

Laboratory test-ordering patterns carry an early prognostic signal in electronic health records

Shakuntala Baichoo

Shakuntala.Baichoo@uhn.ca

Univeristy Health Network <https://orcid.org/0000-0002-9335-1939>

Ali Abedi

University Health Network <https://orcid.org/0000-0002-7393-1362>

Barry Rubin

Toronto General Research Institute

Bo Wang

University of Toronto

Article

Keywords: electronic health records, laboratory test ordering, test-ordering patterns, clinical decision-making, measurement patterns, clinical risk prediction, early risk stratification, sepsis, in-hospital mortality

Posted Date: June 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10117205/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Laboratory test-ordering patterns carry an early prognostic signal in electronic health records

Shakuntala Baichoo^{1*}, Ali Abedi¹, Barry Rubin¹, Bo Wang¹

¹Peter Munk Cardiac Centre, University Health Network, 585 University Ave, Toronto, M5G 2N2, Ontario, Canada.

*Corresponding author(s). E-mail(s): shakuntala.baichoo@uhn.ca;
Contributing authors: ali.abedi@uhn.ca; barry.rubin@uhn.ca;
bo.wang@uhn.ca;

Abstract

Clinical care generates patient measurements and records physician decisions, workflows, and measurement practices. This study evaluated physician laboratory test-ordering behavior as an early prognostic signal in care. The primary analysis used 305,732 admissions from MIMIC-IV, a publicly available hospital electronic health records database. Laboratory and vital-sign value features were extracted from the first 48 hours, together with test-ordering pattern features capturing frequency, intensity, breadth, and temporal escalation. These features were used to train machine-learning models based on measured values, ordering patterns, and their combination. The ordering-pattern model, containing no measured laboratory or vital-sign values, predicted adverse outcomes with AUROC values approaching those of the laboratory and vital-sign values model, although measured values retained a small, statistically significant advantage. For mortality, AUROC was 0.905 versus 0.932; for sepsis, 0.912 versus 0.924. Ordering patterns ranked among the most informative data streams, identified a high-risk “deceptively normal” phenotype, provided information beyond severity scores, and remained informative across subgroup, dose-response, and sensitivity analyses. Within four hours of admission, when many laboratory results remained unavailable, test-ordering patterns approached value-based predictive performance. External validation in the eICU-CRD and MC-MED datasets suggested generalizability despite domain shift. Laboratory test-ordering patterns may support earlier risk stratification if prospectively validated.

Keywords: electronic health records, laboratory test ordering, test-ordering patterns, clinical decision-making, measurement patterns, clinical risk prediction, early risk stratification, sepsis, in-hospital mortality

1 Introduction

Clinical care generates not only measurements of patient physiology, but also a continuous record of clinical decisions. Electronic health records (EHRs) are widely used for observational research, clinical prediction, quality improvement, and decision support because they capture laboratory results, vital signs, diagnoses, medications, procedures, orders, timestamps, and clinical documentation [1, 2]. However, these data are not neutral observations of disease alone. They are produced through clinical judgment, workflows, institutional routines, documentation practices, order-entry systems, and decisions about what to measure, repeat, or omit [1, 3]. Laboratory data illustrate this dual nature: each recorded result reflects a biological measurement, while its presence, absence, timing, and frequency also reflect care-delivery decisions [3, 4]. Therefore, routinely collected laboratory data may contain both physiological information and process information about how clinicians investigate, monitor, and manage patients.

Laboratory test ordering is one of the most frequent, structured, and clinically meaningful decisions recorded in hospital EHRs. Diagnostic tests are ordered to establish or exclude diagnoses, monitor clinical trajectories, assess treatment response, detect complications, and manage diagnostic uncertainty [5, 6]. The decision to order or repeat a test can therefore reflect a clinician’s evolving assessment of risk, sometimes before the result is available. At the same time, laboratory ordering is influenced by local norms, electronic ordering systems, defensive practice, clinical workload, test availability, standing orders, and concerns about missing clinically important disease [5–8]. Laboratory stewardship research has focused on overuse, duplicate testing, low-yield testing, and strategies to reduce unnecessary orders through education, audit and feedback, computerized order entry, and decision support [7–10]. This literature establishes test ordering as an important clinical behavior with consequences for cost, patient burden, and downstream care, but it has generally treated ordering as a target for utilization management rather than as a potential prognostic signal.

Several related lines of work suggest that clinical test ordering and measurement behavior can encode meaningful information. Clinical order patterns learned from EHR data can generate decision-support content aligned with guideline-based and real-world practice patterns, indicating that order data can reflect clinically meaningful behavior [11]. Clinical order data have also been used for data-mined order recommendation and outcome prediction, suggesting that temporal relationships among orders can encode information about patient state and care decisions [12]. Healthcare process modeling has further interpreted clinician interactions with information systems as proxies for decisions, knowledge, and expertise [13], and nursing surveillance patterns have been shown to signal clinical concern before physiological deterioration becomes fully apparent [14]. More specifically for laboratory data, laboratory-test-ordering patterns have been used to identify clinically similar intensive care unit subpopulations [15], and the presence or absence of laboratory measurements has been shown to predict mortality [16]. Broader work on informative presence and observation has emphasized that the presence, absence, and timing of observations can be informative because data collection is driven by patient needs, clinician decisions, and site-level practice [17, 18]. Recent work on informatively collected laboratory data has further

shown that accounting for non-random laboratory measurement patterns can improve clinical prediction models, supporting the view that laboratory collection itself carries information about clinical concern [19]. Methodological studies of missingness features also suggest that indicators of whether data were observed can improve prediction for clinically relevant outcomes, although such approaches generally do not model detailed ordering intensity, breadth, and escalation patterns [20].

Previous work has not systematically evaluated physician laboratory test-ordering behavior as an independent prognostic signal across the full sequence of evidence needed for clinical prediction. Prior studies have shown that orders, measurement presence, missingness, and observation patterns are informative, but they have generally focused on order recommendation, clustering, missing-data handling, laboratory stewardship, or informative observation. They have not directly compared the impact of physician laboratory ordering metadata with actual measured laboratory and vital-sign values across multiple adverse clinical outcomes, tested whether ordering patterns identify high-risk patients whose measured values appear reassuring, evaluated whether ordering-pattern models remain informative before many laboratory results return, or externally validated the signal across independent health systems. This gap is clinically important because early hospital risk assessment often occurs under uncertainty. At admission, triage, or early inpatient evaluation, clinicians must make decisions while symptoms are evolving, prior information is incomplete, and laboratory results may still be pending [6, 21, 22]. In this setting, ordering tests may precede the availability of measured values used by conventional risk models, many of which rely on vital signs and laboratory results to identify patients at risk of deterioration [3]. A clinician who orders repeated creatinine measurements, serial lactate testing, or broader laboratory panels may be responding to bedside cues, comorbidity risk, treatment response, diagnostic uncertainty, or concern for deterioration before abnormal results appear in the record [6, 22]. Conversely, reduced ordering may reflect clinical stability, therapeutic de-escalation, or a judgment that further testing is unlikely to change management [5, 22]. These patterns make ordering behavior clinically interpretable but methodologically challenging. A rigorous evaluation must therefore distinguish ordering signal from measured physiology, baseline comorbidity, demographic factors, illness severity, institutional ordering culture, workflow timing, and time-dependent availability of test results [1, 3].

This study evaluates whether physician laboratory test-ordering behavior, captured through routine measurement-count metadata in EHRs, provides an early and independent prognostic signal in hospital care (Figure 1). The primary analysis uses the Medical Information Mart for Intensive Care IV (MIMIC-IV), a publicly available EHR database from Beth Israel Deaconess Medical Center [2]. Laboratory and vital-sign measurements were extracted from the first 48 hours of admission, together with corresponding ordering features that quantify how often tests were ordered, how broadly testing was performed, and whether ordering intensity escalated over time. Five related research questions were addressed: (1) whether ordering-pattern models without measured laboratory or vital-sign values predict adverse hospital outcomes with performance close to models based on measured laboratory and vital-sign values; (2) whether ordering behavior contributes meaningful prognostic information

relative to other clinical data streams; (3) whether ordering metadata identifies a high-risk “deceptively normal” phenotype among patients with relatively reassuring measured values, whose clinical risk appears higher than their laboratory values alone would suggest; (4) whether ordering features are clinically interpretable and robust to alternative explanations, including conventional severity scores, comorbidity, demographic subgroup, admission timing, dose-response patterns, calibration, and hospital practice variation; and (5) whether ordering-pattern models without measured laboratory or vital-sign values are informative during the earliest admission window, before many laboratory results are available, and can be generalized across independent care settings. External validation was performed using the eICU Collaborative Research Database (eICU-CRD) [23] and Multimodal Clinical Monitoring in the Emergency Department (MC-MED) [24]. Together, these analyses test whether laboratory test-ordering and measurement patterns are interpretable and potentially clinically useful signals already embedded in routine EHR data.

2 Results

The analyses were organized to test whether physician laboratory test-ordering behavior is predictive, clinically interpretable, independent of measured values, robust to alternative explanations, and generalizable across care settings. First, study cohorts and feature sets were defined. Second, machine learning models based on laboratory and vital-sign value features, test-ordering pattern features, and their combination were compared. Third, the contribution of test-ordering patterns was examined relative to other clinical data streams. Fourth, a high-risk phenotype with apparently normal laboratory and vital-sign values was evaluated. Fifth, clinical interpretability and robustness were assessed through severity-score, trajectory, subgroup, dose-response, calibration, and hospital practice variation analyses. Finally, early prediction of adverse outcomes and external validation were tested in independent datasets.

2.1 Study cohorts and analytical framework

Three datasets were analyzed. MIMIC-IV [2] served as the primary cohort for machine learning model development and internal validation. After applying the prespecified eligibility criteria described in the Methods, the MIMIC-IV cohort comprised 305,732 admissions from 138,113 unique patients, corresponding to 70.9% retention of the full database; the remaining admissions were excluded because they did not meet eligibility criteria. Median patient age was 61 years (interquartile range (IQR), 47–74), and 52.1% were female. Outcome prevalences were: in-hospital mortality 2.3% ($N = 7,009$), 30-day readmission 21.6% ($N = 65,918$), acute kidney injury (AKI) 27.7% ($N = 84,817$), and sepsis 8.6% ($N = 26,411$), consistent with published rates for academic medical centers [25, 26]. The median number of distinct laboratory tests ordered per admission was 20 (IQR, 16–24), and the median total number of recorded laboratory and vital-sign measurements was 37 (IQR, 22–52) over the first 48 hours after hospital admission.

Two additional datasets supported external validation. eICU-CRD [23], a multi-site intensive care unit (ICU) dataset, included 108,324 ICU stays across 206 hospitals

(restricted from the full cohort of 200,859 stays to those with the laboratory and vital-sign streams shared with MIMIC-IV) and was used for zero-shot external validation, with models trained exclusively on MIMIC-IV and applied without retraining. The full cohort was used for the dose-response analysis, while the instrumental-variable analysis used a restricted subset of 200,430 stays across 188 hospitals. MC-MED [24], an emergency department (ED) dataset, included 118,385 ED visits and supported cross-context replication, with models trained de novo and evaluated on an unseen held-out test set of 11,792 visits. Compared with the MIMIC-IV inpatient cohort, eICU-CRD represented a higher-acuity ICU setting, while MC-MED represented an ED setting with inpatient admission and ICU admission as the principal outcomes. Full cohort characteristics for all three datasets, including demographics, comorbidities, and outcome prevalences, are provided in Supplementary Table 1.

Across these cohorts, the analyses used two parallel representations of routine clinical data: laboratory and vital-sign value features, and test-ordering pattern features derived from measurement-count metadata. This design allowed direct testing of whether the act of ordering and repeating tests contains prognostic information beyond the measured results themselves.

2.2 Test-ordering patterns predict adverse outcomes comparably to measured clinical values

To test whether test-ordering pattern features carry prognostic information, three Light Gradient Boosting Machine (LightGBM) models were trained on the MIMIC-IV training set and evaluated on the held-out test set: one model trained on laboratory and vital-sign value features, one model trained on test-ordering pattern features, and one combined model trained on both feature sets. All three models also included static baseline features, including demographic, administrative, comorbidity, and medication-class variables, so that comparisons focused on the added prognostic information from measured values, test-ordering patterns, or their combination.

