Supplemental Online Content

Sanchez-Pinto LN, Bennett TD, DeWitt PE, et al; Society of Critical Care Medicine Pediatric Sepsis Definition Task Force. Development and validation of the Phoenix criteria for pediatric sepsis and septic shock. *JAMA.* Published online January 21, 2024. doi:10.1001/jama.2024.0196

This supplemental material has been provided by the authors to give readers additional information about their work.

eAppendix 1. Supplemental methods

Overview

We first established a centralized, multi-center database, including data from pediatric emergency department (ED), inpatient, and intensive care units (ICU), in a HIPAA-compliant Google Cloud environment. We used these data to identify the best performing pediatric organ dysfunction criteria for each individual organ dysfunction in differently resourced settings and care environments. We used the best performing organ dysfunction criteria to develop and validate novel criteria for pediatric sepsis and septic shock. We did this using a stacked regression approach,[1,2] where the best performing organ dysfunction criteria in children with suspected infection were the candidate component models, the final stacked model was based on the component models, and the novel sepsis criteria were integer-based versions of the final model. This work was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) grant number R01HD105939 to TDB and LNS-P. We reported this study using the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines for prediction model development and validation.[3]

Pipeline for Data Harmonization, Data Quality, and Data Analysis

The data pipeline developed for this project is shown in eFigure 1. The seven sites in the development dataset, as well as the external validation site in the U.S., extracted, de-identified, and transferred EHR data using a schema that we provided. The two international sites used for external, secondary validation had different schemas. These were the Pediatric Intensive Care Database (PICDB) from Hangzhou, China,[4] which is a publicly available dataset of pediatric ICU patients only, and the Kenya Medical Research Institute (KEMRI) pediatric dataset from Kilifi, Kenya, which is a clinical registry with limited longitudinal data. From both data sources, we extracted the variables that matched our schema. From the derivation and interval validation sites (HRS 1-5, LRS 1-2), we held out 25% of the data as an untouched internal validation test set. Prior to submission, sites de-identified data by calculating interval durations between admission and the events and removing complete dates and patient identifiers. The exception to this was the data from Kenya which contained actual dates. This was made possible with a data user agreement that allowed for the submission of a coded, limited dataset. Upon receiving each site's data, the study team loaded it into a cloud-based relational database. All processes were automated using Linux bash scripts, GNU Make, and GNU Parallel along with versioning with git to ensure that the entire pipeline was reproducible.[5]

Data Harmonization

The study database schema had individual tables containing information for patients (sex, race/ethnicity), encounters (start and end times, age, disposition, etc.), admission/discharge/transfer (transfer type, hospital locations, etc.), observation and intervention events (vitals, organ support, etc.), tests (laboratory and microbiology test results), medication administration (medication type and dose, administration action, etc.), input and outputs (fluid volume administered, urine output, etc.) and diagnosis codes. All tables, except for the patients table, had timestamps for each observation or event that was harmonized to minutes from the start of the encounter. The primary and foreign key across all tables was the study encounter identifier. Once all tables conformed to the study relational database schema, individual variables were harmonized to the pre-specified study-specific variable names, data types, and units of measurement (https://github.com/CU-DBMI-Peds/phoenix_sepsis_criteria, we will make this repository public on the day of publication). Particular emphasis was placed on harmonizing laboratory tests, vitals, observations, interventions, and medications needed to calculate the individual subscores of the eight previously validated organ dysfunctions scores and criteria used in the analysis (eTable 2), as well as other variables needed to ascertain the patients' infection status and study outcomes. In each case, a

mapping to the source values was maintained and used to stratify variables during the data quality assurance process.

