weekend, season (sine, cosine), calendar time, clock-time shift, three workload variables, tubes drawn simultaneously, interval since first specimen, specimen type, panel size |
| Search | Twelve configurations: 15, 63, 127 or 255 leaves × learning rate 0.015, 0.03 or 0.05; selection by early-stopping-set AUROC; test-set AUROC across configurations 0.776–0.784 |
| Selected | 15 leaves; learning rate 0.05; 400 trees (early stopping, patience 100) |
| Fixed | min_child_samples 100; feature_fraction 0.8; bagging_fraction 0.8 (every iteration); lambda_l2 1.0; max_bin 255; native categorical handling; deterministic mode; seed 2026 |
| **Comparators** | |
| XGBoost | hist; max_depth 6; learning rate 0.05; subsample 0.8; colsample_bytree 0.8; min_child_weight 10; lambda 1.0; early stopping; native categorical handling; seed 2026 |
| Random forest | 300 trees; min_samples_leaf 100; max_features sqrt; max_samples 0.5; one-hot categorical encoding; seed 2026 |
| Logistic regression | L2 penalty (C = 1.0); natural cubic spline for age (4 df); log-transformed workload and panel size; standardized indicator and continuous terms |
| **Interpretation and inference** | |
| SHAP | shap TreeExplainer, tree-path-dependent, raw (log-odds) output; 20,000 test tubes sampled with seed 2026; maximum additivity error $2\times10^{-14}$ |
| Confidence intervals | AUROC: 400 bootstrap replicates resampling presentations (seed 2026); ORs: cluster-robust standard errors by presentation |
| Software | Python 3.13, LightGBM 4.7.0, XGBoost 3.4.1, scikit-learn 1.9.1, statsmodels 0.15.0, shap 0.52.0 |

APTT: Activated partial thromboplastin time; AUROC: Area under the receiver operating characteristic curve; CBC: Complete blood count; LIS: Laboratory information system; OR: Odds ratio; PT/INR: Prothrombin time/international normalized ratio; SHAP: Shapley additive explanations.

**Appendix 8. Results of the re-run for the primary outcome and the specimen-quality outcome**

| | Any rejection | Specimen-quality rejection |
|---|---|---|
| Discrimination and calibration (LightGBM) | | |
| Test-set events, n (%) | 10,144 (3.62) | 8,128 (2.90) |
| AUROC (95% CI) | 0.783 (0.777–0.787) | 0.838 (0.834–0.842) |
| AUROC, specimen type alone | 0.737 | 0.797 |
| AUPRC | 0.141 | 0.140 |
| Calibration slope; CITL | 0.83; 0.20 | 1.03; 0.00 |
| Other algorithms, AUROC (95% CI) | | |
| XGBoost | 0.783 (0.778–0.787) | – |
| Random forest | 0.787 (0.782–0.792) | – |
| Penalized logistic regression | 0.766 (0.761–0.771) | – |
| Adjusted OR (95% CI) | | |
| Biochemistry vs other specimens | 13.41 (12.75–14.11) | 20.87 (19.60–22.23) |
| Pediatric unit (unadjusted OR) | 1.04 (0.97–1.12) (0.82) | 1.03 (0.96–1.10) (0.80) |
| Obstetric unit (unadjusted OR) | 1.02 (0.95–1.10) (0.65) | 0.95 (0.88–1.02) (0.58) |
| Concurrent workload, per SD | 1.04 (1.02–1.05) | 0.98 (0.97–1.00) |
| Weekend | 1.00 (0.97–1.02) | 0.98 (0.96–1.01) |
| Hours 20–24 vs 0–4 after handover | 1.15 (1.09–1.21) | 1.11 (1.05–1.18) |

Specimen-quality rejection treats tubes rejected only for administrative reasons as non-events. Specimen type was assigned from the defining test of each tube (Appendix 7), so the reference category of other specimens is narrower than in the original analysis and the biochemistry OR is correspondingly larger. Unit, workload and weekend ORs are from the model of Table 1; the ORs for hours 20–24 are from the 4-h block model of Appendix 4, which additionally adjusts for panel size, the six-level unit variable and day of week. In biochemistry tubes with hemolysis as the cause, the OR for hours 20–24 was 1.13 (1.07–1.21). AUPRC: Area under the precision–recall curve; AUROC: Area under the receiver operating characteristic curve; CI: Confidence interval; CITL: Calibration-in-the-large; OR: Odds ratio; SD: Standard deviation.