The test-ordering patterns model, which included ordering-pattern and static baseline features but no measured laboratory or vital-sign values, achieved area under the receiver operating characteristic curve (AUROC) values numerically close to the laboratory and vital-sign values model across all four outcomes in MIMIC-IV (Table 1 (a)). For in-hospital mortality, AUROC was 0.905 (95% confidence interval (CI), 0.896–0.913) for the test-ordering patterns model, compared with 0.932 (95% CI, 0.925–0.939) for the laboratory and vital-sign values model, an absolute difference of -0.027 (paired bootstrap $p < 0.001$). For sepsis, AUROC was 0.912 (95% CI, 0.907–0.917) for the test-ordering patterns model, compared with 0.924 (95% CI, 0.919–0.928) for the laboratory and vital-sign values model, an absolute difference of -0.012 ($p < 0.001$).

Measured values retained a small but statistically significant advantage over test-ordering patterns (paired bootstrap $p < 0.001$ for mortality, AKI, and sepsis; $p = 0.013$ for readmission), with absolute AUROC differences ≤ 0.05 ; the ordering model thus recovered most of the values-model discrimination using only measurement metadata. Area under the precision–recall curve (AUPRC) showed a similar pattern for 30-day readmission and sepsis, while larger gaps were observed for mortality and AKI

(Table 1 (b)). These AUPRC differences indicate that ordering-pattern models may be less reliable than value-based models for identifying high-risk patients for some imbalanced outcomes. The combined model did not consistently improve performance over the single-modality models, suggesting substantial overlap between measured values and ordering patterns, with limited incremental gain in some outcomes. Supplementary Table 2 compares LightGBM with alternative machine learning models, demonstrating that the predictive signal from test-ordering patterns is not specific to a single modeling approach. To determine whether test-ordering patterns were a major clinical data stream rather than a marginal correlate, modality selection was framed as a cooperative game with four players: static baseline features, laboratory values, vital-sign values, and test-ordering patterns. Because each modality could either be included or excluded from a model, four modalities yield $2^4 = 16$ possible modality combinations; the empty combination was excluded, leaving $2^4 - 1 = 15$ non-empty coalitions. Across all 15 non-empty coalitions for each outcome, 60 LightGBM models were trained and Shapley values [27] were computed in closed form to attribute marginal AUROC contribution to each modality (Table 2). Test-ordering patterns ranked second for acute kidney injury ($\phi = 0.101$, behind laboratory values $\phi = 0.164$) and third for in-hospital mortality ($\phi = 0.108$, behind laboratory values $\phi = 0.138$ and static baseline features $\phi = 0.114$) and for sepsis ($\phi = 0.110$, behind laboratory values $\phi = 0.121$ and static baseline features $\phi = 0.112$). Vital-sign values consistently ranked last across all four outcomes ($\phi \leq 0.082$). These results suggest that test-ordering patterns were one of the most informative clinical data streams in this predictive modeling framework.

2.3 Test-ordering patterns identify high-risk patients despite relatively reassuring measured values

Having established that test-ordering patterns were predictive at the population level, the next analysis examined where this signal was most clinically informative. Admissions were stratified along two axes: the degree of abnormality in laboratory and vital-sign values, and the intensity of test-ordering patterns. Value abnormality was defined as the mean absolute z -score across all laboratory and vital-sign features. Ordering intensity was defined as the average fraction of 4-hour time bins with at least one recorded measurement, computed across all 46 measurement channels. Both axes were divided into quartiles, with Q1 representing the lowest 25% and Q4 the highest 25%. Four phenotype groups were defined by extreme quartile membership, with the principal comparison between deceptively normal patients (Q1 value abnormality and Q4 ordering intensity) and concordant normal patients (Q1 value abnormality and Q1 ordering intensity).

The deceptively normal group comprised 18,450 admissions, representing 6.0% of the full MIMIC-IV cohort. Despite relatively low measured value abnormality, deceptively normal patients had substantially higher outcome event rates than concordant normal patients: in-hospital mortality 1.64% versus 0.08%, AKI 31.8% versus 3.1%, and sepsis 16.5% versus 0.5%. The mortality difference was observed in the unmatched comparison but could not be reliably estimated after matching because of the low event count; therefore, this finding should be interpreted with caution.

To test whether these outcome differences reflected demographic or comorbidity differences rather than ordering behavior, 1:1 propensity-score matching [28] was performed using age, sex, and Charlson Comorbidity Index [29]. Matching yielded 699 pairs, corresponding to a 3.8% match rate. This low match rate indicates limited covariate overlap between groups, so the matched analysis should be interpreted as a restricted sensitivity analysis rather than a fully generalizable estimate. After matching, outcome differences remained substantial, with matched odds ratios of 7.60 (95% CI, 5.41–10.68; $p < 0.0001$) for AKI and 23.70 (95% CI, 10.36–54.21; $p < 0.0001$) for sepsis. Mortality was not analyzed in the matched cohort owing to the low event count among the 699 matched pairs, which precluded reliable estimation of the odds ratio.

TreeSHAP [30] was applied to the combined model to quantify how each feature group contributed to individual predictions. The proportional importance of each feature group was computed as the sum of its absolute Shapley values divided by the total absolute SHAP magnitude across all features. For sepsis prediction in concordant normal patients, test-ordering features contributed 17.8% of total absolute SHAP magnitude, compared with 42.1% for laboratory values and 40.1% for static baseline features. For deceptively normal patients, test-ordering features contributed 19.7% (laboratory values 53.8%, static baseline 26.5%). These findings indicate that ordering information contributed substantially to model predictions in patients whose measured values appeared relatively reassuring (Supplementary Table 3). Thus, test-ordering patterns identify a clinically important subgroup that may appear low risk when assessed using measured values alone.

2.4 Test-ordering patterns are clinically interpretable and extend beyond conventional severity

Three complementary analyses evaluated whether test-ordering patterns reflect clinically interpretable concern and workflow responses rather than arbitrary measurement frequency. These analyses tested whether ordering features extend beyond standard severity scores, remain informative among patients with lower measured value abnormality, and identify outcome-specific sentinel ordering patterns.

First, to assess whether the predictive value of test ordering reflects illness severity already captured by standard severity scores, three models were compared on the MIMIC-IV test set: a simplified Sequential Organ Failure Assessment (SOFA) score model [31], an ordering-escalation model using temporal trends in test ordering, and a combined SOFA-plus-ordering-escalation model. The ordering-escalation model alone exceeded the SOFA-only model in AUROC for in-hospital mortality (0.858 versus 0.820; difference +0.038; $p < 0.001$), sepsis (0.862 versus 0.830; difference +0.032; $p < 0.001$), and AKI (0.775 versus 0.737; difference +0.038; $p < 0.001$). Adding ordering-escalation features to SOFA increased AUROC to 0.861 for mortality, 0.864 for sepsis, and 0.781 for AKI, corresponding to gains over SOFA alone of +0.041 (mortality), +0.034 (sepsis), and +0.044 (AKI), respectively (Figure 2 (a)). For sepsis, the combined model (0.864) was essentially identical to the ordering-escalation-only model (0.862), indicating that ordering-escalation already captured the SOFA signal. These results indicate that ordering features capture prognostic information beyond what is encoded in standard organ-dysfunction severity scoring.

Second, among admissions with lower measured value abnormality, patients were stratified by ordering-escalation quartile. Mortality varied non-monotonically across quartiles among patients whose laboratory and vital-sign values appeared relatively reassuring: the most de-escalating quartile (Q1) showed the highest in-hospital mortality at 1.36%, the most escalating quartile (Q4) showed 0.84%, and the two intermediate quartiles showed the lowest mortality (Q2: 0.58%, Q3: 0.45%). This U-shaped pattern suggests that ordering trajectory carries mortality signal even among patients with lower measured value abnormality, with both withdrawal and escalation marking risk. Declining test ordering may reflect therapeutic withdrawal or transitions to palliative care in patients with deteriorating clinical status, while increased ordering may reflect physician concern triggered by bedside signs of deterioration not yet apparent in laboratory measurements (Figure 2 (b)).

Third, sentinel ordering patterns were defined as individual tests whose ordering frequency carried a strong univariate prognostic signal for a specific outcome, irrespective of the measured values returned. Each test’s total ordering count over the observation window was evaluated as a univariate predictor for each of the four outcomes. Frequent heart-rate measurement showed a 2.6-fold higher mortality rate in the top ordering quartile compared with the bottom, and frequent respiratory-rate measurement a 2.8-fold enrichment for sepsis; pH (mortality) and lactate (sepsis) ordering also identified elevated risk (~ 2.3 -fold). For AKI, the strongest sentinel tests were creatinine and blood urea nitrogen (BUN), which define the AKI outcome and are therefore not interpreted as independent signals. Together, these analyses suggest that test-ordering patterns may reflect clinically interpretable concern and workflow responses rather than nonspecific data-density effects alone.

2.5 Test-ordering associations are robust across subgroups, timing, and hospital practice variation

The next set of analyses challenged the ordering signal against alternative explanations. Specifically, analyses tested whether the signal was explained by pre-existing comorbidity, demographic subgroup structure, admission timing, model calibration artifacts, dose-response behavior, or hospital-level ordering culture.

To test whether the ordering signal was explained by documented chronic disease burden, the test-set analysis was restricted to patients with no documented comorbidities (Charlson Comorbidity Index = 0; $N = 38,627$ for mortality, AKI, and sepsis, comprising 84.3% of the held-out test set). Within this comorbidity-free stratum, the test-ordering patterns model achieved AUROCs of 0.912 for in-hospital mortality, 0.695 for 30-day readmission, 0.836 for AKI, and 0.918 for sepsis, essentially identical to the corresponding overall test-set values. The test-ordering patterns model also remained close to the laboratory and vital-sign values model across all four outcomes, including 30-day readmission (test-ordering AUROC 0.695). These results suggest that the ordering signal was not eliminated after restricting to admissions without documented Charlson comorbidities, although residual or under-coded comorbidity may remain (Supplementary Table 4).

To assess potential disparities in predictive performance, AUROC for the test-ordering patterns model was evaluated separately within racial and insurance categories. For sepsis prediction, AUROC was tightly consistent across racial groups: 0.909 for White patients, 0.906 for Black patients, 0.910 for Hispanic patients, 0.915 for Asian patients, 0.928 for Other race, and 0.915 for Unknown race, spanning a 0.022 AUROC range. Native American patients were excluded from this comparison due to small sample size, with $N = 12$ sepsis cases. For in-hospital mortality, AUROC ranged from 0.888 in White patients to 0.928 in patients of Other race, a 0.040 range, with wider confidence intervals in smaller groups (for example, only 29 mortality events in Hispanic patients). Insurance type showed similarly modest variation, with AUROC ranges ≤ 0.023 for sepsis. These subgroup analyses suggest that discrimination was broadly similar across major demographic and insurance strata, although some estimates were limited by small event counts (Supplementary Table 5).

To test whether ordering behavior reflected structural patterns related to when patients were admitted, AUROC for the test-ordering patterns model was compared between night and day admissions and between weekday and weekend admissions. Across all four outcomes, the AUROC difference between night and day admissions was small, with the largest difference observed for 30-day readmission (difference -0.018, 95% CI -0.034 to -0.006; $p = 0.005$). For sepsis prediction, AUROC was numerically higher for night admissions than for day admissions (0.917 versus 0.909; difference +0.008, 95% CI -0.004 to +0.017; $p = 0.200$), but this difference was not statistically significant. Weekend versus weekday differences were similarly small, with the largest gap of 0.018 AUROC for in-hospital mortality and overlapping CIs across all outcomes. These results suggest that the ordering signal was broadly stable across admission timing strata and was not primarily explained by broad staffing-pattern differences (Supplementary Table 6).