Data Quality and ad hoc *Variable Curation*

Once harmonized, we performed multiple data quality assurance checks using the Kahn framework and queried sites as necessary.[6] These included checking the files and individual variables for conformance, completeness, and plausibility. Clinicians on the study team identified valid ranges and units for each variable and we performed additional *ad hoc* variable curation when necessary. Examples of this *ad hoc* curation included:

- Transforming variables to the correct unit for the organ dysfunction scoring (e.g. transform lactate in mg/dL to mmol/L)
- Changing numeric formats (e.g. "12,6" in decimal comma notation from some sites to "12.6" in decimal period notation)
- Transforming numeric values to allow for computation (e.g. transforming "platelet count <5" by multiplying the integer 5 by a factor of 0.9 to be able to compute a value of 4.5 in the organ dysfunction scoring)
- Calculating derivative values (e.g. if no mean arterial pressure was recorded when there was a systolic and diastolic blood pressure, one was calculated using 1/3*systolic + 2/3*diastolic pressure in mmHg)
- Mapping non-GCS neurologic assessments like the Adelaide pediatric coma scale to GCS when GCS was not recorded.[7]
- Generating a boolean value for presence or absence of an intervention (e.g. mechanical ventilation, continuous renal replacement therapy [CRRT], ECMO, etc.) using variables and values commonly associated with the interventions (e.g. positive end expiratory pressure and mean airway pressure for mechanical ventilation, replacement fluid rate for CRRT, etc.)
- Generating new label variables like "suspected infection", which was based on the presence of a systemic antimicrobial (i.e. antibacterials, antimalarials, antimycotics, antivirals, antimycobacterials, or anthelmintics) and microbiological testing (bacterial and fungal cultures, viral testing, parasite testing, etc.) within the same 24 hour period.

Data quality checks were performed by analyzing proportions (e.g. percentage of encounters with a given variable recorded), distribution of values (median, range, etc.), as well as visualizations values (box and violin plot visualizations using R Shiny). Individual variables were checked by stratifying by source name and study site and comparisons of proportions and distributions were performed across these strata to identify outliers. Potentially erroneous or incomplete data were flagged and sites queried in an interactive manner.

Data Framework and Handling of Missing Data

Once the data was harmonized and had undergone quality assurance, it was merged into a single dataset. We then established the data framework as a "time course" multivariable time series. In this time course dataset, every minute timestamp for all observations, interventions, results, medications, etc. was included and filled with known values. Encounters with no observation or intervention events (3.0% of the total) were excluded from the dataset at this step. In cases where two values for the same variable in the same encounter were recorded at the exact same minute, we implemented a "tie breaker" approach in which the worse value (depending on the organ dysfunction scoring system in which it was used [e.g. lowest platelet level in most scoring systems]) was kept, since this would reflect the most likely scenario in a real-time organ dysfunction scoring scenario. We used 'last observation carried forward' (LOCF) for physiologically appropriate time windows. This was a pragmatic choice, as LOCF approximates real-time dynamic data more realistically than other imputation methods.[8] Missing medications and

interventions were assumed to not have been used, similar to the approach used by the adult Sepsis-3 study and subsequent validation studies.[9,10] Physiologically appropriate time windows for LOCF included: (1) entire encounter (patient weight); (2) 24 hours (white blood cell count, troponin, thyroxine, prothrombin time, partial thromboplastin time , platelets, INR, hemoglobin, gamma-glutamyl transferase, fibrinogen, D-dimer, creatinine, blood urea nitrogen, bilirubin direct and total, aspartate aminotransferase, alanine aminotransferase, absolute neutrophil count, and absolute lymphocyte count); (3) 12 hours (Glasgow Coma Scale score, inotropes and vasopressors, glucose, continuous renal replacement therapy, ECMO, and prolonged capillary refill time); and (4) 6 hours (base deficit, PaO_2, PCO_2, pH, lactate, pupils, blood pressure, temperature, pulse, respiratory rate, SpO_2, O_2 flow, non-invasive and invasive ventilation, and FiO_2).

Organ Dysfunction Scoring

Once the time course data were filled with known values, LOCF values, or completely missing values that were treated as non-additive to organ dysfunction scoring, we calculated the subscores for all eight organ dysfunction scores and criteria used in the study (eTable 2). The calculation was done using a conservative approach such that values would only be additive for a given subscore or total score if they were recorded at (or were LOCF to) the same time. Calculation of PaO_2/FiO_2 ratio, SpO_2/FiO_2 ratio, oxygenation index, and oxygen saturation index required that the PaO_2 or SpO_2 were recorded at the same time or *after* the FiO_2 and mean airway pressure. We used normal values for age and sex to estimate the baseline creatinine using published reference values for scores that required it.[11] We calculated the organ dysfunction scores and criteria using a pragmatic approach by including variables easily ascertained in structured EHR data and excluding variables that are rare or difficult to ascertain using structured EHR data. The modifications to the scores and criteria were as follows:

- IPSCC Organ Dysfunction Criteria:
    - We did not include the criteria "Despite administration of isotonic intravenous fluid bolus greater than or equal to 40 mL/kg in 1 hour" for cardiovascular dysfunction given the complexity and associated inaccuracy in ascertaining this using EHR data. In addition, the Task Force requested that we not use fluid administration in our final criteria because of the context-specific nature of fluid management in sepsis.[12] We also did not determine core-to-peripheral temperature gap > 3°C for cardiovascular dysfunction as this is rarely recorded in EHRs. We used the remaining cardiovascular dysfunction criteria including systolic blood pressure, vasoactive agents, acidosis, lactatemia, oliguria, and capillary refill.
    - We did not exclude cyanotic heart disease or preexisting lung disease for respiratory dysfunction, as this information would not reliably be available in EHRs for patients who have not been cared for at an institution previously. We used the remaining criteria including hypoxemia, hypercapnia, and mechanical ventilation.
    - We did not calculate acute change in mental status with a decrease in GCS equal or greater to 3 points from abnormal baseline for neurologic dysfunction, as baseline neurologic function is infrequently recorded. We used the criteria based on a fixed GCS cutoff.
    - We did not calculate a decline of 50% in platelet count from the highest value recorded over the past 3 days for hematologic dysfunction. We used the criteria based on fixed platelet count and INR cutoffs.
    - The renal and hepatic criteria were implemented as pre-specified.
- PODIUM Criteria:
    - We did not include ventricular assist devices, cardiac arrest, or echocardiographic estimation of left ventricular ejection fraction, nor did we exclude children with underlying cyanotic congenital heart disease and cardiopulmonary bypass during

the episode of care for cardiovascular dysfunction. We used the remaining criteria including veno-arterial ECMO, heart rate, systolic blood pressure, vasoactive agents, lactatemia, and troponin.
    - We did not calculate an eGFR decrease to <35 mL/min/1.73m^2 or fluid overload equal to or greater than 20% by 48 to 72 hours given the focus on the first 24 hours of admission. We used the remaining criteria including urine output, CRRT, and serum creatinine.
    - We did not calculate gastrointestinal dysfunction because none of the variables (bowel perforation, *pneumatosis intestinalis*, ischemia present on gross inspection or by radiologic imaging, or sloughing of gut) are routinely captured as a structured data element EHRs.
    - We operationalized hepatic encephalopathy as a GCS equal to or less than 8 in patients with biochemical evidence of hepatic dysfunction. We did not exclude patients with evidence of chronic liver disease. We used the remaining criteria including biochemical evidence of acute liver injury and elevated INR.
    - We did not curate serum cortisol levels pre- and post-ACTH stimulation test for endocrine dysfunction. Only a small fraction of patients at a small number of sites had serum total thyroxine measured. We used the criteria based on blood glucose.
    - We did not include CD4+ T-lymphocyte count, CD4+ T-lymphocyte percentage of total lymphocytes, or monocyte HLA-DR expression for immunologic dysfunction. We used the criteria based on absolute neutrophil and lymphocyte counts.
    - The respiratory, coagulation, hematologic, and neurologic criteria were implemented as pre-specified.
- Proulx Criteria
    - We did not include cardiac arrest in cardiovascular dysfunction. We used the rest of the criteria based on systolic blood pressure, heart rate, acidosis, and vasoactive agents.
    - We did not include gastroduodenal bleeding or gastric/duodenal surgery for gastrointestinal dysfunction.
    - We did not exclude icterus due to breastfeeding for hepatic dysfunction. We used the criteria based on total bilirubin level.
    - We did not exclude pre-existing renal disease for renal dysfunction. We used the rest of the criteria based on blood urea nitrogen, serum creatinine, and dialysis.
    - We did not exclude mechanical ventilation without restriction to >24 hours in postoperative patients nor cyanotic congenital heart disease for respiratory dysfunction. We used the rest of the criteria based on tachypnea, hypercapnea, and hypoxemia.
    - The hematologic and neurologic criteria were implemented as pre-specified.

