

MultiDiseasePred: An Interpretable Multi-Disease Risk Prediction Framework and Its Application for Early Risk Stratification in the Emergency Department

Yanwei Jin

University of Minnesota

Ke Xu

University of Minnesota

Yinzhaoh Wang

University of Minnesota

Chenyin Chu

Monash University

Xinnie Mai

University of Minnesota

Zechen Zhou

University of Minnesota

Andrew Meyer

University of Minnesota

Yumeng Chen

University of Minnesota

Geetha Saarunya

University of Minnesota Medical School

R. Adams Dudley

University of Minnesota Medical School

Nicholas E. Ingraham

University of Minnesota Medical School

David A. Wacker

University of Minnesota Medical School

Mingquan Lin

University of Minnesota

Michael A. Puskarich

Hennepin Healthcare

Feng Xie

xie00469@umn.edu

University of Minnesota

Article

Keywords:

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10963805/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

At emergency department triage, clinicians must assess multiple overlapping acute conditions simultaneously, yet many existing risk tools are developed for individual conditions and provide outputs that are difficult to compare across diseases with substantially different frequencies. We developed and validated MultiDiseasePred, an interpretable multi-task learning framework for simultaneous risk estimation across multiple non-mutually exclusive clinical targets, and demonstrated its application for early multi-condition risk profiling in emergency department triage. MultiDiseasePred estimated risks for seven acute conditions (sepsis, ARDS, COPD/asthma exacerbation, acute coronary syndrome, viral pneumonia, pulmonary embolism, acute kidney injury) and two disposition endpoints (hospitalization; a critical outcome) in a single inference pass. A per-target gate weighted a shared expert pool differently for each target, allowing the model to learn shared clinical and physiological signals across related conditions while preserving target-specific prediction pathways. To reduce feature acquisition burden while maintaining model performance, a parsimonious feature selection procedure identified 12 variables available at triage from routinely available candidate predictors, without requiring laboratory results, imaging, or downstream clinical documentation. Calibrated probabilities were summarized using prevalence-adjusted lift measures to enable comparison across targets with different baseline prevalence, and each risk estimate was accompanied by target-specific feature attributions indicating which triage findings contributed to the prediction. Across 2,203,437 encounters from three health systems, the model achieved a macro-AUROC of 0.900 (95% CI, 0.897–0.903) for the seven acute conditions and 0.852 (0.849–0.856) for the two disposition endpoints in internal validation, outperforming NEWS and generally matching or exceeding single-task machine-learning baselines. In external validation without local retraining or recalibration, acute-condition macro-AUROC was 0.793 at Beth Israel Deaconess Medical Center and 0.866 at Stanford. These findings support MultiDiseasePred as a generalizable framework for interpretable, early multi-condition risk profiling and motivate prospective evaluation in emergency care workflows.

Introduction

Emergency department (ED) presentations often involve overlapping symptoms and physiological abnormalities that may reflect multiple acute conditions, such as infection, respiratory failure, cardiovascular disease, and renal dysfunction¹. Traditional diagnostic and risk assessment approaches combine clinical variables to estimate risk for an individual condition and are generally not designed to assess multiple co-occurring conditions simultaneously, despite this reflecting the broader clinical decision-making process in emergency care². Widely implemented scores, such as qSOFA for sepsis risk assessment or the HEART score for acute coronary syndrome, offer transparent and practical decision support but rely on fixed thresholds and disease-specific assumptions^{3,4}. Although machine-learning models have expanded the capacity for clinical prediction⁵, many remain focused on a single condition or narrowly defined clinical application⁶. This single-task paradigm creates an important unmet need in early clinical risk assessment: clinicians often require a coordinated risk profile across multiple plausible,

non-mutually exclusive conditions, whereas separately trained models may produce fragmented outputs, learn overlapping predictors independently, fail to leverage information shared across related conditions, and provide limited support for prioritizing competing risks at bedside. These limitations are particularly relevant for low-prevalence but high-acuity conditions, where shared physiological signals across related conditions may improve representations⁷.

This gap is particularly consequential during ED triage⁸, when clinicians must make time-sensitive decisions before definitive diagnostic information is available⁹. Initial assessment often relies on triage vital signs, chief complaints, and readily accessible clinical history, whereas laboratory results, imaging, and downstream documentation typically become available later in the encounter. Models that depend heavily on these downstream data may have limited utility for immediate risk stratification and may be less transportable across institutions with different diagnostic workflows and resource availability¹⁰. Triage is also a point at which care diverges along lines unrelated to illness severity. Emergency Severity Index assignment has been reported to differ by race, ethnicity, and primary language after adjustment for clinical factors and department conditions^{11,12}, and several of the conditions predicted here, including sepsis and acute coronary syndrome, show documented differences in recognition and time to treatment across these groups. A model applied at this point therefore has to be examined for whether it narrows or widens those differences rather than assumed to be neutral. These constraints highlight the need for a model that uses data available at triage to provide simultaneous early risk profiling across multiple acute conditions and disposition endpoints. To address these challenges, we developed MultiDiseasePred, an interpretable, task-adaptive multi-task learning framework for simultaneous risk estimation across multiple acute conditions and disposition endpoints at ED triage. The framework uses a parsimonious set of routine variables available at triage or from prior clinical history to estimate these risks without requiring laboratory results, imaging, or downstream documentation. MultiDiseasePred also incorporates calibrated probabilities, prevalence-adjusted cross-target comparison, and target-specific SHapley Additive exPlanations (SHAP)-based feature attribution to support interpretable risk profiling across heterogeneous prediction targets. The framework was developed in a large multicenter ED cohort, and evaluated for transportability in two independent external health systems to assess whether a unified multi-task architecture could deliver clinically interpretable early risk predictions across diverse emergency-care settings.

Methods

MultiDiseasePred Framework

MultiDiseasePred is an interpretable, task-adaptive multi-task learning framework designed for simultaneous risk estimation across multiple non-mutually exclusive clinical prediction targets¹³. It is built in four modules. The first establishes that the targets share predictive structure, the second reduces the input set to variables available at triage, the third learns all nine risks jointly over that input set, and the fourth converts the model outputs into calibrated estimates that can be compared across targets. An

overview of framework development, validation, and intended clinical use is provided in **Figure 1**. MultiDiseasePred has been released as an open-source package (<https://github.com/fengx13/MultiDiseasePred>) that accommodates structured datasets with differing target sets and feature configurations, with the intended bedside use summarized in **Figure 1d**. This study is reported in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence (TRIPOD+AI) statement¹⁴.

Module 1: Outcome Co-occurrence Assessment

We first characterized the co-occurrence structure of the seven acute-condition targets to evaluate whether the modeled conditions showed empirical dependence and shared clinical structure. Within the UMN development and BIDMC external cohorts, we derived pairwise joint and marginal event counts from the outcome-indicator matrix and summarized each pair by the phi coefficient, which for two binary indicators is the Pearson correlation between them. The co-occurrence analysis provided empirical context for the shared-expert design by identifying condition pairs with overlapping clinical structure while preserving the need for target-specific prediction pathways.

Module 2: Variable Selection by Greedy Forward Selection

Shared predictive structure is only useful if the inputs can be obtained at the moment the prediction is needed. To reduce feature acquisition burden and support triage-stage implementation, we derived a parsimonious predictor set from the 62 candidate variables using greedy forward selection. Variables were added one at a time using only the development data. At each step, separate gradient-boosted tree models were fitted for the seven acute-condition targets using a random 400,000-encounter subsample of the development cohort, and candidate variables were evaluated on the held-out selection split. Among the remaining candidates, the variable selected at each step was the one that produced the largest improvement in macro-averaged AUROC across the seven acute-condition targets, conditional on the variables already selected. The final number of variables was determined by a prespecified performance-retention rule that took the smallest value of k reaching at least 98% of the AUROC obtained with all 62 candidate variables, which selected 12 variables. The resulting 12-variable set was then fixed before final model training, external validation, and comparison with baseline models. The selected variables and their definitions are listed in **eTable 1**.

Module 3: Multi-task Modeling with Mixture-of-Experts Representation

With the input set now fixed, the co-occurrence structure characterized in Module 1 motivated a shared-representation design over those twelve variables. The framework jointly models the nine binary prediction targets within a single multi-task, multi-label architecture, allowing each encounter to be associated with multiple acute conditions and disposition outcomes. The retained variables are passed to a shared pool of 13 feedforward expert subnetworks, following a Mixture-of-Experts formulation^{15–17}. For each target, expert outputs are combined using a learned gate over the shared pool, formed by averaging two target-specific logit matrices and normalizing at a fixed temperature. The gate does not

depend on the input, so each target applies the same weighting of experts to every encounter. The resulting task-adapted representation is passed to a target-specific linear output layer, which accommodates differences in baseline prevalence and decision boundaries across endpoints. All nine risks are therefore generated in a single forward pass per encounter. The full network specification, regularization terms, and training hyperparameters are provided in Supplementary **eMethods 2**.