Because clinical deployment also requires reliable uncertainty estimates, split conformal prediction [32] was applied to the combined model. For in-hospital mortality, coverage varied systematically with ordering intensity: patients in the highest-ordering quartile had only 74.1% coverage against a 90% target, while patients in the lowest-ordering quartile were over-covered at 99.2%. Mondrian conformal prediction, which performs separate calibration within each ordering-intensity quartile, restored coverage to within the target range across all four subgroups (89.8–90.3%; Supplementary Table 7). These findings indicate that standard calibration may be inadequate in high-ordering patients and that subgroup-stratified calibration would be required before any clinical deployment, particularly because the highest-ordering patients may receive the least reliable uncertainty estimates under standard conformal prediction.

Dose-response analyses then examined whether outcome rates rose with ordering intensity across independent datasets. In MIMIC-IV ($N = 305,732$), outcome rates rose strictly monotonically with ordering intensity. Mortality increased from 0.19% in Q1 to 7.26% in Q4 (Q2: 0.53%, Q3: 1.26%); sepsis increased from 1.01% to 26.45% (Q2: 2.62%, Q3: 4.45%); and AKI increased from 7.29% to 49.46% (Q2: 22.55%, Q3: 31.58%). Individual-level Spearman rank correlations (rank-biserial form for binary outcomes) confirmed positive associations ($\rho = 0.181$ for mortality, $\rho = 0.344$ for sepsis, and $\rho = 0.354$ for AKI; all $p < 0.001$). In eICU-CRD ($N = 200,859$), the

overall gradient was replicated, although mortality was elevated in the lowest-ordering quartile (Q1: 5.18%, Q2: 3.00%, Q3: 3.64%, Q4: 9.99%; $\rho = 0.077$). This pattern may reflect palliative or comfort-care decisions concentrating in the lowest-ordering quartile. In MC-MED ($N = 118,385$), inpatient admission rates similarly showed a U-shaped pattern, with the lowest-ordering group elevated relative to the middle group and the highest-ordering group highest overall (G1: 20.9%, G2: 11.3%, G3: 57.5%; $\rho = 0.231$, $p < 0.001$). Across all three datasets, ordering intensity was associated with downstream outcome risk, although the direction and shape of the association varied by care context, with expected deviations in care contexts where de-escalation or zero-order floor effects were prominent (Figure 3).

To assess whether the observed association between ordering intensity and adverse outcomes was biased by unmeasured illness severity, an instrumental-variable analysis using two-stage least squares (2SLS) regression was applied within eICU-CRD ($N = 200,430$ ICU stays across 188 hospitals). The instrument was constructed as the leave-one-out mean ordering intensity at each patient’s hospital, capturing institutional ordering culture, specialty mix, and protocol-level practice variation. Instrument strength was assessed using the first-stage partial F-statistic, which was high ($F = 26,894$), indicating a strong association between the instrument and patient-level ordering intensity. The 2SLS estimate of the ordering-outcome association was $\hat{\beta}_{2SLS} = 0.035$ (95% CI, 0.032–0.038; $p < 0.001$), statistically indistinguishable from the corresponding ordinary least squares (OLS) estimate $\hat{\beta}_{OLS} = 0.032$ (95% CI, 0.031–0.034). A complementary hospital-level analysis showed a strong positive correlation between institutional ordering culture and hospital mortality rate across the 188 hospitals (Spearman $\rho = 0.560$, $p < 0.001$). These sensitivity analyses provide partial reassurance that the ordering signal is not driven by a single measured demographic, workflow, comorbidity, or hospital-practice factor, although residual confounding and violations of instrumental-variable assumptions cannot be excluded.

Because creatinine measurements partly define acute kidney injury and antimicrobial administration partly defines sepsis, we tested whether the ordering signal was an artifact of this overlap by re-training the models after removing the outcome-defining inputs from the features (Supplementary Table 8). Removing the creatinine measurement-count channel changed the AKI ordering-model AUROC by only -0.0015 (0.835 to 0.833). For sepsis, the only label-overlapping input was a static antimicrobial-medication indicator shared by all three models; blood-culture collection is recorded in the microbiology table and is absent from the laboratory ordering features. Removing the antimicrobial indicators lowered the ordering model from 0.912 to 0.897 and the laboratory and vital-sign values model from 0.924 to 0.915. After these exclusions, the ordering model retained strong discrimination and remained close to the values model, indicating that the AKI and sepsis signals are not artifacts of outcome ascertainment.

Because some in-hospital deaths occur within the 48-hour feature window, we repeated the mortality analysis restricted to patients still alive at a 48-hour landmark; discrimination was essentially unchanged (laboratory and vital-sign values model 0.932 to 0.927; test-ordering patterns model 0.905 to 0.899), indicating that the signal is not driven by truncated measurement records near death.

We next asked whether the three coarse per-channel summaries (total count, fraction of bins with a measurement, and linear trend) discarded temporal information that a more expressive model could exploit, such as non-linear escalation, bursting, or de-escalation. A gated-recurrent-unit (GRU) network trained on the raw per-4-hour-bin measurement-count sequences, with the same static features, did not meaningfully exceed the gradient-boosted summary model on any outcome (mean paired AUROC difference ≤ 0.0062 in magnitude, per-seed maximum 0.0072, across five random initializations; Supplementary Table 9). The ordering signal, therefore, saturates at simple frequency summaries, and a sequence model provided no additional discrimination, supporting the use of an interpretable gradient-boosted tree model as the primary analysis.

2.6 Test-ordering patterns enable early prediction and show generalizability across datasets

The final set of analyses evaluated whether test-ordering patterns are clinically useful in early prediction and whether the signal generalizes beyond MIMIC-IV. To assess early triage utility, the test-ordering patterns model and the laboratory and vital-sign values model were evaluated using only data available during the first 4 hours after admission. At this early timepoint, only 26.6% of admissions had any laboratory result recorded in the EHR. Despite this limited availability of measured laboratory values, the test-ordering patterns model achieved AUROC of 0.819 (95% CI 0.806–0.832) for in-hospital mortality, approaching the laboratory and vital-sign values model’s AUROC of 0.831 (0.818–0.844). For sepsis, the test-ordering patterns model achieved AUROC of 0.780 (0.771–0.789), compared with 0.796 (0.788–0.804) for the laboratory and vital-sign values model. At this 4-hour timepoint the ordering model remained informative but trailed measured values by a small margin (≈ 0.012 – 0.016 AUROC). These results suggest that test-ordering patterns observed within the first hours of admission encode prognostic information similar to the laboratory and vital-sign data available during the same time window, supporting early triage decision support before many laboratory results are available for review (Figure 4).

To test cross-system transfer under domain shift, LightGBM models trained exclusively on MIMIC-IV, using only the 30 laboratory tests and 8 vital-sign streams shared with eICU-CRD, were applied to eICU-CRD without retraining or fine-tuning. On eICU-CRD, the test-ordering patterns model achieved AUROCs of 0.715 for in-hospital mortality, 0.651 for AKI, and 0.719 for sepsis, while the laboratory and vital-sign values model achieved corresponding AUROCs of 0.833, 0.838, and 0.801. The laboratory and vital-sign values model outperformed the test-ordering patterns model for all three outcomes, with AUROC differences of $+0.118$ for mortality (95% CI $+0.113$ to $+0.123$; $p < 0.001$), $+0.187$ for AKI (95% CI $+0.183$ to $+0.192$; $p < 0.001$), and $+0.083$ for sepsis (95% CI $+0.080$ to $+0.086$; $p < 0.001$). Both feature types showed substantial attenuation relative to MIMIC-IV internal performance on the same shared feature subset (Figure 5 (a)), and the ordering-pattern model showed larger performance gaps than the value-based model for some outcomes, particularly AKI, indicating that both ordering metadata and measured values are affected by domain shift. Per-hospital analysis, restricted to hospitals with at least 30 evaluable

cases and 3 positive cases per outcome, yielded median test-ordering patterns AUROC of 0.732 for mortality, 0.643 for AKI, and 0.713 for sepsis across evaluable hospitals (Figure 5 (a–b)).

To test whether ordering-pattern features could be replicated in a different care context, four LightGBM models were trained de novo on MC-MED and evaluated on the held-out test set. The four models used triage vital-sign features, ED laboratory values, ED test-ordering features, and all features combined. For inpatient admission, the ED test-ordering model achieved AUROC of 0.876 (95% CI 0.870–0.881), outperforming both the triage vital-sign model (0.832; 0.825–0.840; difference +0.044, 95% CI +0.035 to +0.052; $p < 0.001$) and the ED laboratory values model (0.783; 0.773–0.793; difference +0.093, 95% CI +0.084 to +0.102; $p < 0.001$). For ICU admission, the triage vital-sign model was the strongest single modality (0.911; 0.896–0.926), outperforming the ED test-ordering model (0.874; 0.856–0.892; difference +0.037, 95% CI +0.016 to +0.059; $p = 0.002$), consistent with ICU triage decisions being driven primarily by acute physiological derangement. However, the ED test-ordering model still outperformed the ED laboratory values model for ICU admission (0.874 versus 0.831; difference +0.043, 95% CI +0.014 to +0.072; $p = 0.002$). Combining triage vital signs and ED test-ordering features produced substantial gains over either modality alone, with AUROC reaching 0.957 for ICU admission and 0.933 for inpatient admission (Figure 5 (c–f)). These findings support test-ordering pattern features as low-burden EHR signals that remain informative across independent care settings.

3 Discussion

The central finding of this study is that physician laboratory test-ordering behavior contains an early and clinically meaningful prognostic signal in EHR data. In MIMIC-IV, and in zero-shot eICU-CRD validation and de novo MC-MED replication, test-ordering patterns predicted adverse hospital outcomes with performance close to models based on measured laboratory and vital-sign values, contributed meaningful information relative to other clinical data streams, identified high-risk patients with relatively reassuring measured values, remained informative across robustness analyses, and supported early prediction before many laboratory results were available. These findings indicate that EHR data contain not only measurements of patient physiology, but also a structured record of clinical concern, diagnostic uncertainty, monitoring intensity, and workflow practice.

This work extends prior research on informative observation, EHR missingness, clinical process data, and laboratory ordering behavior. Previous studies have shown that the presence, absence, timing, and frequency of observations in routine care are not random and may reflect patient severity, clinician behavior, and site-level practice [3, 17, 18]. Other work has used clinical orders for decision support, order recommendation, patient clustering, or laboratory stewardship [10–12, 15]. The present study builds on this literature by directly evaluating physician laboratory test-ordering behavior as an independent prognostic signal, comparing it with measured laboratory and vital-sign values, testing its contribution across multiple outcomes, examining a

high-risk lower-abnormality phenotype, and validating the signal across independent care settings.

The most clinically distinctive finding relates to the deceptively normal phenotype. These patients had relatively reassuring measured values but high ordering intensity, and they experienced substantially higher rates of mortality, AKI, and sepsis than concordant normal patients with relatively reassuring measured values and low ordering intensity. In a restricted propensity-score matched sensitivity analysis based on age, sex, and Charlson Comorbidity Index, the increased risk of AKI and sepsis persisted. This phenotype is important because it represents a subset of patients that may be missed by early warning systems driven mainly by abnormal measured values [33, 34]. In these cases, laboratory values may still appear reassuring, but the care process and clinician judgment have already signaled concern through repeated or broader testing and measurement. Careful evaluation of ordering metadata, therefore, makes an implicit bedside signal explicit and measurable.

Several analyses support the interpretation that the ordering signal is clinically meaningful rather than a statistical artifact. Ordering-escalation features added prognostic information beyond a simplified SOFA score, and ordering trajectory stratified mortality even among patients with lower measured value abnormality. The observed U-shaped pattern, in which both increasing and decreasing ordering trajectories were associated with risk, is clinically plausible. Increased ordering may reflect concern for deterioration not yet captured by measured values, whereas decreased ordering in high-risk patients may reflect therapeutic withdrawal, comfort-focused care, or a judgment that further testing is unlikely to change management. Sentinel ordering patterns further supported interpretability, as repeated ordering of specific tests mapped onto plausible outcome-specific clinical concerns.

The robustness analyses addressed several alternative explanations. The ordering signal remained strong among patients without documented comorbidities, indicating that it was not simply a proxy for chronic disease burden. Predictive performance was broadly consistent across racial, insurance, day-night, and weekday-weekend strata, suggesting that the signal was not dominated by a single demographic or workflow subgroup. Dose-response analyses showed that adverse outcome rates generally increased with ordering intensity across MIMIC-IV, eICU-CRD, and MC-MED, although deviations were observed in settings where palliative de-escalation or zero-order floor effects were expected. Hospital-level analyses provided a sensitivity analysis for severity-related confounding, although they cannot exclude residual confounding or violations of instrumental-variable assumptions.