Stacked Regression

One way to make prediction robust is to combine or average the predictive power of many models or information sources. There are several methodological pathways to combine information in predictive models, e.g., bagging methods such as Bayesian model averaging, ensemble learning, or stacked regression, boosting methods such as ADABOOST or XGBOOST, or methods where the parameters at different layers of the model are jointly estimated such as deep learning with neural networks. Of these, the most transparent methods where there can be simple and direct relationships between predictions and the weighting or contribution of predictors are the bagging class of methods. And, among bagging methods, stacked regression or generalization can be made to be particularly interpretable and transparent using, e.g., linear or logistic regression, while also allowing for several regularization or model selection approaches.

Model stacking is an appropriate approach for the current work for several reasons. First, it allows for robust, accurate, and interpretable evaluation of the underlying component models (in this case, organ dysfunction subcomponents).[13] Each organ dysfunction subcomponent has its own parameter in the stacked model so that the relative impact of organ dysfunctions can be estimated. Second, the stacked model never performs worse than the most accurate component model.[2]

Stacked regression begins with a set of models (i.e. organ dysfunction scores) with differing input variables that all make the same prediction (probability of mortality, in this case). The outputs (predictions) of these models are then stacked as input variables for a second regression model that also predicts mortality. We are developing sepsis criteria by "stacking" organ dysfunction score subcomponents. Mortality is the target outcome, as it was in Sepsis-3. Many stacked layers are possible, but each layer must predict the same feature in the same units. In this way, a stacked regression takes other component models as covariates and estimates the regression weights — the relative contribution of each respective model's prediction to the overall prediction — in accordance with the various models' predictive power.[2] Stacked regression is similar to deep learning but is more transparent, easier to interpret and decompose, and often requires substantially less data to estimate. Because we are using well-established and highly interpretable organ dysfunction model subcomponents, we anticipated correlation between the component (organ dysfunction) models and their variables. This is expected and stacked regression approaches generally assume this type of correlation is present. To manage the correlation between component models, we regularized the top-level model to limit overfitting and reduce or eliminate redundancy. Because they are highly interpretable, we used regularized regression models in the top level. We first used ridge regularized (L2) logistic regression as the top-level stacked model. Ridge regression at the top level will tend to reduce similar component models' weights equally without eliminating any models. Ridge models tolerate correlated inputs better than another commonly used form of penalized regression, L1 or the least absolute shrinkage and selection operator (LASSO). LASSO eliminates redundant component models in a rank-ordered fashion according to their predictive power. This led to the most parsimonious models. However, when component models are correlated and have similar predictive power, LASSO effectively randomly selects which to eliminate. By comparing ridge and LASSO regularized models, we gained insight into which component models were producing unique versus redundant predictive power. Finally, we evaluated elastic net regularization, which balances ridge and LASSO regularization, but did not find benefit beyond ridge or LASSO. This approach allowed us to directly quantify and understand which component models were contributing to the mortality prediction.

Bootstrapping and Cross-validation

Bootstrapping was used to create confidence intervals for AUPRC point estimates when identifying the best-performing organ subscores for each organ system. Ten-fold cross-validation was used to select the regularization parameter lambda in the stacked models that minimized deviance for each value of alpha (0 = ridge, 1 = LASSO, between 0 and 1 = elastic net). For each model and value of alpha, we selected the largest value of lambda such that the deviance was within one standard deviation of the minimum deviance value. This was to encourage parsimony in the LASSO and elastic net models.

Software

For data processing, statistical analysis, and data visualization we used:

- GNU Parallel[5]
- Google BigQuery (https://cloud.google.com/bigquery/)
- R Version 4.3.1 (R Foundation for Statistical Computing, Vienna, Austria)

Python Version 3.10.12 (https://www.python.org)

eTable 1. Site Characteristics

City	Hospital	Number of years	EHR
Baltimore	JHCC	5	Epic
Chicago	Lurie	8	Epic
Denver	CHCO	10	Epic
Dhaka	ICDDR,B	11	local
Hangzhou	Zhejiang	9	n/a^
Medellin	HGM	11	SAP
Nairobi/Kilifi	KEMRI	10	KIDMS
Philadelphia	CHOP	10	Epic
Pittsburgh	CHP	7	Cerner
Seattle	SCH	10	Cerner*