Module 4: Calibration and Interpretation

Predicted probabilities were calibrated separately for each target by isotonic regression on the logit scale¹⁸, fitted in the development cohort and then applied unchanged to the internal and external validation cohorts without local recalibration. Calibrated probabilities were treated as the primary risk estimates. To facilitate comparison across targets with different baseline prevalence, each calibrated probability was additionally transformed into a log-lift ratio, defined as the natural logarithm of the ratio between the predicted probability and the target-specific baseline prevalence. Feature attribution was computed using expected gradients on the logit scale, a sampling-based SHAP-related attribution method, with estimates averaged across three independently trained model seeds. The four modules together return, for each encounter, nine calibrated probabilities, the lift of each over its own baseline prevalence, and the variable contributions behind each one.

Data Sources and Study Cohorts

The University of Minnesota (UMN) cohort was derived from M Health Fairview (MHFV), an integrated multi-hospital academic health system affiliated with UMN. It included ED encounters from 11 hospitals in Minnesota between 2019 and 2025, ranging from a Level III academic trauma center to community EDs. The cohort was used for model development and internal validation. Use of these data was approved by the University of Minnesota Institutional Review Board (STUDY00023702), which waived the requirement for informed consent because the analysis was retrospective and used data already collected in the course of clinical care.

External validation was conducted in two independent cohorts. The first was the Medical Information Mart for Intensive Care IV, Emergency Department (MIMIC-IV-ED) database, a de-identified dataset from Beth Israel Deaconess Medical Center (BIDMC), Boston, comprising ED attendances between 2011 and 2019^{19,20}. The second external dataset was derived from Stanford Health Care (Stanford), comprising ED encounters between 2020 and 2022²¹. Access to both datasets was obtained through PhysioNet under its credentialing and data use agreement²². Both cohorts were used exclusively for external validation and contributed no data to model development. The five-level Emergency Severity Index (ESI) was available across the three cohorts and was used only for cohort characterization; it was excluded a priori from the candidate predictor set.

Across the three cohorts we assembled 62 candidate variables spanning demographics, triage vital signs, chief complaints, prior ED utilization and comorbidity indicators (**Figure 1b**). Comorbidities were

encoded as binary indicators derived from the Charlson and Elixhauser comorbidity groupings^{23–25}. The complete ICD-9-CM and ICD-10-CM code sets defining each comorbidity group are provided in **eTable 2**. The nine prediction targets are defined in the Prediction Targets section.

Data Preprocessing

We included adult encounters (age ≥ 18 years) and excluded records that lacked a documented arrival time and contained no information beyond the encounter identifier. The UMN cohort was divided temporally: encounters from 2019–2023 comprised the development cohort, whereas encounters from 2024–2025 were reserved as the internal validation set. Within the development cohort, a random 20% ($n = 262,687$) was held out for variable and model selection and was not used to fit the final model. Each external cohort was retained in its entirety for external validation.

Preprocessing was specified by variable type. Triage vital-sign values outside prespecified physiological ranges were clipped to the nearest boundary. Missing continuous triage vital signs were imputed using variable-specific medians computed on the development cohort, while missing Glasgow Coma Scale scores, pain score, and comorbidity indicators were assigned prespecified deterministic values. Full details of the missingness and the imputation protocol are provided in **eMethods 1**. After applying the cohort-selection criteria, the UMN cohort comprised 1,664,277 encounters. Missingness for each retained variable is reported in **eTable 3**. The same preprocessing rules were applied to every encounter regardless of age, sex, or site, and no step was conditioned on a demographic characteristic. Whether the resulting imputation rates differ across demographic groups was not examined.

Prediction Targets

We defined nine binary prediction targets for each index ED encounter, including seven primary acute-condition targets and two secondary disposition endpoints. The disposition endpoints were hospitalization and a composite critical outcome. Hospitalization was defined as admission from the ED to an inpatient ward. The critical outcome was defined as ICU admission within 12 hours of ED arrival or all-cause in-hospital mortality. These components were combined because each was infrequent and both represent severe clinical trajectories relevant to early ED risk stratification; similar composite outcomes have been used in previous ED risk-prediction studies^{26,27}.

The seven acute condition targets were sepsis, acute respiratory distress syndrome (ARDS), chronic obstructive pulmonary disease (COPD)/asthma exacerbation, acute coronary syndrome (ACS), viral pneumonia (VP), pulmonary embolism (PE), and acute kidney injury (AKI). These targets were selected as clinically important, time-sensitive conditions that may present with overlapping, nonspecific symptoms and physiological abnormalities at triage and may co-occur within the same encounter. AKI was included as a common and clinically consequential co-occurring condition, particularly in systemic and critical illness. This target set was intended as a proof-of-principle application of a framework that

can accommodate broader sets of non-mutually exclusive clinical risks. Across sites, target labels were harmonized using diagnosis codes and, where applicable, disposition, procedure, and care-location data. Detailed target definitions and coding rules are provided in **eTable 4**.

Evaluation and Statistical Analysis

All ED encounters meeting the eligibility criteria in each data source were used, so the study size was determined by the available data rather than by a target precision. Precision is limited by the rarest endpoint rather than by cohort size. Baseline characteristics were summarized by cohort using means with standard deviations or medians with interquartile ranges for continuous variables, and counts with percentages for categorical variables. Given the large cohort sizes, between-cohort differences were described without formal hypothesis testing as negligible differences could yield very small P values. Model performance was evaluated in the UMN internal validation set and the two independent external validation cohorts from BIDMC and Stanford. Discrimination was quantified using the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC), with AUPRC emphasized when interpreting low-prevalence targets²⁸. No resampling, class weighting, or synthetic minority oversampling was applied at any stage. The observed prevalence of each endpoint was preserved during training so that the predicted probabilities remain interpretable on the absolute scale and require no post hoc correction. Because disposition outcomes reflect operational and system-level influences in addition to patient characteristics, performance was summarized separately for the seven acute-condition targets and the two secondary disposition endpoints. For each target group, unweighted macro-averages were calculated by averaging target-specific metrics. Performance was additionally examined within strata of age band, sex, and ESI in each validation cohort. No fairness metric was specified in advance and no mitigation procedure was applied. Race and ethnicity were not extracted or harmonized across the three data sources and could not be examined, which is a limitation of this analysis rather than a property of the model.

Calibration was assessed graphically using calibration curves and quantitatively using the calibration slope (ideal value 1), calibration intercept (calibration-in-the-large; ideal value 0), and Brier score²⁹. These metrics were computed for the calibrated predictions and were summarized per target and as macro-averages for the acute-condition labels and disposition endpoints separately. Operating characteristics were calculated on the calibrated probability scale at the highest threshold meeting a pre-specified sensitivity of 90% for each target; sensitivity rather than positive predictive value was fixed because prevalence was below 6% for every target except hospitalization. Operating points at 95% sensitivity were additionally evaluated for sepsis and critical outcome. At each operating point, we report sensitivity, specificity, positive/negative predictive value, and the proportion of encounters exceeding the threshold. Decision-curve analysis was performed in all three cohorts to estimate net benefit relative to treat-all and treat-none strategies across a range of threshold probabilities.

MultiDiseasePred was compared with the National Early Warning Score (NEWS)³⁰ and five single-task machine-learning baselines: logistic regression, random forest³¹, LightGBM³², XGBoost³³ and histogram-based gradient boosting³⁴. Each single-task model was trained separately for each target using the same 12 variables, development data and calibration procedure as MultiDiseasePred. The full comparison is reported in **eTable 5**. Uncertainty was quantified using 1,000 bootstrap resamples to estimate 95% confidence intervals for each performance metric. Model comparisons were descriptive, and no formal paired hypothesis tests were performed. As a conservative descriptive criterion, differences were highlighted only when the corresponding 95% confidence intervals did not overlap. All analyses were conducted using R version 4.3.1 and Python version 3.9.11.