The external validation and early-window findings are central to the practical relevance of this work. Within the first four hours of admission, when many laboratory results were not yet available, test-ordering patterns approached laboratory and vital-sign value models for mortality and sepsis prediction, with measured values retaining a small advantage. This indicates that the act of ordering tests may provide useful information before the results of those tests are recorded. Validation in eICU-CRD showed that the signal transferred across a multi-site ICU dataset, although with expected performance attenuation that may be due to differences in case mix, workflow, institutional ordering culture, and coding practices. De novo replication in MC-MED

showed that ordering-pattern features were also informative in an ED setting, particularly when combined with triage vital signs. Together, these findings suggest that ordering-pattern signals are not unique to MIMIC-IV, although performance varies across databases and care contexts and local calibration and prospective evaluation will be required before deployment.

Clinical implications. The main clinical implication is that ordering metadata could be translated into real-world decision support with minimal additional burden or cost. Most hospitals already store laboratory timestamps and measurement-event records, and some also store detailed order-entry logs, in their EHR systems. These data could be used to compute ordering intensity, test diversity, repeated-order frequency, and escalation patterns in real time during the first hours of admission. Such features could complement early warning systems, triage dashboards, sepsis surveillance tools, AKI risk models, or patient-prioritization workflows. For example, a patient with relatively reassuring early measured values but unusually high ordering activity could be flagged for closer monitoring, repeat risk assessment, or earlier senior clinical review. Importantly, ordering metadata should not replace laboratory values, vital signs, bedside assessment, or clinician judgment. Its most appropriate role is as an additional layer that captures aspects of clinical decision-making, monitoring intensity, and workflow before structured results become available. Because ordering behavior is influenced by local practice, implementation should include site-specific validation, prospective silent evaluation, workflow integration, and calibration across ordering-intensity strata.

Several strengths support the reliability of these findings. The study used a large primary cohort, multiple clinically relevant outcomes, one zero-shot external validation dataset, and one independent de novo replication dataset. Ordering features were separated from measured laboratory and vital-sign values, allowing direct comparison between clinical measurements and the decision to obtain them. The analyses also evaluated predictive performance, relative contribution across data streams, phenotype discovery, propensity-score matching, severity-score comparison, trajectory and sentinel-ordering patterns, subgroup robustness, dose-response relationships, hospital practice variation, early prediction, zero-shot external validation, and de novo cross-context replication. This breadth of evidence supports the interpretation that measurement and test-ordering patterns are clinically interpretable EHR signals, while prospective validation is needed to determine clinical utility.

Limitations. Several limitations should be noted. First, ordering features were derived from measurement event counts rather than complete physician order-entry records. Measurement frequency is informative but may reflect a mixture of physician-entered orders, nurse-initiated measurements, protocolized testing, standing orders, automatically scheduled measurements, and local documentation practices. Second, the propensity-score matched analysis of the deceptively normal phenotype achieved a low match rate, limiting statistical power for matched comparisons. Third, the SOFA score used here was a simplified version derived from available structured data. Fourth, AKI and sepsis labels relied partly on diagnostic coding and rule-based definitions, which may introduce label noise. Because AKI and sepsis ascertainment may overlap with the 48-hour feature extraction window, model performance may partly reflect recognition, documentation, or monitoring of evolving events rather

than purely prospective prediction. This concern is particularly relevant for sepsis because blood culture collection and intravenous antibiotic administration were part of the outcome definition and may overlap with ordering and medication-class features. Similarly, AKI prediction may be influenced by creatinine monitoring intensity because creatinine measurements contribute to both AKI ascertainment and ordering-pattern features. Fifth, eICU-CRD external validation involved an ICU-only cohort and therefore represents a difficult domain-shift setting rather than a direct replication of the mixed inpatient MIMIC-IV cohort; performance attenuation, especially for AKI, indicates that local recalibration or retraining may be required. Sixth, hospital-level instrumental-variable analyses addressing severity-related confounding depend on assumptions that cannot be fully verified; hospital ordering culture may be associated with case mix, staffing, protocols, and other hospital-level determinants of mortality. Finally, all analyses were retrospective, and the study does not establish that presenting ordering-based predictions to clinicians would improve patient outcomes.

Future work should evaluate ordering metadata prospectively in silent-mode deployment before testing its impact in clinical workflows. Such studies should assess discrimination, calibration, alert burden, clinician interpretability, fairness across patient groups, and incremental value beyond existing early warning systems. Future models should also distinguish physician-entered orders from protocolized or automatically scheduled measurements where detailed order-entry data are available.

In conclusion, physician laboratory test-ordering behavior is an early, reproducible, interpretable, and routinely available prognostic signal in EHR data. The findings show that when and how often tests and measurements are recorded carry information beyond the measured values themselves, identify high-risk patients with relatively reassuring measured values, and remain informative across sensitivity analyses, zero-shot external validation, and de novo cross-context replication, although performance varies across settings. If prospectively validated, ordering metadata could provide a low-burden way to improve early risk stratification using information already generated during routine clinical care.

4 Methods

The analytical workflow followed the sequence of the main study questions. First, study cohorts and outcome labels were defined. Second, parallel laboratory and vital-sign value features and test-ordering pattern features were constructed. Third, predictive models based on measured values, ordering patterns, and their combination were trained and compared. Fourth, the contribution of ordering patterns was quantified relative to other clinical data streams. Fifth, the deceptively normal phenotype was defined and evaluated. Sixth, clinical interpretability was assessed using severity-score, trajectory, and sentinel-ordering analyses. Seventh, robustness was evaluated across comorbidity, demographic subgroup, admission timing, calibration, dose-response, and hospital-level practice variation analyses. Finally, early triage performance and external validation were assessed in independent EHR datasets.

4.1 Study design, data sources, cohorts, and outcomes

This retrospective cohort study analyzed admissions and visits from three independent EHR datasets. MIMIC-IV v2.2 [2], the primary dataset, comprises 431,231 hospital admissions from 180,733 unique patients at Beth Israel Deaconess Medical Center from 2008 to 2019. eICU-CRD [23] comprises 200,859 ICU stays across 208 US hospitals from 2014 to 2015. After filtering for the laboratory and vital-sign streams shared with MIMIC-IV, the zero-shot external validation cohort included 108,324 ICU stays across 206 hospitals. MC-MED [24] comprises 118,385 adult ED visits from 70,545 patients at Stanford University Medical Center from 2020 to 2022. Together, these datasets span three distinct care settings: a single academic hospital system, a multi-site ICU network, and a single academic ED.

MIMIC-IV primary cohort. From MIMIC-IV, adult inpatient admissions were included if the patient was aged ≥ 18 years, the hospital length of stay was ≥ 24 hours, and at least one laboratory result was recorded within the first 48 hours. This yielded 305,732 admissions, corresponding to 70.9% retention, from 138,113 unique patients. Admissions were split at the subject level into 70% training, 15% validation, and 15% test sets, ensuring no patient appeared in more than one partition (the split was not stratified).

eICU-CRD external validation cohort. From eICU-CRD, equivalent age and length-of-stay criteria were applied. For the primary zero-shot external validation, stays were restricted to those with at least one of the 30 laboratory tests and 8 vital-sign streams shared with MIMIC-IV, yielding 108,324 stays across 206 hospitals. No internal split was performed. The entire eICU-CRD cohort served as a held-out external test set, with models trained exclusively on MIMIC-IV applied without retraining or fine-tuning. The dose-response analysis used the full eICU-CRD cohort of 200,859 stays. The hospital instrumental-variable analysis used 200,430 stays from 188 hospitals after hospitals with insufficient stays for stable instrument estimation were excluded. Per-hospital AUROC analyses required ≥ 30 observations and ≥ 3 positive cases, yielding 176, 179, and 172 evaluable hospitals for mortality, AKI, and sepsis, respectively, with 183 unique hospitals across tasks.

MC-MED de novo replication cohort. From MC-MED, the full cohort of 118,385 ED visits was split at the patient medical record number level into 80% training, 10% validation, and 10% test sets. This yielded a held-out test set of 11,792 visits used for all reported MC-MED AUROCs. Models were trained de novo on the MC-MED training set.

For MIMIC-IV, four binary outcomes were defined: in-hospital mortality, 30-day readmission, AKI, and sepsis. AKI was defined using International Classification of Diseases, 10th Revision (ICD-10) code N17.x, or Kidney Disease: Improving Global Outcomes (KDIGO) creatinine rise ≥ 0.3 mg/dL within 48 hours [35]. Sepsis was defined using ICD-10 code A41.x with blood culture and intravenous antibiotics within 72 hours, consistent with the Third International Consensus Definitions for Sepsis and Septic Shock [36, 37]. The eICU-CRD external validation used corresponding in-hospital mortality, AKI, and sepsis outcomes where available. In MC-MED, the evaluated outcomes were in-hospital mortality, 30-day readmission, ICU admission, and inpatient admission from the ED.

Data access for all three datasets complied with PhysioNet credentialing and data use agreement requirements. Because all records are de-identified and publicly available, additional institutional ethics approval was not required.

4.2 Feature construction

Two parallel feature sets were constructed from the first 48 hours of each admission to allow separate modeling of measured clinical values and physician test-ordering behavior. A shared set of static baseline features was included in all primary models. **Laboratory and vital-sign value features.** Laboratory and vital-sign value features comprised 184 features, defined as four summary statistics (mean, minimum, maximum, and last observed value) for each of 35 laboratory tests and 11 vital-sign streams. These features represented measured physiological values recorded during the first 48 hours.

Test-ordering pattern features. Test-ordering pattern features comprised 146 features derived exclusively from per-bin measurement event counts, recording the number of times each laboratory test or vital-sign was logged in each 4-hour bin. These features contained no measured laboratory or vital-sign values. Three per-channel ordering metrics were computed for each of the 35 laboratory tests and 11 vital-sign streams, yielding 138 features: total measurement count over 48 hours, fraction of bins with at least one measurement, and linear temporal trend. These were complemented by eight admission-level summary features computed once per admission by aggregating across all 46 measurement channels. These summary features captured the patient’s overall ordering profile, including ordering intensity, ordering diversity, ordering breadth, and ordering escalation. Because each laboratory measurement requires a corresponding clinical order, laboratory measurement counts served as a proxy for physician laboratory ordering frequency.

Static baseline features. Static baseline features comprised 70 patient-level variables, including 33 demographic and administrative features (age, sex, race, insurance, and admission type), 17 Charlson Comorbidity Index categories [29], and 20 medication class indicators. Full feature definitions are provided in the Supplementary Methods.

This feature design allowed measured clinical values and measurement-pattern metadata to be modeled separately, enabling direct evaluation of whether ordering metadata contains prognostic information independent of the laboratory and vital-sign values themselves.

4.3 Primary predictive modeling and statistical evaluation

LightGBM [38], a gradient-boosted decision tree framework, was used for all primary predictive modeling because of its strong performance on large structured tabular datasets with mixed continuous and categorical features [39–41]. Three LightGBM classifiers were trained for each MIMIC-IV outcome: a laboratory and vital-sign values model, a test-ordering patterns model, and a combined model. The laboratory and vital-sign values model included laboratory and vital-sign value features plus

static baseline features (254 features). The test-ordering patterns model included test-ordering pattern features plus static baseline features (216 features). The combined model integrated all feature sets (400 features).

Each model was trained for up to 2,000 boosting iterations with a learning rate of 0.05. Each tree was limited to a maximum of 127 leaves, and training was halted when validation AUROC no longer improved. Class imbalance was addressed by weighting positive-class admissions inversely to their prevalence in the training set. Models were trained on the MIMIC-IV training set, hyperparameters were selected on the validation set, and final performance was evaluated on the held-out test set ($N = 45,805$ admissions). Random seeds were fixed at 42 across all experiments to support reproducibility.