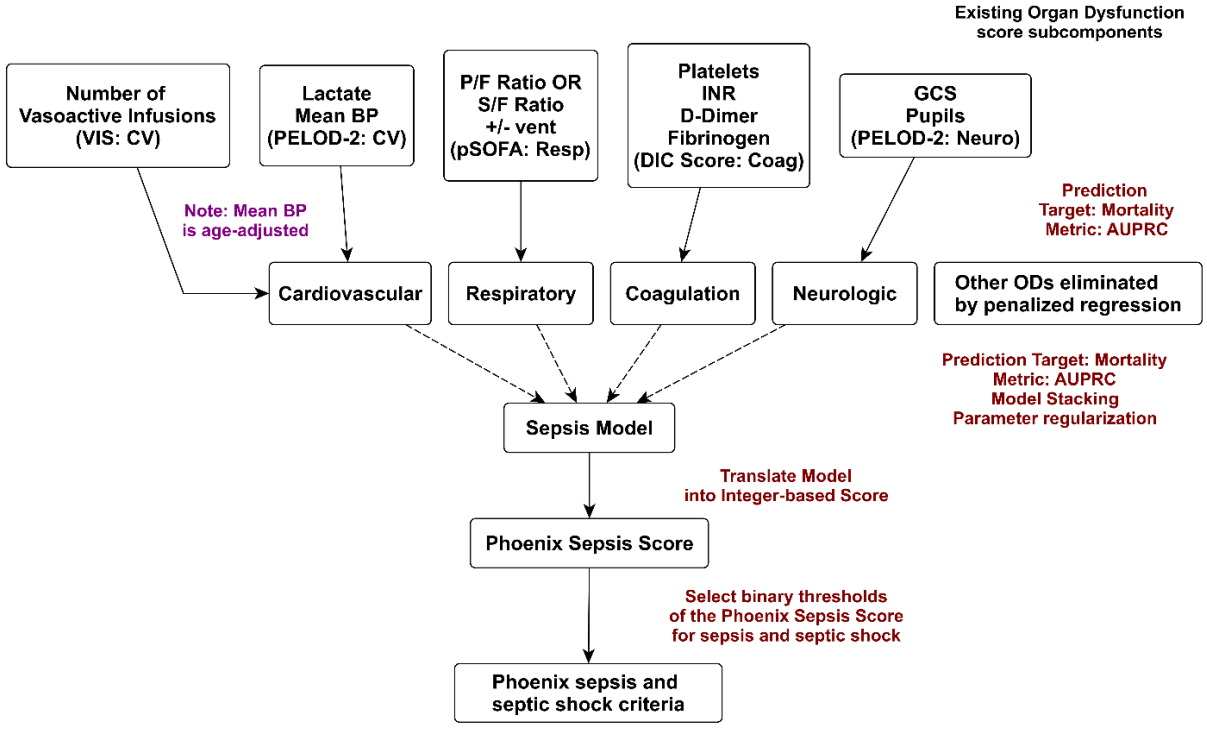
CHCO, Children's Hospital Colorado, USA; *CHOP*, Children's Hospital of Philadelphia, USA; *CHP*, University of Pittsburgh Medical Center Children's Hospital of Pittsburgh, USA; *EHR*, Electronic Health Record; *HGM*, Hospital General de Medellin, Colombia; *ICDDR, B*, International Centre for Diarrheal Disease Research, Bangladesh; *JHCC*, Johns Hopkins Children's Center, USA; *KEMRI*, Kenya Medical Research Institute; *KIDMS*, Kilifi Integrated Data Management System; *Lurie*, Ann & Robert H. Lurie Children's Hospital, USA; *SCH*, Seattle Children's Hospital, USA; *Zhejiang*, Children's Hospital, Zhejiang University (ICU only), China; ^The PICDB data were harmonized to a common data model (OMOP) prior to being made available; *Cerner until October 2020, now Epic.