Results

Multicenter Cohort Characteristics

This study included 2,203,437 ED encounters across three health systems: 1,664,277 from UMN, 422,114 from BIDMC, and 117,046 from Stanford, after application of the cohort-selection criteria shown in **Figure 2**. The four analysis cohorts differed in demographic composition, recorded comorbidities, and outcome prevalence (**Table 1**), reflecting substantial between-cohort heterogeneity. Median age ranged from 49.9 years (IQR 33.2–67.8) in the UMN development cohort to 54.0 years in both the BIDMC (IQR 35.0–69.0) and Stanford (IQR 36.0–70.0) cohorts. Chronic pulmonary disease was recorded in 22.2% of the UMN development cohort compared with 9.5% at Stanford, and fluid and electrolyte disorders in 28.7% and 12.0%, respectively. Hospitalization prevalence ranged from 19.8% in the UMN internal validation cohort to 47.9% at BIDMC, and the critical outcome ranged from 0.4% to 6.6%. The prevalence of recorded comorbidities also varied markedly across cohorts (**eTable 6**).

The acute-condition targets likewise varied in prevalence across cohorts. Sepsis prevalence was similar in the UMN development, UMN internal validation, and BIDMC cohorts (2.5–2.7%) but lower at Stanford (0.7%), whereas ARDS remained rare across all cohorts (0.03–0.10%). Despite these prevalence differences, pairwise outcome-correlation patterns were broadly similar in the UMN development and BIDMC cohorts (**Figure 3**). Sepsis and AKI showed the strongest association in both cohorts, with weaker associations among sepsis, ACS, and AKI. Overall, correlations were modest (16 of 21 pairs at UMN and 19 of 21 at BIDMC had $|\rho| < 0.10$). The Stanford cohort was excluded from the correlation analysis because several condition-specific event counts were too low to support stable pairwise estimates.

Model Performance and Parsimony

The UMN internal validation set comprised 350,842 encounters from 2024–2025. The 12-variable MultiDiseasePred model achieved a macro-averaged AUROC of 0.900 (95% CI, 0.897–0.903) and a macro-averaged AUPRC of 0.213 (0.208–0.218) across the seven acute conditions. Across the two

disposition endpoints, the corresponding macro-averaged AUROC and AUPRC were 0.852 (0.849–0.856) and 0.307 (0.298–0.316), respectively (**Table 2**). Performance against NEWS and five single-task machine-learning baselines trained on the same variables is reported in **eTable 5**.

The endpoint-level performance for the MultiDiseasePred model in internal validation was summarized in **Table 2**. Across the nine targets, AUROC ranged from 0.790 (95% CI, 0.788–0.792) for hospitalization to 0.964 (0.961–0.968) for pulmonary embolism, whereas AUPRC ranged from 0.022 (0.003–0.040) for ARDS to 0.505 (0.501–0.509) for hospitalization. For sepsis, AUROC was 0.844 (0.840–0.848) and AUPRC 0.190 (0.182–0.198). For ARDS, which had a prevalence of 0.06% in the internal validation cohort, AUROC was 0.897 (0.877–0.916), whereas AUPRC was 0.022, approximately 38-fold above the prevalence-based no-skill baseline. Receiver operating characteristic and precision–recall curves for all nine endpoints are shown in **eFigures 1 and 2**.

In external validation, discrimination varied across targets and sites (**Table 3**). Across the seven acute-condition targets, the macro-averaged AUROC and AUPRC were 0.793 and 0.116 at BIDMC and 0.866 and 0.082 at Stanford, respectively. For the two disposition endpoints, the corresponding values were 0.699 and 0.426 at BIDMC and 0.710 and 0.260 at Stanford. Endpoint-level AUROCs ranged from 0.744 for pulmonary embolism to 0.848 for COPD/asthma exacerbation in BIDMC and from 0.787 for acute kidney injury to 0.932 for pulmonary embolism in Stanford. AUPRCs ranged from 0.008 for ARDS to 0.673 for hospitalization in BIDMC and from 0.004 for ARDS to 0.414 for hospitalization in Stanford, consistent with marked variation in target prevalence across sites. Endpoint-level results for all comparator models at each external site are provided in **eTable 5**, with additional results for the two disposition endpoints in **eTable 7**; corresponding curves are shown in **eFigures 3 and 4**.

The number of retained variables was determined using a criterion specified before inspection of the selection curve: the smallest value of k achieving at least 98% of the macro-AUROC obtained with all 62 candidates. This criterion selected 12 variables, yielding a macro-averaged AUROC of 0.912 compared with 0.928 for the full candidate set on the selection split (**Figure 4, eTable 9**). The marginal AUROC gain per additional variable declined approximately twenty-fold between $k = 2$ and $k = 15$, without a clear inflection point or plateau, the 12-variable set reflected the prespecified performance-retention criterion. Applying thresholds of 95% and 97% would have selected 7 and 10 variables, respectively. For descriptive comparison, the same feature-count sweep was repeated in each of the three health-system cohorts. In the internal validation set, the 12-variable model reached a macro-AUPRC of 0.213 (0.208–0.218) against 0.274 for the same architecture trained on all 62 candidates. Reductions were more pronounced for AUPRC among the rarer targets. For sepsis, AUROC was 0.844 (0.840–0.848) versus 0.873 and AUPRC 0.190 (0.182–0.198) versus 0.235; for pulmonary embolism, AUROC was 0.964 (0.961–0.968) versus 0.969 and AUPRC 0.282 (0.263–0.300) versus 0.411 (**eTable 11**). Thus, the parsimonious model retained most of the full model's discrimination as measured by AUROC but incurred larger reductions in precision–recall performance, particularly for rare targets.

Model Interpretation and Calibration

Expected-gradient attribution analysis showed distinct target-specific patterns across the nine prediction targets (**Figure 5**). Age had the highest mean absolute attribution overall (0.603, compared with 0.387 for diastolic blood pressure) and ranked first across all three training seeds. Attribution profiles differed across the respiratory targets: for ARDS, the largest attributions were observed for diastolic blood pressure (0.73), fluid and electrolyte disorders (0.61), and oxygen saturation (0.49), whereas viral pneumonia showed the largest attributions for oxygen saturation (0.47) and age (0.44). Chronic pulmonary disease had the largest attribution for COPD/asthma exacerbation (1.76, the largest target-specific value in the attribution matrix). For acute coronary syndrome, age had the largest attribution (1.08), followed by prior myocardial infarction (0.47). Pulmonary hypertension had relatively small attributions for eight targets (0.05–0.11) but the largest attribution for pulmonary embolism (0.69).

Two pre-existing comorbidity indicators accounted for a substantial proportion of discrimination for their corresponding targets (**eTable 12**). Pulmonary hypertension alone achieved an AUROC of 0.929 for pulmonary embolism, representing 92% of the full model's AUROC gain above 0.5. Chronic pulmonary disease alone achieved an AUROC of 0.883 for COPD/asthma exacerbation, representing 83% of the corresponding gain achieved by the full model. Because the comorbidity look-back period ended before the index encounter, these findings were not attributable to direct temporal leakage; however, they indicate that predictions for these two targets depended substantially on a single recorded historical feature.

In the UMN internal validation set, calibration curves generally showed close agreement between predicted and observed event rates across targets (**Figure 6**). For hospitalization, predicted probabilities extended to approximately 0.6 and closely followed the ideal calibration line. Similar patterns were observed for sepsis, viral pneumonia, and AKI, although calibration estimates for rare outcomes, including ARDS and the critical outcome, were based on comparatively few events. The mean calibration slope across the nine targets was 1.00, with target-specific values ranging from 0.93 to 1.20, and the macro-averaged Brier score was 0.027. Calibration intercepts ranged from -0.23 to +0.14 for eight targets. Viral pneumonia showed a larger negative intercept (-1.04), consistent with overprediction after its prevalence decreased from 0.90% in the development cohort to 0.38% in the internal validation set. Target-specific calibration metrics for all four cohorts are reported in **eTable 14**, with corresponding curves in **eFigures 6 and 7**. Decision-curve analysis showed positive net benefit relative to treat-all and treat-none strategies across the evaluated threshold ranges in all three validation cohorts (**eFigures 8 and 9**).

Using the calibrated probability scale, we evaluated target-specific operating points corresponding to 90% sensitivity. Across the seven acute-condition targets, macro-averaged specificity was 73.2%. For pulmonary embolism, positive predictive value was 10.6%, compared with a prevalence of 0.74%, at a specificity of 94.4%. The corresponding values were 5.5% versus 2.7% for sepsis at 57.6% specificity and 0.9% versus 0.4% for the composite critical outcome at 57.2% specificity. Positive predictive values

remained modest in absolute terms for targets with prevalence below 3%. Negative predictive value was 99.07% for AKI and exceeded 99.5% for the other six acute-condition targets, although these values should be interpreted in the context of their low baseline prevalence. The same thresholds were applied without modification to the two external cohorts, with results reported in **eTable 15**; target- and model-specific operating characteristics are provided in **eTable 16**.