Model discrimination was assessed using AUROC and AUPRC. Ninety-five percent CIs were computed using 1,000 bootstrap resamples of the test set. Pairwise comparisons between models used two-sided paired bootstrap tests. All analyses followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) reporting guidelines for multivariable prediction models [42]. The MIMIC-IV train, validation, and test split was held constant across experiments using a subject-level 70%/15%/15% split, preventing admissions from the same patient from appearing in more than one split (the split was not stratified).

4.4 Contribution of test-ordering patterns relative to other clinical data streams

To quantify the marginal contribution of each clinical data modality to predictive performance, modality selection was framed as a cooperative game. Four feature modalities were treated as players: static baseline features, laboratory values, vital-sign values, and test-ordering patterns. For each outcome, LightGBM models were trained across all $2^4 - 1 = 15$ non-empty modality coalitions. Across the four outcomes, this yielded 60 coalition models.

For each modality, the cooperative game-theoretic Shapley value [27] was computed in closed form to estimate its average marginal AUROC contribution across all possible modality combinations. This analysis differs from individual-level SHAP attribution because it evaluates the contribution of entire data streams to predictive performance rather than the contribution of individual features within a fixed model. The resulting modality-level contributions were used to determine whether test-ordering patterns provided substantial predictive information relative to measured laboratory values, vital signs, and static baseline features.

4.5 Deceptively normal phenotype analysis

To evaluate whether ordering behavior identifies high-risk patients whose measured values appear reassuring, all MIMIC-IV admissions were stratified along two axes: laboratory and vital-sign value abnormality, and test-ordering intensity. Value abnormality was defined as the mean absolute z -score across all laboratory and vital-sign value features. Ordering intensity was defined as the average fraction of 4-hour time bins with at least one recorded measurement, computed across all 46 measurement

channels. Each axis was divided into quartiles. The deceptively normal phenotype was defined as Q1 value abnormality combined with Q4 ordering intensity, representing patients with relatively low measured value abnormality but high ordering activity. Concordant normal patients were defined as Q1 value abnormality combined with Q1 ordering intensity.

Outcome event rates were compared between deceptively normal and concordant normal patients. To assess whether differences reflected demographic or comorbidity imbalance, 1:1 propensity-score matching [28] was performed using logistic regression on age, sex, and Charlson Comorbidity Index. Greedy nearest-neighbor matching was performed with a caliper of 0.20 standard deviations of the propensity score. Post-matching odds ratios and CIs were estimated for outcomes with sufficient event counts.

To evaluate whether ordering features contributed disproportionately to predictions in these phenotypes, TreeSHAP [30] was applied to the combined model. For each phenotype, proportional feature-group importance was computed as the sum of absolute SHAP values within each feature group divided by the total absolute SHAP magnitude across all features. This analysis linked phenotype discovery to model-level attribution by testing whether ordering features were especially influential among patients whose laboratory and vital-sign values appeared relatively reassuring.

4.6 Clinical interpretability and severity-adjusted analyses

Three complementary analyses examined whether test-ordering patterns captured clinically interpretable information beyond measured values.

Incremental value over objective severity. To test whether ordering patterns added predictive information beyond a standard organ-dysfunction score, a simplified SOFA score [31] was constructed from available laboratory values and vital signs. Three models were compared for in-hospital mortality, AKI, and sepsis: a SOFA-only model, an ordering-escalation-only model, and a combined SOFA-plus-ordering-escalation model. The incremental AUROC contributed by ordering escalation was quantified as the difference between SOFA-only and combined model performance.

Ordering escalation in stable patients. To test whether ordering trajectory predicted outcomes even when laboratory and vital-sign values appeared reassuring, admissions with value abnormality below the cohort median were identified. Value abnormality was defined as the mean absolute z -score across all laboratory and vital-sign value features, yielding approximately half of the MIMIC-IV cohort ($N \approx 152,866$). Within this stable-value subgroup, admissions were stratified by ordering-escalation quartile, defined as the linear temporal slope of ordering frequency over the 48-hour window. Outcome event rates were compared across the four quartiles.

Sentinel ordering patterns. To identify individual tests whose ordering behavior carried strong prognostic signal, each test’s total ordering count over 48 hours was evaluated as a univariate predictor of each outcome. The count threshold maximizing sensitivity plus specificity was identified using Youden’s J statistic. Outcome enrichment between high-ordering and low-ordering strata was then used to identify clinically interpretable sentinel ordering patterns.

4.7 Robustness, subgroup, and calibration analyses

Several analyses tested whether the ordering signal was robust to patient characteristics, workflow timing, and calibration differences.

Comorbidity-free subgroup analysis. To test whether ordering signal reflected pre-existing chronic disease burden, the held-out MIMIC-IV test set was restricted to patients with no documented comorbidities, defined as Charlson Comorbidity Index = 0. The test-ordering patterns model and laboratory and vital-sign values model were evaluated in this subgroup using the same AUROC procedures as in the primary analysis.

Race and insurance subgroup analysis. To assess potential disparities in predictive performance, AUROC was evaluated separately within racial and insurance categories. Subgroups with insufficient positive events were excluded from formal comparison. Ninety-five percent bootstrap CIs were computed within each subgroup where sample size permitted.

Admission timing analysis. To assess whether the ordering signal reflected structural workflow differences related to admission timing, performance was compared between night and day admissions and between weekday and weekend admissions. AUROC differences and bootstrap CIs were used to assess whether the ordering signal was stable across timing strata.

Conformal prediction and subgroup calibration. Predictive uncertainty was quantified using split conformal prediction [32] at significance level $\alpha = 0.10$, corresponding to a target empirical coverage of 90%. Standard split conformal prediction was first applied to the combined model. Coverage was then evaluated across ordering-intensity quartiles to determine whether uncertainty estimates were equally reliable across ordering strata. Mondrian conformal prediction was subsequently applied by calibrating separate conformal predictors within each ordering-intensity quartile, restoring subgroup-specific coverage where standard split conformal prediction was miscalibrated.

4.8 Dose-response and instrumental-variable analyses

To test whether higher physician ordering intensity was associated with higher downstream outcome risk, admissions in each dataset were ranked by total ordering intensity and partitioned into ordered strata. MIMIC-IV and eICU-CRD admissions were partitioned into four equal-sized quartiles. MC-MED admissions were partitioned into three groups instead because 39.8% of ED visits had zero ordering events and could not be split further. Outcome event rates with 95% bootstrap CIs were computed within each stratum. Complementing this group-level analysis, individual-level Spearman rank correlations between ordering intensity and each binary outcome were computed on the full MIMIC-IV cohort ($N = 305,732$). Computing the correlation at the individual level avoided the artifactual perfect correlation that can arise when summary statistics are correlated across strata defined by ordering intensity.

To test whether the association between ordering intensity and in-hospital mortality could be explained by illness severity confounding rather than ordering behavior itself, an instrumental-variable analysis using 2SLS regression was performed in the

eICU-CRD cohort. The instrument was each hospital’s mean ordering intensity, calculated as the average number of laboratory orders per admission across the hospital after excluding the index patient. This leave-one-out hospital mean ordering intensity was intended to capture institutional ordering culture, specialty mix, and protocol-level practice variation, which vary across hospitals for reasons not determined by any individual patient’s severity.

OLS estimates, which may be biased in the presence of severity confounding, were compared with 2SLS estimates, which may reduce such confounding if the instrumental-variable assumptions hold. Both ordering intensity and the binary mortality outcome were standardized before estimation. Concordance between the two estimates was interpreted as a sensitivity analysis suggesting limited bias under the instrumental-variable assumptions. Instrument strength was assessed using the first-stage partial F-statistic, with values greater than 10 conventionally indicating sufficient strength. Ninety-five percent CIs were computed using 1,000 hospital-stratified bootstrap resamples.

4.9 Outcome-ascertainment sensitivity analysis

To confirm that the predictive value of test-ordering patterns for AKI and sepsis was not an artifact of the outcome definitions, the AKI and sepsis models were re-trained after excluding the outcome-defining inputs. For AKI (defined in part by a creatinine rise ≥ 0.3 mg/dL within 48 hours), the creatinine measurement-count channel was removed, and its three per-channel ordering features (total count, fraction of 4-hour bins with a measurement, and linear temporal trend) together with the admission-level ordering summaries were recomputed without that channel. For sepsis (defined in part by intravenous antibiotic administration), the antimicrobial medication-class indicators (antibiotic, antifungal, antiviral) were removed from the static feature set for all models; blood-culture collection is recorded in the microbiology table and was not among the laboratory ordering features. The subject-level split, hyperparameters, class weighting, and early-stopping procedure were identical to the primary models. Full-feature and ascertainment-restricted AUROCs were compared using a two-sided paired bootstrap (2,000 resamples).

4.10 Deep temporal model

To test whether richer temporal modeling captured ordering information beyond the engineered summaries, a GRU network was trained on the raw measurement-count sequences. Each admission was represented as a sequence of twelve 4-hour windows spanning the first 48 hours of admission; each window was a 46-dimensional vector of measurement counts (35 laboratory tests and 11 vital-sign channels). The sequence was processed by a single-layer GRU with 128 hidden units, which updates a hidden state window by window. The hidden state after the final window was concatenated with the 70 static baseline features and passed to a multilayer-perceptron head (layer normalization over the 198-dimensional hidden-plus-static vector, dropout 0.2, a $198 \rightarrow 64$ linear layer with ReLU activation, and a $64 \rightarrow 1$ output). Counts and static features were standardized using training-set statistics only. The network was trained with AdamW

(learning rate 1×10^{-3} , weight decay 1×10^{-4}) and binary cross-entropy weighted by inverse class prevalence, with early stopping on validation AUROC (patience 8). The GRU was compared with the LightGBM ordering-summary model on the identical subject-level split using a two-sided paired bootstrap, repeated across five random initializations (mean $\pm$ standard deviation reported).

4.11 External validation, de novo replication, and early triage analysis

The final set of analyses tested zero-shot transfer to eICU-CRD, de novo replication in MC-MED, and the usefulness of test-ordering patterns during the earliest hours of admission, before many laboratory results were available.

Zero-shot transfer to eICU-CRD. The three MIMIC-IV-trained models, comprising the laboratory and vital-sign values model, the test-ordering patterns model, and the combined model, were applied directly to the eICU-CRD cohort without retraining or fine-tuning. Feature extraction was restricted to the 30 laboratory tests and 8 vital-sign streams shared between MIMIC-IV and eICU-CRD to ensure feature-space compatibility. Discrimination was assessed by AUROC for in-hospital mortality, AKI, and sepsis. Per-hospital AUROC was additionally computed for hospitals with at least 30 admissions and at least 3 positive cases for the outcome of interest.

De novo training on MC-MED. Four LightGBM models were trained from scratch on the MC-MED training set: a triage vital signs model, an ED test-ordering patterns model, a laboratory values model, and a combined model integrating triage vital signs with ED ordering. Discrimination was assessed by AUROC for inpatient admission and ICU admission outcomes on the held-out test set. This procedure tested cross-context replication in ED care, complementing eICU-CRD zero-shot validation, which tested cross-system transfer of MIMIC-IV-trained models.

Triage-window analysis. To assess the predictive value of ordering patterns in the earliest hours of admission, separate LightGBM models were trained and evaluated on progressively wider observation windows. The narrowest window comprised the first 4 hours of admission, corresponding to time bin 0 only. Subsequent windows extended by 4-hour increments to the full 48-hour observation period, corresponding to time bins 0–11 and yielding twelve nested windows. For each window, three models were compared: a laboratory and vital-sign values model, a test-ordering patterns model, and a demographic baseline model using only age, sex, and admission type. Each model was evaluated by AUROC at each window width. To contextualize early-window results, feature sparsity was quantified as the fraction of admissions with at least one laboratory measurement recorded within each 4-hour bin.

Data Availability

The data analyzed in this study are publicly available through PhysioNet, subject to the respective data use agreements, credentialing requirements, and required training. The primary analysis used MIMIC-IV v2.2, available at <https://physionet.org/content/mimiciv/2.2/>. External validation used the eICU Collaborative Research

Database v2.0, available at <https://physionet.org/content/eicu-crd/2.0/>. Cross-context replication used Multimodal Clinical Monitoring in the Emergency Department (MC-MED) v1.0.1, available at <https://physionet.org/content/mc-med/1.0.1/>. No new patient-level dataset was generated as part of this study.