To illustrate patient-level output, MultiDiseasePred was applied to a constructed, composite ED encounter involving a 67-year-old man presenting with fever, tachycardia, tachypnea, hypotension, and hypoxemia, with prior myocardial infarction, chronic pulmonary disease, renal disease and shortness of breath as the chief complaint (NEWS score 7; **Figure 7**). Using the 12-variable set, MultiDiseasePred generated calibrated probabilities, prevalence-adjusted log-lift ratios, and target-specific feature attributions across the nine prediction targets. Sepsis, AKI, and ACS had the highest prevalence-adjusted rankings. This example demonstrates the presentation and interpretability of model outputs rather than providing independent evidence of predictive validity or clinical benefit.

Discussion

In this large multicenter study, we developed and externally validated MultiDiseasePred, a general interpretable multi-task learning framework for simultaneous multi-condition risk prediction, and demonstrated its clinical utility in an emergency department triage application involving seven acute conditions and two disposition endpoints. By combining a shared expert pool with target-specific routing, the Mixture of Experts framework selectively shares predictive information across related targets while preserving target-specific patterns. Its calibrated probabilities, prevalence-adjusted cross-target summaries, and target-specific feature attributions provide a unified and interpretable risk profile before laboratory or imaging results are available. Performance in the internal validation set and two independent external health systems supports the feasibility of this approach across heterogeneous ED settings without local retraining or recalibration. Against single-task models trained using the same 12 variables, MultiDiseasePred was competitive but not uniformly superior across targets and sites. Its principal contribution therefore lies not in maximizing discrimination for every individual endpoint, but in generating a coordinated and interpretable nine-target risk profile in a single inference pass. Variation in cross-site performance and calibration, together with substantial dependence on individual historical variables for some targets, indicates that MultiDiseasePred should be viewed as a prioritization aid requiring local assessment and prospective workflow evaluation rather than as a deployment-ready diagnostic system²⁸.

Beyond the single-task baselines evaluated here, MultiDiseasePred also differs from established disease-specific assessment pathways. Many existing tools are designed to refine risk assessment once a particular condition has entered the differential diagnosis. For myocardial infarction, machine-learning approaches applied to cardiac troponin can estimate individualized risk rather than relying on a single universal threshold, but they still require biomarker results to be available³⁵. For sepsis, screening models that do not require laboratory values can be applied early, including at triage, but they remain

focused on a single syndrome³⁶. For pulmonary embolism, diagnostic evaluation is typically stepwise, with pretest clinical probability guiding the use of D-dimer testing and imaging³⁷. MultiDiseasePred addresses an earlier and broader triage question: among several plausible and potentially co-occurring acute conditions, which risks should be prioritized before definitive testing is available? Its discrimination was broadly comparable to that reported for dedicated disease-specific models, although these comparisons are indirect because published models were developed in different populations, often used laboratory or longitudinal data rather than triage variables alone, and in some cases addressed related but non-identical targets^{38–45}, with particular relevance for low-prevalence targets³⁸. By producing simultaneous calibrated risks for seven acute conditions and two disposition endpoints from the same triage-available inputs, MultiDiseasePred provides a unified risk profile that is more aligned with the early, nonspecific nature of ED decision-making than a sequence of separate disease-specific assessments.

Joint modeling of co-occurring or clinically related conditions is emerging across different data modalities. ProtAIDe-Dx, a deep joint-learning model trained on plasma proteomics from 17,187 individuals, generated simultaneous probabilistic diagnoses for six dementia-associated conditions and identified shared as well as disease-specific protein networks, illustrating how a single model can represent both common and condition-specific signals⁴⁶. Similarly, representation-learning approaches based on coded electronic health records have shown that one learned patient representation can support multiple downstream prediction tasks⁴⁷. These studies support the broader premise of multi-condition modeling, but their clinical settings differ from ED triage: proteomic testing requires a dedicated assay, and longitudinal EHR representation learning depends on accumulated historical records. MultiDiseasePred extends this principle to a low-latency emergency-care setting, where the relevant question must be addressed using information already available at triage and before laboratory or imaging results return.

The practical value of MultiDiseasePred lies in this early triage setting. Before laboratory or imaging results are available, the triage team faces a broad question: which plausible acute conditions should be prioritized? MultiDiseasePred uses only triage-available information to generate a calibrated, interpretable risk profile for seven acute conditions and two disposition endpoints in a single inference pass. For each target, the model reports not only the estimated risk but also the features that contributed to that estimate. Because all targets are modeled within a shared framework, the same patient feature can be interpreted across conditions, revealing whether it is broadly informative, condition-specific, or essentially irrelevant for a given target. This unified output is more aligned with early ED decision-making than a set of separate disease-specific models.

The parsimonious feature set reduces input requirements, although this simplification was accompanied by lower precision–recall performance for some rare targets compared with the 62-variable model. Excluding the ESI from the candidate predictors also limits circular reuse of clinician-assigned triage acuity. MultiDiseasePred is therefore intended as clinician-in-the-loop decision support rather than an

autonomous diagnostic or disposition system, and its calibrated probabilities and target-specific explanations may help clinicians organize multiple plausible risks and identify conditions warranting further assessment. However, ease of integration, usability, alert burden, and effects on clinical decision-making were not evaluated, and cross-site variation indicates that local performance and calibration should be assessed before implementation. Two practical points follow for implementation. Values that are missing at the moment of prediction are filled with the same development-cohort medians used during training, so a deployed instance behaves identically to the evaluated model rather than adopting a local imputation that was never tested, and values outside the prespecified physiological ranges are clipped to the nearest boundary as they were during development. The released software does not silently impute absent or non-numeric columns and instead reports which inputs are unusable, because an input error corrected at triage is preferable to a risk profile computed from a value the model never saw. Using the model requires no additional data entry and no expertise beyond routine triage practice, because all twelve variables are either recorded by the triage nurse at arrival or already present in the record. Interpretation of the resulting risk profile, including the variable attributions, is intended for clinicians rather than direct patient-facing communication.

Many emergency departments already use protocolized, nurse-initiated pathways for time-sensitive conditions such as sepsis and chest pain, in which standing orders allow selected early steps to begin before physician evaluation. A familiar example is the triage electrocardiogram obtained for patients reporting chest pain. If validated prospectively, MultiDiseasePred could support selected early components of these pathways, such as electrocardiography, laboratory testing, blood cultures, or lactate measurement before provider assessment⁴⁸. Because these actions are generally lower risk than definitive treatment or disposition decisions, they could in principle shorten time-to-results and time-to-treatment, including door-to-antibiotic time in sepsis and door-to-electrocardiogram time in suspected acute coronary syndrome^{49–51}. We emphasize that this remains a hypothesis to be tested. These triage-initiated steps concern early diagnostic workup rather than disposition, which is typically considered later in the encounter. Appropriate operating thresholds, the specific actions suitable for protocolized initiation, and the net effect on triage decisions, diagnostic efficiency, resource allocation, and patient outcomes should be established in prospective implementation studies co-designed with emergency clinicians. Displaying nine risks simultaneously also requires careful presentation: calibrated absolute probability should remain the primary output because absolute risk is what clinicians act on, whereas the log-lift ratio serves as a secondary aid for comparing risks across targets with different baseline prevalence. The divergence between discrimination and precision at Stanford illustrates the same issue: under prevalence shift, a well-ranked model can still yield few true positives per alert. Therefore, site-specific calibration assessment, subgroup evaluation (as shown in **eTable 17**, **eFigure 10**), and prospective workflow testing should precede implementation.

This study has several limitations. First, MultiDiseasePred was evaluated retrospectively, so its effect on ED decision-making, workflow efficiency, resource allocation, and patient outcomes is unknown; prospective silent validation and implementation studies are needed before clinical deployment. In

addition, our comparisons against single-task baselines evaluate the end-to-end modeling pipeline rather than the architecture in isolation. An ablation holding variables and training protocol fixed is reported in **eTable 18**, and the gain from averaging seeds rather than ensembling them in **eTable 19**, but the separate contributions of the shared expert pool, task-specific routing and expert dropout were not individually quantified, which we identify as important future work. Second, the variable budget was chosen against the seven acute conditions alone, with no vote from the two disposition endpoints; a set selected with disposition weighted equally would probably differ, and the reduction from 62 variables to 12 was not evaluated for the disposition endpoints. Third, two comorbidity flags account for most of the discrimination on their own endpoint, so for pulmonary embolism and COPD/asthma exacerbation the model is substantially re-expressing a single recorded history item. Fourth, the model was deliberately restricted to structured triage and historical variables to enable early prediction and therefore does not incorporate laboratory results, imaging, medication data, or unstructured clinical notes; future work could develop dynamic models that update risk estimates as additional information accrues during the encounter. Fifth, target labels were derived from routinely collected data and may be influenced by site-specific coding, documentation, and diagnostic practices. Sixth, missing data were handled using predefined rule-based and median imputation, and the effect of the imputation strategy was not re-examined for the final twelve-variable set; because missingness in emergency care may itself be informative, future analyses should examine missingness mechanisms and compare alternative imputation strategies. Finally, although external validation supported transportability across two independent health systems, local calibration, subgroup performance, fairness across demographic groups⁵², and clinician-facing workflow integration remain to be evaluated prospectively.

Declarations

Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

Y.J. curated the data, developed and ran the final models, produced the figures, and wrote the manuscript. K.X. and Y.W. developed the initial model and drafted the first version of the manuscript. C.C. verified the model and revised the manuscript. X.M., Z.Z., A.M., Y.C., G.S., R.A.D., N.E.I., D.A.W., M.L., M.A.P. and F.X. reviewed the code and revised the manuscript. F.X. conceived and supervised the study, interpreted the findings, and revised the manuscript. All authors read and approved the final manuscript. Y.J., K.X. and Y.W. contributed equally to this work.

Competing interests

The authors declare no competing interests.

Data availability

The MIMIC-IV-ED and MC-MED datasets analysed in this study are publicly available through PhysioNet at <https://physionet.org/content/mimiciv/3.1/> and <https://physionet.org/content/mc-med/1.0.0/>, subject to the credentialing and data use agreements of each resource. The University of Minnesota / M Health Fairview electronic health record data cannot be shared. They contain protected health information, and the institutional data use agreement and IRB approval under which they were obtained permit analysis only within the University of Minnesota secure computing environment, from which record-level data cannot be exported. Investigators seeking access should contact the University of Minnesota Institutional Review Board; access requires a separate institutional agreement and is granted only within that environment.

Code availability

MultiDiseasePred is released as an open-source Python package at <https://github.com/fengx13/MultiDiseasePred>, together with the scripts used for model development, training, calibration and evaluation and the source code for the demonstration dashboard.

Protocol and registration

No study protocol was prepared and the study was not registered.

Patient and public involvement

Patients and members of the public were not involved in the design, conduct, reporting, or dissemination of this study.

Ethics Declaration

The use of University of Minnesota / M Health Fairview electronic health record data was approved by the University of Minnesota Institutional Review Board (STUDY00023702), which waived the requirement for informed consent because the study was retrospective and used data collected in the course of routine clinical care.

Tables

Tables 1-3 are available in the Supplementary Files section.

References

1. Friedemann Smith, C. *et al.* Patient characteristics, serious disease diagnoses, and incidental findings in individuals with non-specific symptoms referred to the Suspected CANcer (SCAN) Pathway: a prospective cohort study in England. *Lancet Prim. Care* 1, 100019 (2025).
2. Müller-Riemenschneider, F. *et al.* Barriers to routine risk-score use for healthy primary care patients: survey and qualitative study: Survey and qualitative study. *Arch. Intern. Med.* 170, 719–724 (2010).
3. Singer, M. *et al.* The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* 315, 801–810 (2016).
4. Six, A. J., Backus, B. E. & Kelder, J. C. Chest pain in the emergency room: value of the HEART score. *Neth. Heart J.* 16, 191–196 (2008).
5. Xie, F. *et al.* Development and validation of an interpretable machine learning scoring tool for estimating time to emergency readmissions. *EClinicalMedicine* 45, 101315 (2022).
6. Rakers, M. M. *et al.* Availability of evidence for predictive machine learning algorithms in primary care: A systematic review: A systematic review. *JAMA Netw. Open* 7, e2432990 (2024).
7. Krones, F., Marikkar, U., Parsons, G., Szmul, A. & Mahdi, A. Review of multimodal machine learning approaches in healthcare. *Inf. Fusion* 114, None (2025).
8. Xie, F. *et al.* Development and assessment of an interpretable machine learning triage tool for estimating mortality after emergency admissions. *JAMA Netw. Open* 4, e2118467 (2021).
9. Draeger, L. *et al.* Barriers and facilitators to optimal sepsis care - a systematized review of healthcare professionals' perspectives. *BMC Health Serv. Res.* 25, 591 (2025).
10. Xie, P., Hu, Y., Li, J., Ma, Y. & Xiao, J. Unlocking the potential of real-time ICU mortality prediction: redefining risk assessment with continuous data recovery. *NPJ Digit. Med.* 8, 733 (2025).
11. Joseph, J. W. *et al.* Race and ethnicity and primary language in emergency department triage. *JAMA Netw. Open* 6, e2337557 (2023).
12. Patel, M. D. *et al.* Patient sex, racial and ethnic disparities in emergency department triage: A multi-site retrospective study. *Am. J. Emerg. Med.* 76, 29–35 (2024).
13. Caruana, R. Multitask learning. *Mach. Learn.* 28, 41–75 (1997).
14. Collins, G. S. *et al.* TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385, e078378 (2024).
15. Shazeer, N. *et al.* Outrageously large neural networks: The Sparsely-Gated Mixture-of-experts layer. *arXiv [cs.LG]* (2017) doi:10.48550/arXiv.1701.06538.
16. Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. Adaptive mixtures of local experts. *Neural Comput.* 3, 79–87 (1991).
17. Hazimeh, H. *et al.* DSelect-k: Differentiable selection in the Mixture of experts with applications to multi-task learning. *arXiv [cs.LG]* (2021) doi:10.48550/arXiv.2106.03760.

18. Zadrozny, B. & Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, New York, NY, USA, 2002). doi:10.1145/775047.775151.
19. Johnson, A. et al. MIMIC-IV-ED. PhysioNet <https://doi.org/10.13026/5NTK-KM72> (2023).
20. Xie, F. et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Sci. Data* 9, 658 (2022).
21. Kansal, A., Chen, E., Jin, T., Rajpurkar, P. & Kim, D. Multimodal Clinical Monitoring in the Emergency Department (MC-MED). PhysioNet <https://doi.org/10.13026/WVYW-G663> (2025).
22. Johnson, A. et al. MIMIC-IV. PhysioNet <https://doi.org/10.13026/KPB9-MT58> (2024).
23. Quan, H. et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care* 43, 1130–1139 (2005).
24. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.* 40, 373–383 (1987).
25. Elixhauser, A., Steiner, C., Harris, D. R. & Coffey, R. M. Comorbidity measures for use with administrative data. *Med. Care* 36, 8–27 (1998).
26. Raita, Y. et al. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit. Care* 23, 64 (2019).
27. Yen, C.-C., Ma, C.-Y. & Tsai, Y.-C. Interpretable machine learning models for predicting critical outcomes in patients with suspected urinary tract infection with positive urine culture. *Diagnostics (Basel)* 14, 1974 (2024).
28. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 10, e0118432 (2015).
29. Van Calster, B. et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 17, 230 (2019).
30. Williams, B. The National Early Warning Score: from concept to NHS implementation. *Clin. Med.* 22, 499–505 (2022).
31. Breiman, L. Random forests. *Mach. Learn.* 45, 5–32 (2001).
32. Ke, G. LightGBM: a highly efficient gradient boosting decision tree. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 3149–3157 (Curran Associates Inc, Red Hook, NY, USA, 2017).
33. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv [cs.LG]* (2016) doi:10.48550/arXiv.1603.02754.
34. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *arXiv [cs.LG]* (2012) doi:10.48550/arXiv.1201.0490.
35. Doudesis, D. et al. Machine learning for diagnosis of myocardial infarction using cardiac troponin concentrations. *Nat. Med.* 29, 1201–1210 (2023).
36. Jiang, S. et al. Development and validation of a screening tool for sepsis without laboratory results in the emergency department: a machine learning study. *EClinicalMedicine* 80, 103048 (2025).