Code Availability

The code used for data preprocessing, feature construction, model training, statistical analysis, and figure generation is available at <https://github.com/ShakunBaichoo/physician-ordering-signal>.

References

- [1] Callahan, A., Shah, N.H., Chen, J.H.: Research and reporting considerations for observational studies using electronic health record data. *Annals of internal medicine* **172**(11_Supplement), 79–84 (2020)
- [2] Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., *et al.*: Mimic-iv, a freely accessible electronic health record dataset. *Scientific data* **10**(1), 1 (2023)
- [3] Agniel, D., Kohane, I.S., Weber, G.M.: Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj* **361** (2018)
- [4] Overmars, L.M., Niemantsverdriet, M.S., Groenhouf, T.K.J., De Groot, M.C., Hulsbergen-Veelken, C.A., Van Solinge, W.W., Musson, R.E., Ten Berg, M.J., Hoefer, I.E., Haitjema, S.: A wolf in sheep’s clothing: reuse of routinely obtained laboratory data in research. *Journal of Medical Internet Research* **24**(11), 40516 (2022)
- [5] Whiting, P., Toerien, M., Salis, I., Sterne, J.A., Dieppe, P., Egger, M., Fahey, T.: A review identifies and classifies reasons for ordering diagnostic tests. *Journal of clinical epidemiology* **60**(10), 981–989 (2007)
- [6] Weijden, T., Bokhoven, M.A., Dinant, G.-J., Hasselt, C.M., Grol, R.P.: Understanding laboratory testing in diagnostic uncertainty: a qualitative study in general practice. *The British Journal of General Practice* **52**(485), 974 (2002)
- [7] Freedman, D.B.: Towards better test utilization—strategies to improve physician ordering and their impact on patient outcomes. *Ejifcc* **26**(1), 15 (2015)
- [8] Zhi, M., Ding, E.L., Theisen-Toupal, J., Whelan, J., Arnaout, R.: The landscape of inappropriate laboratory testing: a 15-year meta-analysis. *PloS one* **8**(11), 78962 (2013)

- [9] Müskens, J.L., Kool, R.B., Dulmen, S.A., Westert, G.P.: Overuse of diagnostic testing in healthcare: a systematic review. *BMJ quality & safety* **31**(1), 54–63 (2022)
- [10] Xu, S., Hom, J., Balasubramanian, S., Schroeder, L.F., Najafi, N., Roy, S., Chen, J.H.: Prevalence and predictability of low-yield inpatient laboratory diagnostic tests. *JAMA network open* **2**(9), 1910967 (2019)
- [11] Wang, J.K., Hom, J., Balasubramanian, S., Schuler, A., Shah, N.H., Goldstein, M.K., Baiocchi, M.T., Chen, J.H.: An evaluation of clinical order patterns machine-learned from clinician cohorts stratified by patient mortality outcomes. *Journal of biomedical informatics* **86**, 109–119 (2018)
- [12] Chen, J.H., Podchiyska, T., Altman, R.B.: Orderrex: clinical order decision support and outcome predictions by data-mining electronic medical records. *Journal of the American Medical Informatics Association* **23**(2), 339–348 (2016)
- [13] Rossetti, S.C., Knaplund, C., Albers, D., Dykes, P.C., Kang, M.J., Korach, T.Z., Zhou, L., Schnock, K., Garcia, J., Schwartz, J., *et al.*: Healthcare process modeling to phenotype clinician behaviors for exploiting the signal gain of clinical expertise (hpm-expertsignals): development and evaluation of a conceptual framework. *Journal of the American Medical Informatics Association* **28**(6), 1242–1251 (2021)
- [14] Rossetti, S.C., Dykes, P.C., Knaplund, C., Kang, M.-J., Schnock, K., Garcia Jr, J.P., Fu, L.-H., Chang, F., Thai, T., Fred, M., *et al.*: The communicating narrative concerns entered by registered nurses (concern) clinical decision support early warning system: protocol for a cluster randomized pragmatic clinical trial. *JMIR Research Protocols* **10**(12), 30238 (2021)
- [15] Sharafoddini, A., Dubin, J.A., Lee, J.: Finding similar patient subpopulations in the icu using laboratory test ordering patterns. In: *Proceedings of the 2018 7th International Conference on Bioinformatics and Biomedical Science*, pp. 72–77 (2018)
- [16] Sharafoddini, A., Dubin, J.A., Maslove, D.M., Lee, J.: A new insight into missing data in intensive care unit patient profiles: observational study. *JMIR medical informatics* **7**(1), 11605 (2019)
- [17] Sisk, R., Lin, L., Sperrin, M., Barrett, J.K., Tom, B., Diaz-Ordaz, K., Peek, N., Martin, G.P.: Informative presence and observation in routine health data: a review of methodology for clinical risk prediction. *Journal of the American Medical Informatics Association* **28**(1), 155–166 (2021)
- [18] Tan, A.L., Getzen, E.J., Hutch, M.R., Strasser, Z.H., Gutiérrez-Sacristán, A., Le, T.T., Dagliati, A., Morris, M., Hanauer, D.A., Moal, B., *et al.*: Informative missingness: What can we learn from patterns in missing laboratory data in the

- electronic health record? *Journal of biomedical informatics* **139**, 104306 (2023)
- [19] Sun, M., Engelhard, M.M., Bedoya, A.D., Goldstein, B.A.: Incorporating informatively collected laboratory data from ehr in clinical prediction models. *BMC Medical Informatics and Decision Making* **24**(1), 206 (2024)
 - [20] Singh, J., Sato, M., Ohkuma, T.: On missingness features in machine learning models for critical care: observational study. *JMIR Medical Informatics* **9**(12), 25022 (2021)
 - [21] Liu, V.X., Lu, Y., Carey, K.A., Gilbert, E.R., Afshar, M., Akel, M., Shah, N.S., Dolan, J., Winslow, C., Kipnis, P., *et al.*: Comparison of early warning scoring systems for hospitalized patients with and without infection at risk for in-hospital mortality and transfer to the intensive care unit. *JAMA network open* **3**(5), 205191 (2020)
 - [22] Khattak, Z.E., El Sharu, H., Bhutta, B.S.: Overview on ordering and evaluation of laboratory tests. In: *StatPearls* [Internet]. StatPearls Publishing, ??? (2023)
 - [23] Pollard, T.J., Johnson, A.E., Raffa, J.D., Celi, L.A., Mark, R.G., Badawi, O.: The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data* **5**(1), 180178 (2018)
 - [24] Kansal, A., Chen, E., Jin, T., Rajpurkar, P., Kim, D.: Multimodal clinical monitoring in the emergency department (mc-med). Published online March **3** (2025)
 - [25] Rudd, K.E., Johnson, S.C., Agesa, K.M., Shackelford, K.A., Tsoi, D., Kievlan, D.R., Colombara, D.V., Ikuta, K.S., Kissoon, N., Finfer, S., *et al.*: Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet* **395**(10219), 200–211 (2020)
 - [26] Tomašev, N., Glorot, X., Rae, J.W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., *et al.*: A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**(7767), 116–119 (2019)
 - [27] Shapley, L.S., *et al.*: A value for n-person games (1953)
 - [28] Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* **46**(3), 399–424 (2011)
 - [29] Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R.: A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases* **40**(5), 373–383 (1987)

- [30] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I.: From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence* **2**(1), 56–67 (2020)
- [31] Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C.K., Suter, P., Thijs, L.G.: The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix). *Intensive care medicine* **22**(7), 707–710 (1996)
- [32] Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning* **16**(4), 494–591 (2023)
- [33] Subbe, C.P., Kruger, M., Rutherford, P., Gemmel, L.: Validation of a modified early warning score in medical admissions. *Qjm* **94**(10), 521–526 (2001)
- [34] Prytherch, D.R., Smith, G.B., Schmidt, P.E., Featherstone, P.I.: Views—towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation* **81**(8), 932–937 (2010)
- [35] Kellum, J.A., Lameire, N., Group, K.A.G.W.: Diagnosis, evaluation, and management of acute kidney injury: a kdigo summary (part 1). *Critical care* **17**(1), 204 (2013)
- [36] Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.-D., Coopersmith, C.M., *et al.*: The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama* **315**(8), 801–810 (2016)
- [37] Seymour, C.W., Liu, V.X., Iwashyna, T.J., Brunkhorst, F.M., Rea, T.D., Scherag, A., Rubenfeld, G., Kahn, J.M., Shankar-Hari, M., Singer, M., *et al.*: Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama* **315**(8), 762–774 (2016)
- [38] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **30** (2017)
- [39] Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. *Information fusion* **81**, 84–90 (2022)
- [40] Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems* **35**, 507–520 (2022)

- [41] Lei, J., Zhai, J., Zhang, Y., Qi, J., Sun, C.: Supervised machine learning models for predicting sepsis-associated liver injury in patients with sepsis: development and validation study based on a multicenter cohort study. *Journal of Medical Internet Research* **27**, 66733 (2025)
- [42] Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G.: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Journal of British Surgery* **102**(3), 148–158 (2015)

Author Contributions

Shakuntala Baichoo designed the study, conducted the experiments, analyzed the data, and prepared the initial manuscript draft, with support from Ali Abedi. Ali Abedi revised the draft and restructured the manuscript. Barry Rubin and Bo Wang supervised the study. All authors reviewed and approved the final manuscript.

Competing Interests

The authors declare no competing interests.

Figure Legends

Figure 1. Conceptual overview of laboratory test-ordering patterns as early prognostic signals in electronic health records. After patient admission, clinicians order laboratory tests based on clinical concern, diagnostic uncertainty, and monitoring needs. These ordering patterns are recorded before or alongside laboratory results and may provide early prognostic information while test results are still pending. In MIMIC-IV, models using test-ordering pattern features achieved performance close to models using measured laboratory and vital-sign values across multiple adverse outcomes, supporting ordering behavior as a routinely available EHR signal for early risk stratification.

Figure 2. (a) Incremental discrimination contribution of test-ordering features over the SOFA severity score, shown as AUROC for SOFA-only, Ordering-only, and combined SOFA + Ordering LightGBM models for in-hospital mortality, sepsis, and acute kidney injury (AKI). Error bars indicate 95% bootstrap confidence intervals. Gains of the combined model over SOFA-only are +0.041, +0.034, and +0.044 for mortality, sepsis, and AKI, respectively. (b) In-hospital mortality across ordering-escalation quartiles among patients with relatively reassuring laboratory values (value abnormality below the cohort median). Mortality follows a U-shaped pattern across quartiles: the most de-escalating quartile (Q1, 1.36%) and the most escalating quartile (Q4, 0.84%) show higher mortality than the two intermediate quartiles (Q2, 0.58%; Q3, 0.45%), indicating that ordering trajectory carries mortality signal independent of laboratory values. Declining test ordering may reflect therapeutic withdrawal or transitions to palliative care in patients with deteriorating clinical status, while increasing ordering

may reflect physician concern triggered by bedside signs not yet apparent in laboratory measurements.

Figure 3. Dose-response of ordering intensity versus adverse outcomes across three independent datasets: (a) In-hospital mortality in MIMIC-IV; (b) Sepsis in MIMIC-IV; (c) AKI in MIMIC-IV; (d) In-hospital mortality in eICU-CRD; (e) ICU admission in MC-MED; and (f) Inpatient admission in MC-MED. Event rates are reported by ordering-intensity quartile (MIMIC-IV, eICU-CRD; equal-frequency quartiles) or by ordering-intensity group (MC-MED; not equal-frequency; 39.8% of MC-MED ED visits have zero orders, which constitutes a structural floor in the low-ordering group). Error bars indicate 95% bootstrap confidence intervals. In MIMIC-IV ($N = 305,732$), event rates rise strictly monotonically with ordering intensity for all three outcomes. In eICU-CRD ($N = 200,859$), the gradient replicates but mortality is elevated in the lowest-ordering quartile (Q1), a pattern consistent with palliative or comfort-care decisions concentrating in this group. In MC-MED ($N = 118,385$), inpatient admission rates are also elevated in the lowest-ordering group owing to the zero-order floor effect.