37. Roussel, M. *et al.* D-Dimer thresholds for diagnosis of pulmonary embolism based on a single question: is it the most likely diagnosis? A prospective, multicentre, open-label, single-arm interventional study. *Lancet Respir. Med.* 14, 29–37 (2026).
38. Le, S. *et al.* Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). *J. Crit. Care* 60, 96–102 (2020).
39. Alkhamis, M. A., Al Jarallah, M., Attur, S. & Zubaid, M. Interpretable machine learning models for predicting in-hospital and 30 days adverse events in acute coronary syndrome patients in Kuwait. *Sci. Rep.* 14, 1243 (2024).
40. Rapuri, S., Gong, K., Harris, C. & Stevens, R. D. A machine learning model to identify pulmonary embolism in patients admitted to intensive care. *Comput. Biol. Med.* 204, 111548 (2026).
41. Abady, E. *et al.* Artificial intelligence-driven triage in pediatric emergency departments: Accuracy, bias, and impact on clinical outcomes: A narrative review. *Sage Open Pediatr.* 13, (2026).
42. Moor, M. *et al.* Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *EClinicalMedicine* 62, 102124 (2023).
43. Jin, Y. *et al.* Development and multicenter external validation of A Data-Driven Scoring System for Early and Rapid Identification of Sepsis in Emergency Departments. *Emergency Medicine* (2025).
44. Feng, Y. *et al.* Characterization of risk prediction models for acute kidney injury: A Systematic Review and Meta-analysis. *JAMA Netw. Open* 6, e2313359 (2023).
45. Liang, Q. *et al.* Severe acute kidney injury predicting model based on transcontinental databases: a single-centre prospective study. *BMJ Open* 12, e054092 (2022).
46. An, L. *et al.* A deep joint-learning proteomics model for diagnosis of six conditions associated with dementia. *Nat. Med.* 32, 1852–1864 (2026).
47. Guo, L. L. *et al.* A multi-center study on the adaptability of a shared foundation model for electronic health records. *NPJ Digit. Med.* 7, 171 (2024).
48. Burgess, L., Kynoch, K., Theobald, K. & Keogh, S. The effectiveness of nurse-initiated interventions in the Emergency Department: A systematic review. *Australas. Emerg. Care* 24, 248–254 (2021).
49. Seymour, C. W. *et al.* Time to treatment and mortality during mandated emergency care for sepsis. *N. Engl. J. Med.* 376, 2235–2244 (2017).
50. Coopersmith, C. M. *et al.* The surviving sepsis campaign: Research priorities for Coronavirus disease 2019 in critical illness. *Crit. Care Med.* 49, 598–622 (2021).
51. Gulati, M. *et al.* 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR guideline for the evaluation and diagnosis of Chest Pain: Executive summary: A report of the American college of cardiology/American heart association joint committee on clinical practice guidelines. *Circulation* 144, e368–e454 (2021).
52. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453 (2019).

Figures

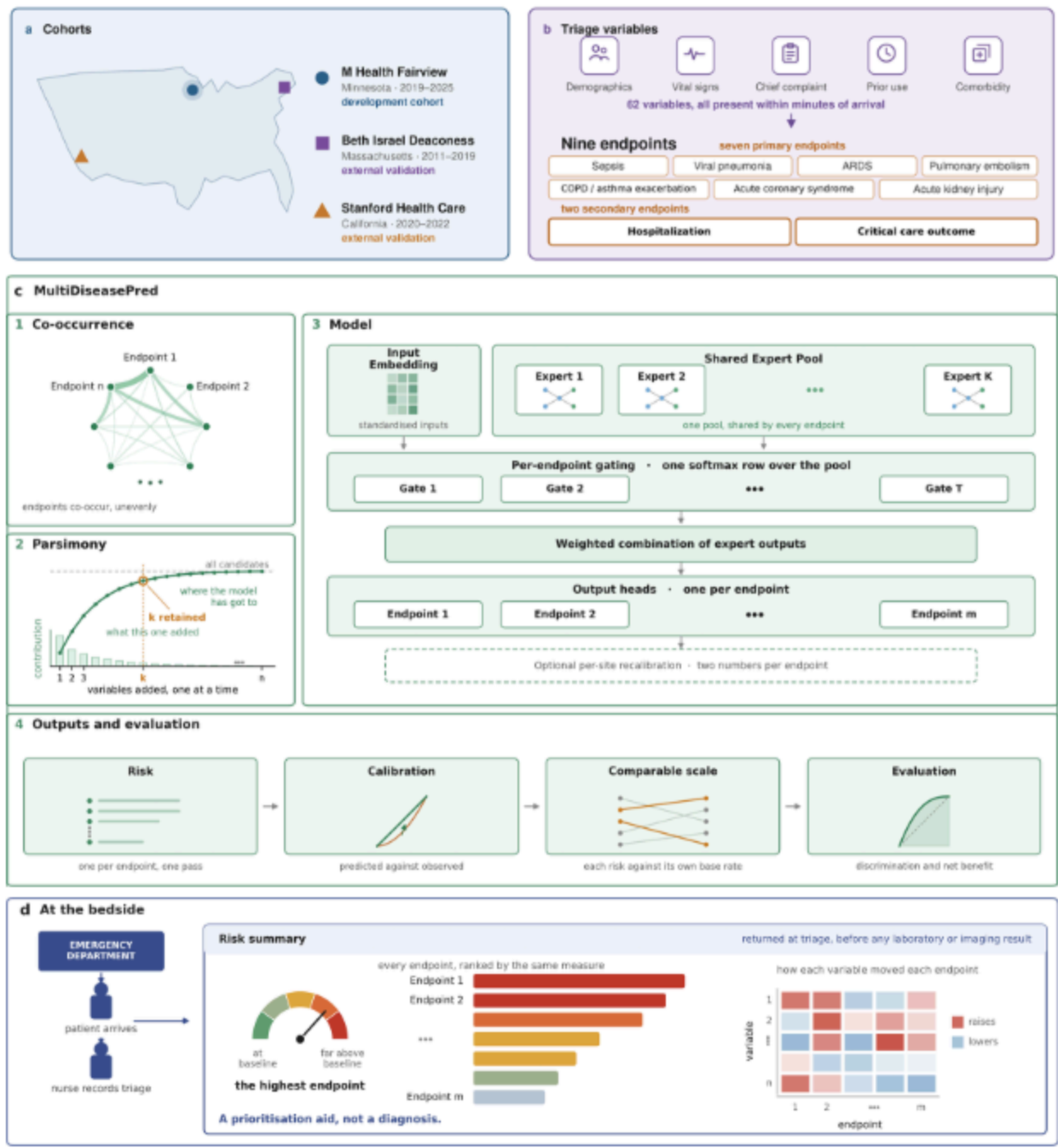


Figure 1

Development, validation, and clinical use of MultiDiseasePred.

Notes: (a) The three contributing health systems, with the period each contributed and its role: M Health Fairview / University of Minnesota for development and internal validation, and Beth Israel Deaconess Medical Center and Stanford Health Care for external validation. (b) The 62 candidate variables, grouped into five domains, all recorded within minutes of arrival, and the nine prediction targets: seven acute conditions treated as primary endpoints and two disposition endpoints treated as secondary. (c) The

framework in four steps. c1, the seven acute conditions co-occur unevenly, which is what motivates a shared representation rather than nine independent models. c2, variables are added one at a time by greedy forward selection, and the number retained is fixed by a rule specified before the curve was examined. c3, the twelve retained variables are standardized and passed to a shared pool of experts; a per-endpoint gate takes one softmax row over that pool, the weighted combination forms a target-specific representation, and one output head per endpoint produces all nine risks in a single forward pass, with optional per-site recalibration of two numbers per endpoint, shown as optional and not used in the reported external validation (Table 3). c4, the outputs are one calibrated risk per endpoint from that single pass, each expressed relative to its own base rate so the nine are comparable, and evaluated for discrimination and net benefit. (d) At the bedside the risk summary is returned at triage, before any laboratory or imaging result: endpoints are ranked by lift over their own base rates rather than by absolute probability, and each variable’s contribution to each endpoint is shown. The output is a prioritization aid, not a diagnosis.

Abbreviations: UMN, University of Minnesota; BIDMC, Beth Israel Deaconess Medical Center; AUROC, area under the receiver operating characteristic curve; NEWS, National Early Warning Score; EG, expected gradient.



Figure 2

Study cohort derivation.

Notes: Inclusion and exclusion criteria and the resulting analysis cohorts at the three participating health systems. The UMN cohort was split temporally into a development cohort (2019–2023) and an internal validation set (2024–2025); within the development cohort a random 20% was held out for variable selection. The BIDMC and Stanford cohorts were reserved in their entirety for external validation. All counts are taken from the analysis files; a screened count is shown only where the source extract could be identified and re-counted.

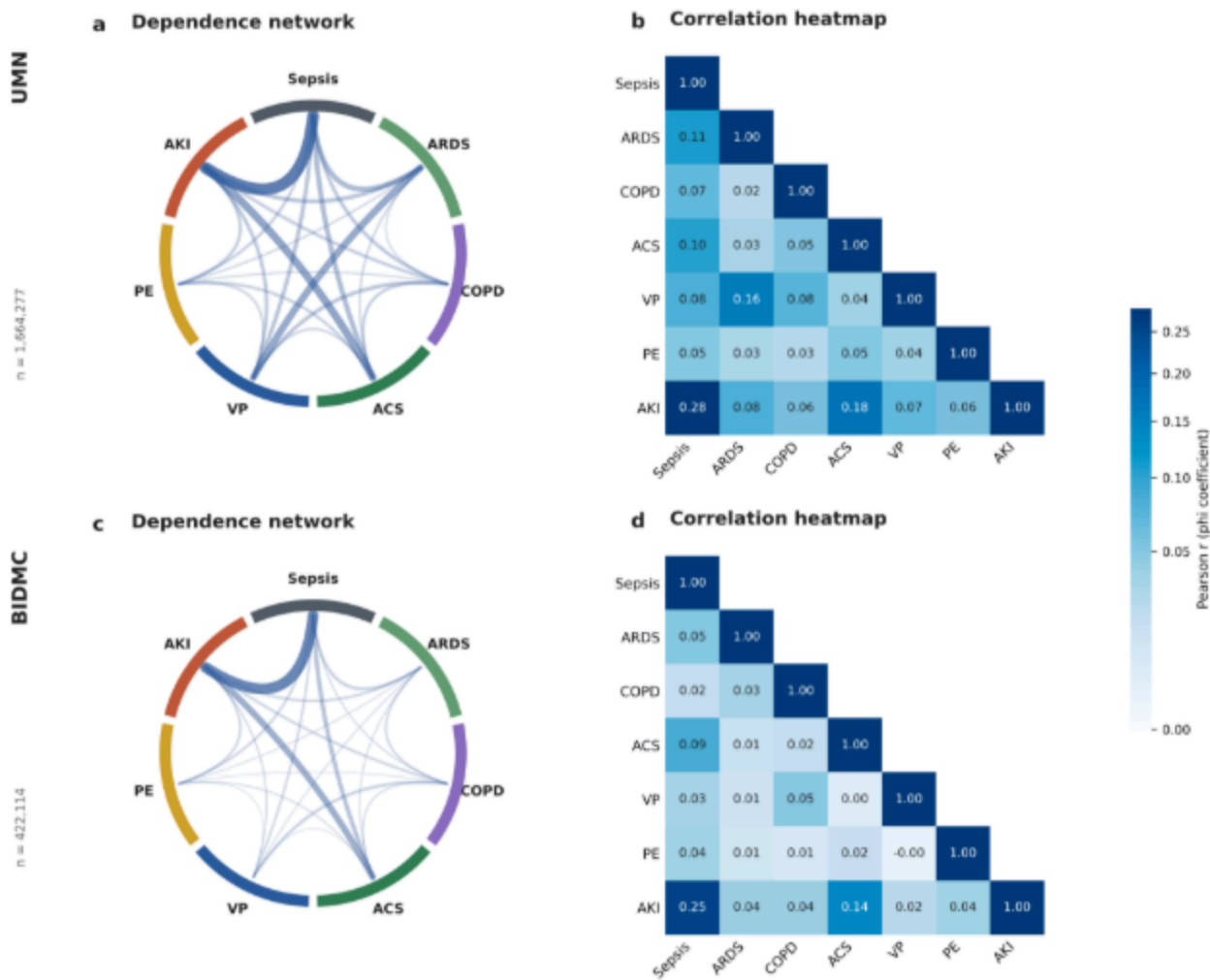


Figure 3

Pairwise correlation structure of the seven acute conditions in the UMN and BIDMC cohorts.

Notes: Each row shows the co-occurrence structure of the seven acute conditions, sepsis, acute respiratory distress syndrome (ARDS), COPD/asthma exacerbation, acute coronary syndrome (ACS), viral pneumonia (VP), pulmonary embolism (PE), and acute kidney injury (AKI), in the UMN development cohort (top, a–b) and the BIDMC external cohort (bottom, c–d). Panels a and c present dependence networks as chord diagrams, in which each perimeter arc represents one condition and each ribbon connects a pair of conditions; ribbon width and color intensity are both proportional to the pairwise correlation between the two conditions. Panels b and d present the corresponding lower-triangular correlation matrices, with the Pearson correlation coefficient (ϕ coefficient for binary outcomes) annotated in each cell. For two binary (0/1) condition indicators, the ϕ coefficient equals the Pearson correlation between them and was computed from the 2×2 co-occurrence table as $\phi = (n_{11} \cdot n_{00} - n_{10} \cdot n_{01}) / \sqrt{[(n_{11} + n_{10})(n_{01} + n_{00})(n_{11} + n_{01})(n_{10} + n_{00})]}$, where n_{11} is the number of encounters positive for both conditions, n_{10} and n_{01} the numbers positive for only one, and n_{00} the number positive for neither. Correlations were computed between binary outcome indicators across all index ED encounters in each cohort. Color encodes the pairwise correlation on a common scale across both panels and

cohorts; a square-root color scale is used to improve the visibility of weak correlations. The strongest association in both cohorts was between sepsis and AKI, with additional correlations among sepsis, ACS, and AKI.

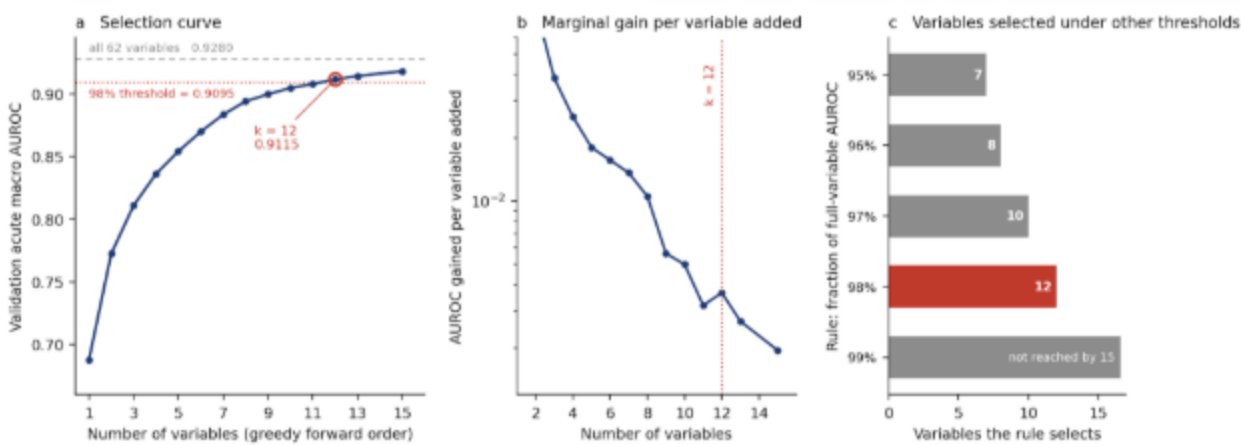


Figure 4

Variable selection by greedy forward selection.

Notes: (a) Macro-averaged AUROC over the seven acute conditions, computed on the 20% split of the UMN development cohort held out for variable selection, as variables are added one at a time by greedy forward selection; the variable added at each step is given in eTable 1. The dashed grey line is the AUROC obtained with all 62 candidate variables (0.9280) and the dotted red line is 98% of it (0.9095), the threshold fixed before the curve was examined; the smallest k reaching that threshold is 12 (0.9115). (b) The AUROC gained per variable added, on a logarithmic axis. The gain decays approximately twenty-fold between k = 2 and k = 15 with no kink and no plateau, so k = 12 is a disclosed choice rather than an inflection point in the data; the single step up, at k = 11 to 12, is smaller than the spread across training seeds. (c) The number of variables the same rule would select at other thresholds: 95% returns 7 variables, 96% returns 8, 97% returns 10, and 99% is not reached within the 15 variables the greedy sweep recorded. The full curve is tabulated in eTable 9, and the same sweep repeated in all three cohorts is shown in eFigure 5 and eTable 10.



Figure 5

Expected-gradient attribution of the twelve retained variables across the nine prediction targets.

Notes: Left, the mean absolute expected gradient of each of the twelve retained variables, averaged over the seven acute conditions, with variables ordered by that quantity; the color block at the far left marks the variable’s domain (purple, demographic; green, triage vital sign; orange, comorbidity; red, chief complaint); whiskers give the spread across three training seeds, and bar color indicates whether the variable’s typical value at triage lies below (blue) or above (red) the baseline used for the attribution. Right, the same quantity computed separately for each of the nine targets and printed in every cell; color is normalized within each row, so a dark cell marks the endpoint a variable matters most for rather than a large absolute value. The rank column gives the variable’s position in each of the three seeds, and the bars above the matrix give the total attribution received by each endpoint. Attributions are expected gradients on the logit scale for the twelve-variable model. Chronic pulmonary disease ranges from 0.06 for hospitalization to 1.76 for COPD/asthma exacerbation, a thirty-fold spread, and illustrates a variable the model uses for one endpoint and effectively ignores for the rest. The underlying values are tabulated in eTable 13.