Figure 4. (a) Triage comparison at the 4-hour timepoint. AUROC is reported for three feature sets across four outcomes on the MIMIC-IV held-out test set, using only data observable within the first 4 hours of admission. Baseline uses admission demographics and administrative variables (age, sex, admission type). Ordering-only uses the test-ordering trajectory in the first 4 h, with no measured values. Values-only uses whichever laboratory and vital-sign results have returned by 4 h. Only 26.6% of admissions have any laboratory result by 4 h, so the values model is operating on a sparse signal. Despite this, the ordering-only model approaches the values-only model for in-hospital mortality (0.819 vs 0.831) and sepsis (0.780 vs 0.796), trailing by a small margin (≈ 0.012 to 0.016 AUROC). (b) AUROC as a function of observation window. Each line plots discrimination of the test-ordering patterns model (solid) and the laboratory and vital-sign values model (dashed) at observation windows ranging from 4 h to 48 h after admission, for in-hospital mortality (blue), 30-day readmission (teal), acute kidney injury (gold), and sepsis (pink). At $t = 4$ h (triage), the two models are close for mortality and sepsis; for AKI and readmission, the values model gains ground as more laboratory results return over the first 24 h. The ordering signal continues to rise monotonically with longer observation windows for all four outcomes.

Figure 5. External validation of the test-ordering signal across two independent datasets. (a, b) Zero-shot transfer: LightGBM classifiers trained on MIMIC-IV were evaluated on eICU-CRD without retraining or fine-tuning, using the 30 laboratory tests and 8 vital-sign streams shared between the datasets. (a) MIMIC-IV held-out internal test set ($N = 45,805$) shows reference internal performance for both models on these 30 shared laboratory tests and 8 vital-sign streams (hence below the full-feature values in Table 1). (b) eICU-CRD external evaluation (108,324 ICU stays across 206 US hospitals) shows substantial attenuation relative to the internal benchmark for both the test-ordering patterns model (Ordering-only) and the laboratory and vital-sign values model (Values-only), indicating that both ordering metadata and measured values are affected by inter-system domain shift. (c–f) De novo replication in the MC-MED Stanford Emergency Department cohort ($N = 11,792$ held-out visits, 2020–2022). Four LightGBM classifiers were trained de novo on MC-MED: a

Triage vitals model (NEWS2-like physiological score), an ED ordering model (orders alone, no measured values), an ED full model (orders combined with triage vitals), and a Lab values model (measured laboratory results). These models were evaluated for (c) in-hospital mortality, (d) 30-day readmission, (e) ICU admission, and (f) inpatient admission. The ED ordering model outperforms both triage vitals and laboratory values for inpatient admission (0.876 vs 0.832 vs 0.783), while triage vitals are strongest for ICU admission (0.911), consistent with ICU triage being driven primarily by acute physiological derangement. Combining test-ordering with triage vitals (ED full) achieves the highest discrimination for all four outcomes, reaching AUROC 0.957 for ICU admission and 0.933 for inpatient admission. Error bars indicate 95% bootstrap confidence intervals.

Tables

Table 1 Discrimination performance of three Light Gradient Boosting Machine (LightGBM) classifiers per outcome, reported as area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC), each with 95% bootstrap confidence interval (CI). Values model: laboratory and vital sign value features with static baseline features. Ordering model: test-ordering pattern features with static baseline features and no measured laboratory or vital sign values. Combined model: laboratory and vital sign value features, test-ordering pattern features, and static baseline features. p -values are from two-sided paired bootstrap tests comparing the values model with the ordering model.

(a) AUROC				
Outcome	Values AUROC (95% CI)	Ordering AUROC (95% CI)	p -value Values vs ordering	Combined AUROC (95% CI)
In-hospital mortality	0.932 (0.925–0.939)	0.905 (0.896–0.913)	< 0.001	0.934 (0.927–0.940)
30-day readmission	0.705 (0.699–0.711)	0.702 (0.696–0.708)	0.013	0.705 (0.699–0.710)
Acute kidney injury	0.882 (0.878–0.885)	0.835 (0.831–0.839)	< 0.001	0.888 (0.885–0.892)
Sepsis	0.924 (0.919–0.928)	0.912 (0.907–0.917)	< 0.001	0.926 (0.921–0.930)

(b) AUPRC				
Outcome	Values AUPRC (95% CI)	Ordering AUPRC (95% CI)	p -value Values vs ordering	Combined AUPRC (95% CI)
In-hospital mortality	0.453 (0.422–0.481)	0.328 (0.298–0.357)	< 0.001	0.456 (0.425–0.486)
30-day readmission	0.423 (0.413–0.433)	0.421 (0.411–0.431)	0.313	0.424 (0.414–0.434)
Acute kidney injury	0.764 (0.758–0.771)	0.667 (0.658–0.675)	< 0.001	0.778 (0.771–0.785)
Sepsis	0.669 (0.654–0.683)	0.644 (0.629–0.658)	< 0.001	0.676 (0.661–0.690)

Table 2 Marginal area under the receiver operating characteristic curve (AUROC) contribution, expressed as Shapley value ϕ , of each clinical data modality, derived from 60 Light Gradient Boosting Machine (LightGBM) models trained across all 15 non-empty modality coalitions and four outcomes. Rankings within each outcome are shown in parentheses.

Modality	In-hospital mortality ϕ (rank)	30-day readmission ϕ (rank)	Acute kidney injury ϕ (rank)	Sepsis ϕ (rank)
Laboratory values	0.138 (1)	0.060 (2)	0.164 (1)	0.121 (1)
Test-ordering patterns	0.108 (3)	0.041 (3)	0.101 (2)	0.110 (3)
Static baseline features	0.114 (2)	0.100 (1)	0.095 (3)	0.112 (2)
Vital sign values	0.075 (4)	0.003 (4)	0.027 (4)	0.082 (4)

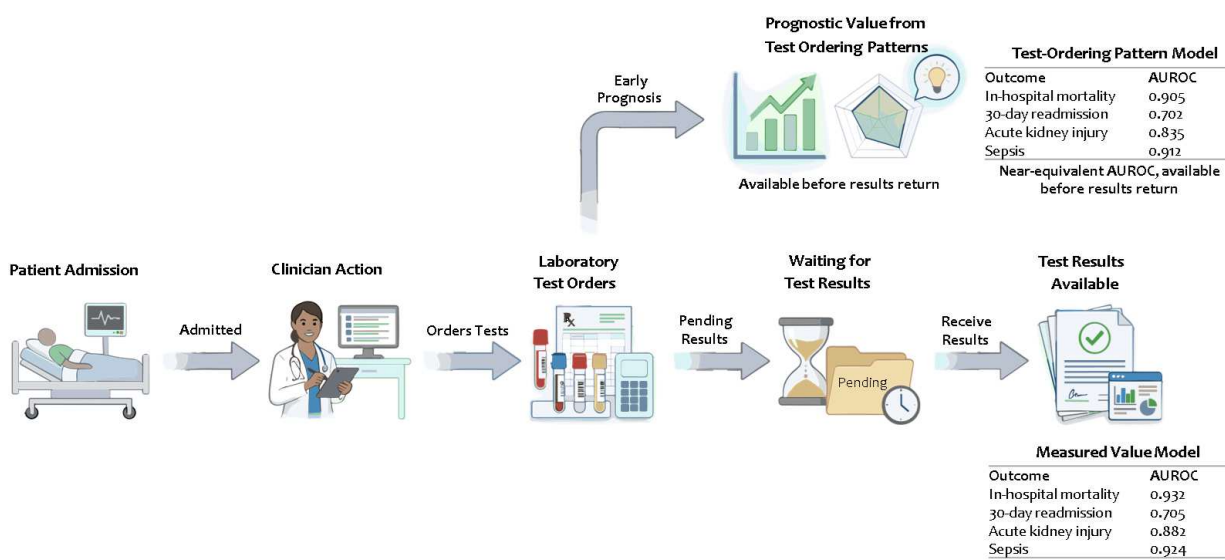


Figure 1. Conceptual overview of laboratory test-ordering patterns as early prognostic signals in electronic health records. After patient admission, clinicians order laboratory tests based on clinical concern, diagnostic uncertainty, and monitoring needs. These ordering patterns are recorded before or alongside laboratory results and may provide early prognostic information while test results are still pending. In MIMIC-IV, models using test-ordering pattern features achieved performance close to models using measured laboratory and vital-sign values across multiple adverse outcomes, supporting ordering behavior as a routinely available EHR signal for early risk stratification.

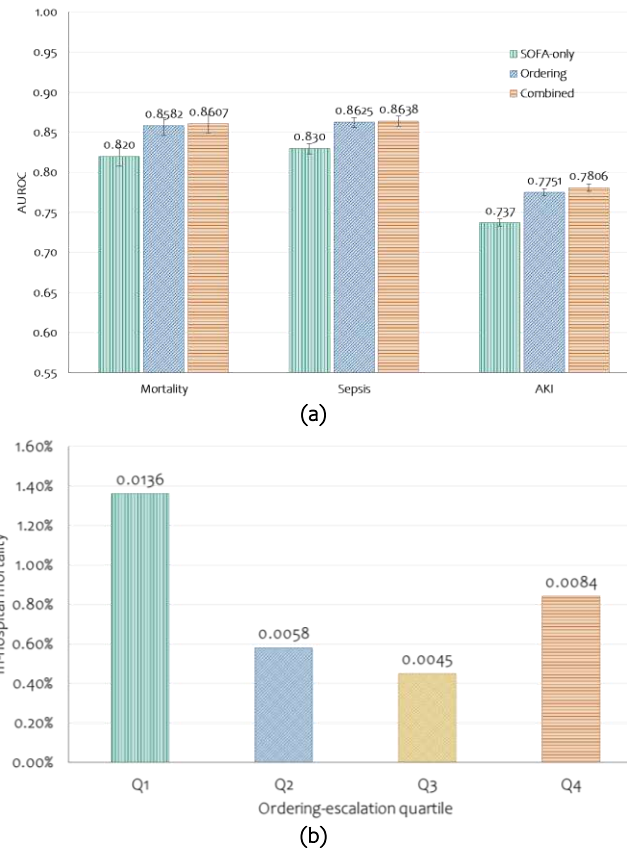


Figure 2. (a) Incremental discrimination contribution of test-ordering features over the SOFA severity score, shown as AUROC for SOFA-only, Ordering-only, and combined SOFA + Ordering LightGBM models for in-hospital mortality, sepsis, and acute kidney injury (AKI). Error bars indicate 95% bootstrap confidence intervals. Gains of the combined model over SOFA-only are +0.041, +0.034, and +0.044 for mortality, sepsis, and AKI, respectively. (b) In-hospital mortality across ordering-escalation quartiles among patients with relatively reassuring laboratory values (value abnormality below the cohort median). Mortality follows a U-shaped pattern across quartiles: the most de-escalating quartile (Q1, 1.36%) and the most escalating quartile (Q4, 0.84%) show higher mortality than the two intermediate quartiles (Q2, 0.58%; Q3, 0.45%), indicating that ordering trajectory carries mortality signal independent of laboratory values. Declining test ordering may reflect therapeutic withdrawal or transitions to palliative care in patients with deteriorating clinical status, while increasing ordering may reflect physician concern triggered by bedside signs not yet apparent in laboratory measurements.