Abbreviations: ACS, acute coronary syndrome; ARDS, acute respiratory distress syndrome; BP, blood pressure; COPD, chronic obstructive pulmonary disease; EG, expected gradient; MI, myocardial infarction.

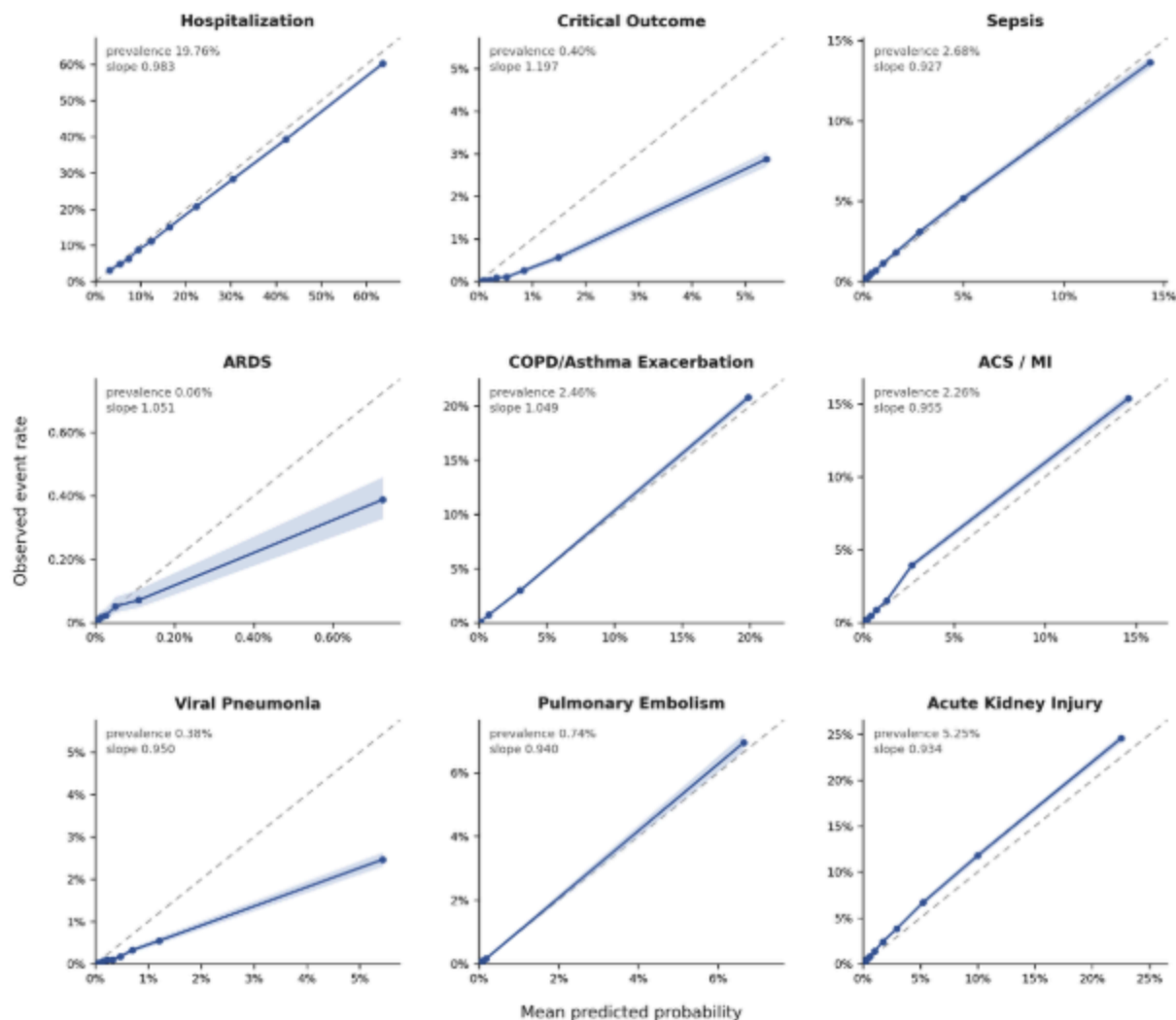


Figure 6

Non-parametric empirical calibration curves across the nine prediction targets.

Notes: Reliability plots map the framework's calibrated risk predictions (x-axis) against the observed event proportion (y-axis) across ten equal-sized bins of predicted risk in the UMN internal validation set ($n = 350,842$). The solid line traces observed against predicted risk and the shaded band is the Wilson 95% interval on the observed proportion; the dashed diagonal represents perfect calibration. Panels cover both the high-prevalence disposition target (hospitalization) and the low-prevalence acute syndromes (sepsis, ARDS, viral pneumonia, pulmonary embolism, COPD/asthma exacerbation, acute coronary syndrome, and acute kidney injury), as well as the composite critical outcome. Axes differ by panel because prevalence spans three orders of magnitude.

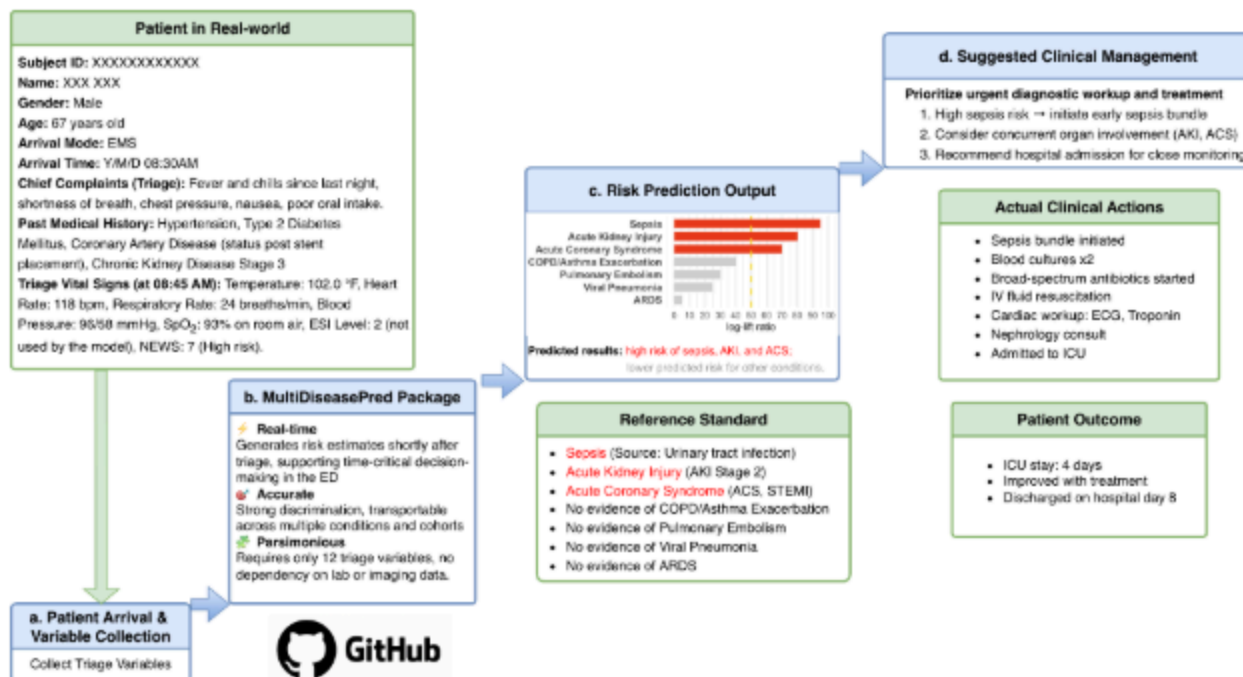


Figure 7

End-to-end real-time inference and clinical decision support for a representative composite patient encounter.

Notes: The operational pipeline is illustrated using a constructed, representative composite case of a 67-year-old male presenting with acute, non-specific symptoms spanning several organ systems; this case was designed to demonstrate the inference pipeline and does not correspond to any single real patient. The framework uses the 12 routinely available variables (panel a) to execute real-time inference (panel b). Predicted risks are contextualized relative to each target's population baseline prevalence using localized log-lift ratios (panel c), which rank sepsis, acute kidney injury (AKI), and acute coronary syndrome (ACS) as the highest-priority emergency hazards. The illustrative downstream course shown in panel d, including the reference diagnoses and representative interventions (sepsis bundle initiation and specialist consults), was constructed to be concordant with the initial triage risk profile. The subject identifier and presentation date and time shown are fictitious placeholders that do not correspond to any real individual. Triage acuity is shown as part of the recorded presentation and is annotated in the figure as not used by the model.

Supplementary Files

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

- [03SupplementaryInformation.pdf](#)
- [Table.docx](#)