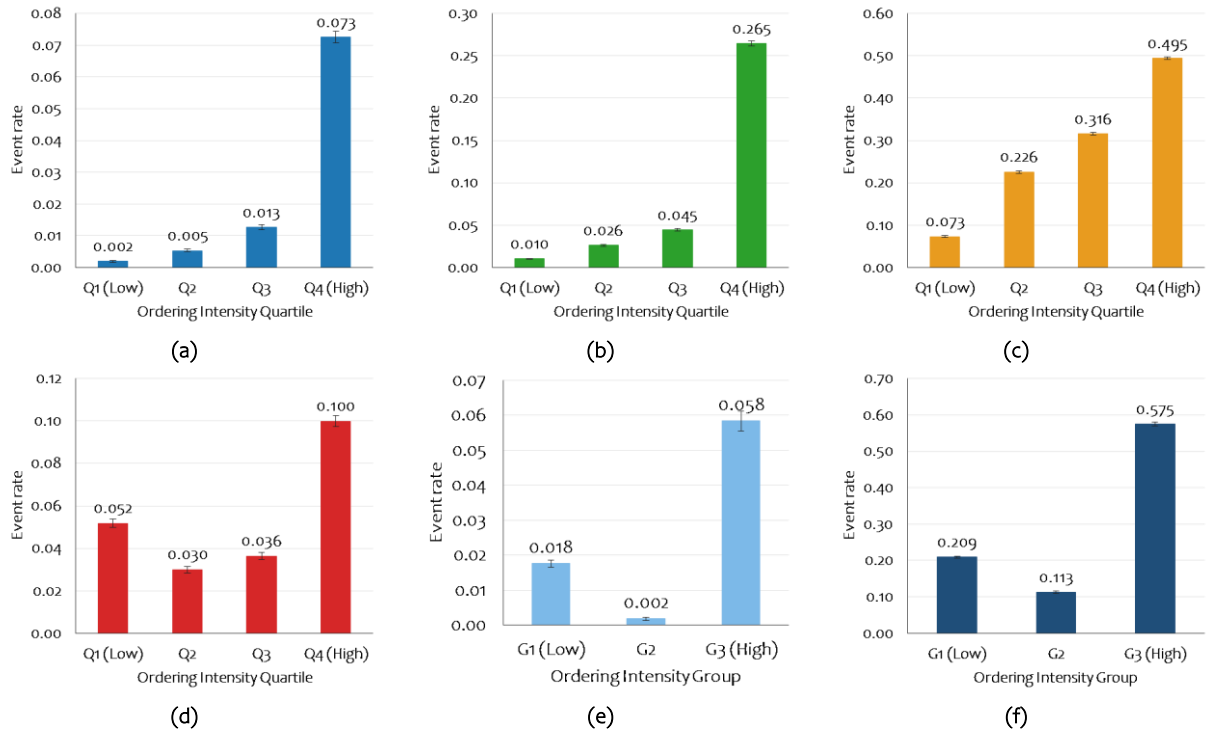


Figure 3. Dose-response of ordering intensity versus adverse outcomes across three independent datasets: (a) In-hospital mortality in MIMIC-IV; (b) Sepsis in MIMIC-IV; (c) AKI in MIMIC-IV; (d) In-hospital mortality in eICU-CRD; (e) ICU admission in MC-MED; and (f) Inpatient admission in MC-MED. Event rates are reported by ordering-intensity quartile (MIMIC-IV, eICU-CRD; equal-frequency quartiles) or by ordering-intensity group (MC-MED; not equal-frequency; 39.8% of MC-MED ED visits have zero orders, which constitutes a structural floor in the low-ordering group). Error bars indicate 95% bootstrap confidence intervals. In MIMIC-IV ($N = 305,732$), event rates rise strictly monotonically with ordering intensity for all three outcomes. In eICU-CRD ($N = 200,859$), the gradient replicates but mortality is elevated in the lowest-ordering quartile (Q1), a pattern consistent with palliative or comfort-care decisions concentrating in this group. In MC-MED ($N = 118,385$), inpatient admission rates are also elevated in the lowest-ordering group owing to the zero-order floor effect.

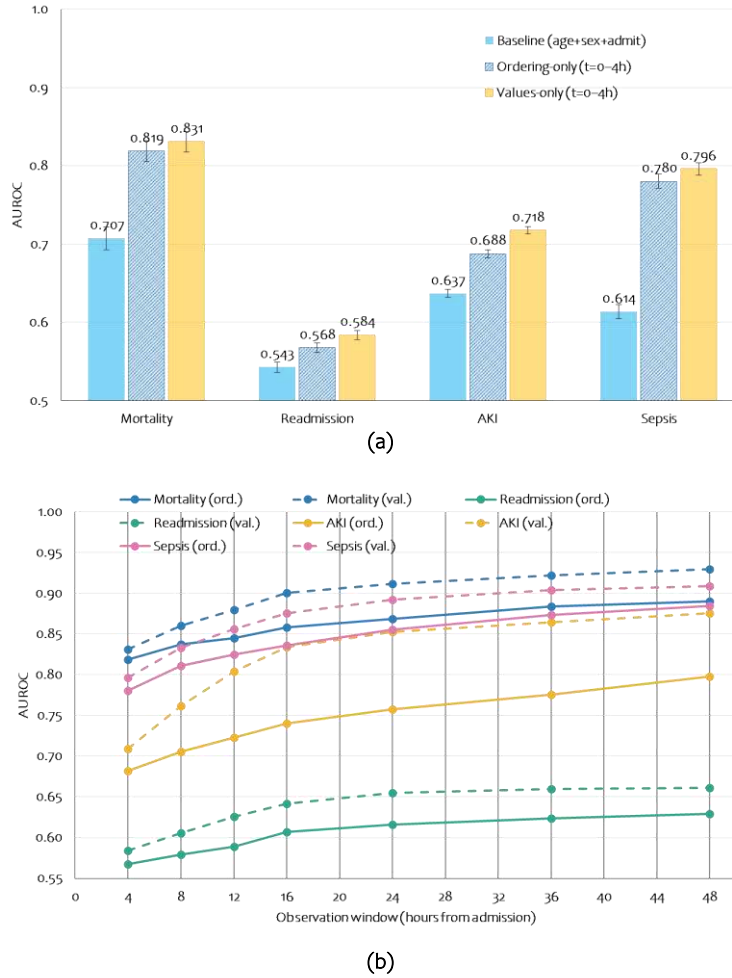


Figure 4. (a) Triage comparison at the 4-hour timepoint. AUROC is reported for three feature sets across four outcomes on the MIMIC-IV held-out test set, using only data observable within the first 4 hours of admission. Baseline uses admission demographics and administrative variables (age, sex, admission type). Ordering-only uses the test-ordering trajectory in the first 4 h, with no measured values. Values-only uses whichever laboratory and vital-sign results have returned by 4 h. Only 26.6% of admissions have any laboratory result by 4 h, so the values model is operating on a sparse signal. Despite this, the ordering-only model approaches the values-only model for in-hospital mortality (0.819 vs 0.831) and sepsis (0.780 vs 0.796), trailing by a small margin (≈ 0.012 – 0.016 AUROC). (b) AUROC as a function of observation window. Each line plots discrimination of the test-ordering patterns model (solid) and the laboratory and vital-sign values model (dashed) at observation windows ranging from 4 h to 48 h after admission, for in-hospital mortality (blue), 30-day readmission (teal), acute kidney injury (gold), and sepsis (pink). At $t = 4$ h (triage), the two models are close for mortality and sepsis; for AKI and readmission, the values model gains ground as more laboratory results return over the first 24 h. The ordering signal continues to rise monotonically with longer observation windows for all four outcomes.

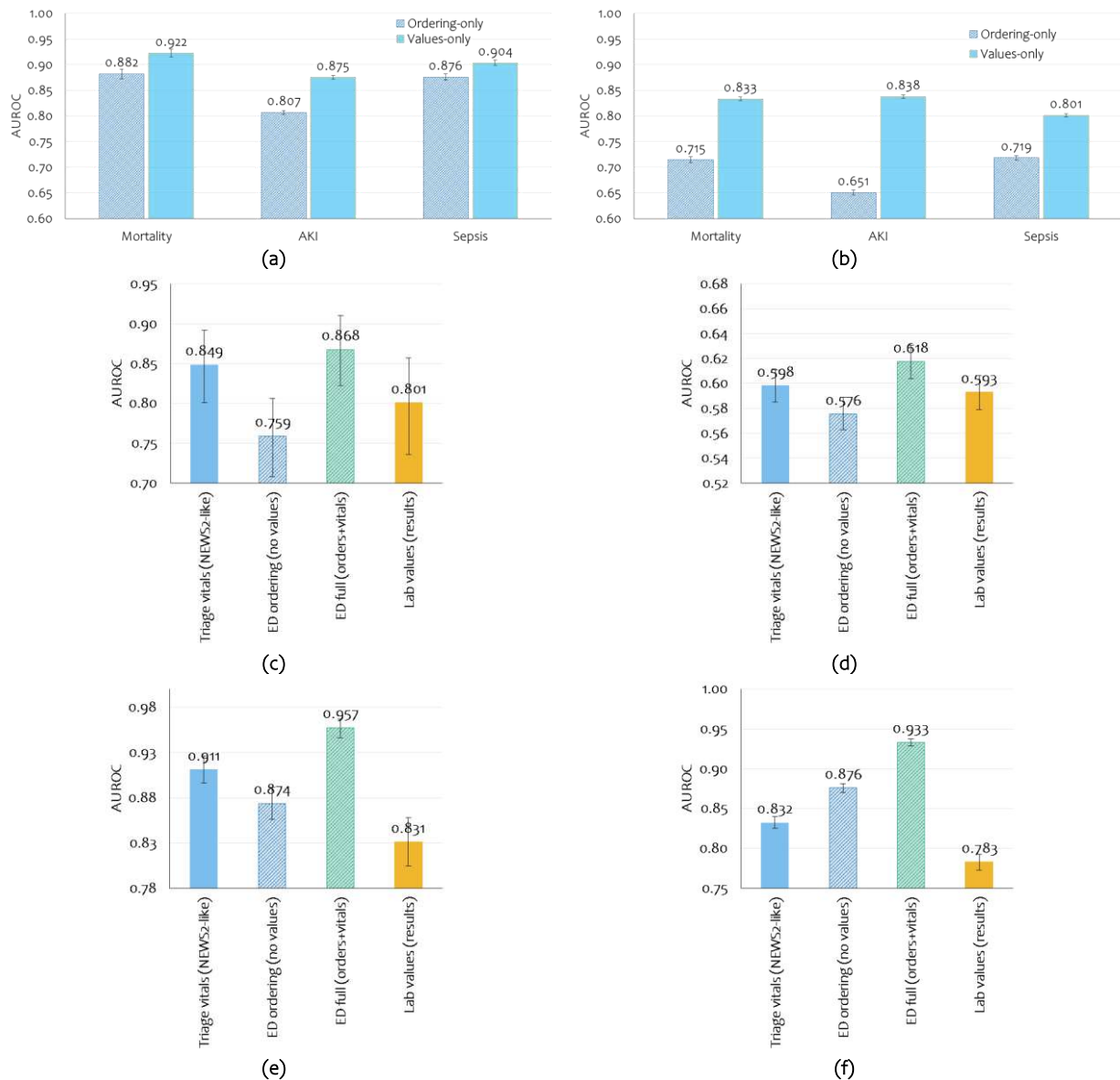


Figure 5. External validation of the test-ordering signal across two independent datasets. (a, b) Zero-shot transfer: LightGBM classifiers trained on MIMIC-IV were evaluated on eICU-CRD without retraining or fine-tuning, using the 30 laboratory tests and 8 vital-sign streams shared between the datasets. (a) MIMIC-IV held-out internal test set ($N = 45,805$) shows reference internal performance for both models on these 30 shared laboratory tests and 8 vital-sign streams (hence below the full-feature values in Table 1). (b) eICU-CRD external evaluation (108,324 ICU stays across 206 US hospitals) shows substantial attenuation relative to the internal benchmark for both the test-ordering patterns model (Ordering-only) and the laboratory and vital-sign values model (Values-only), indicating that both ordering metadata and measured values are affected by inter-system domain shift. (c–f) De novo replication in the MC-MED Stanford Emergency Department cohort ($N = 11,792$ held-out visits, 2020–2022). Four LightGBM classifiers were trained de novo on MC-MED: a Triage vitals model (NEWS2-like physiological score), an ED ordering model (orders alone, no measured values), an ED full model (orders combined with triage vitals), and a Lab values model (measured laboratory results). These models were evaluated for (c) in-hospital mortality, (d) 30-day readmission, (e) ICU admission, and (f) inpatient admission. The ED ordering model outperforms both triage vitals and laboratory values for inpatient admission (0.876 vs 0.832 vs 0.783), while triage vitals are strongest for ICU admission (0.911), consistent with ICU triage being driven primarily by acute physiological derangement. Combining test-ordering with triage

vitals (ED full) achieves the highest discrimination for all four outcomes, reaching AUROC 0.957 for ICU admission and 0.933 for inpatient admission. Error bars indicate 95% bootstrap confidence intervals.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [03supplementarytables.docx](#)