eTable 2. Organ dysfunction scores and criteria used in the study

Organ system	Organ Dysfunction Score/Criteria							
	IPSCC	PELOD-2	PODIUM	Proulx	pSOFA	DIC	VIS	SI
Cardiovascular	X	X	X	X	X		X	X
Respiratory	X	X	X	X	X			
Neurological	X	X	X	X	X			
Renal	X	X	X	X	X			
Hepatic	X		X	X	X			
Heme/Coag	X	X	X	X	X	X		
Immunologic			X					
Endocrine			X					

All the organ dysfunctions scores were developed and/or have primarily been validated to discriminate mortality. In our study we focused on the scores in the first 24 hours. *IPSCC*, International Pediatric Sepsis Consensus Conference; *PELOD-2*, Pediatric Logistic Organ Dysfunction, version 2; *PODIUM*, Pediatric Organ Dysfunction Information Update Mandate; *pSOFA*, pediatric Sequential Organ Failure Assessment; *SI*, shock index; *VIS*, vasoactive-inotrope score; *DIC*, disseminated intravascular coagulation.

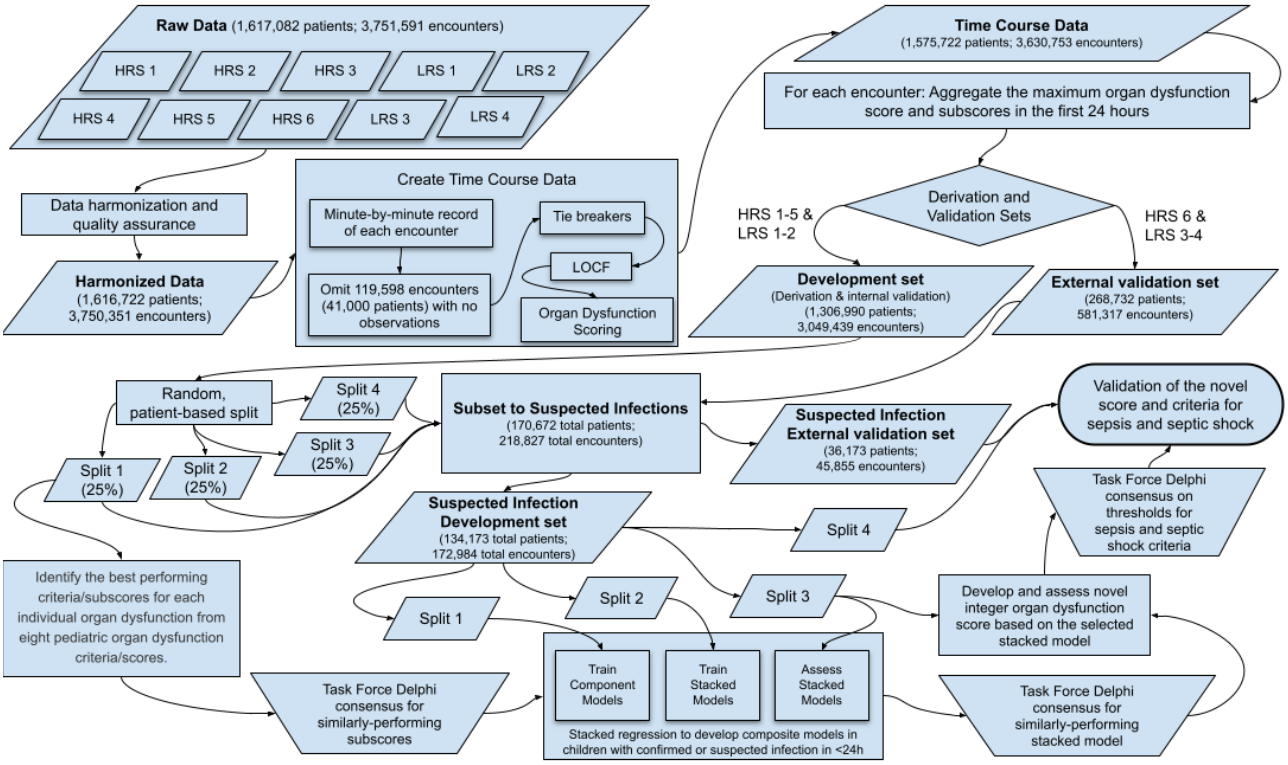
eFigure 1. Conceptual illustration of how stacked regression was used to develop the sepsis criteria



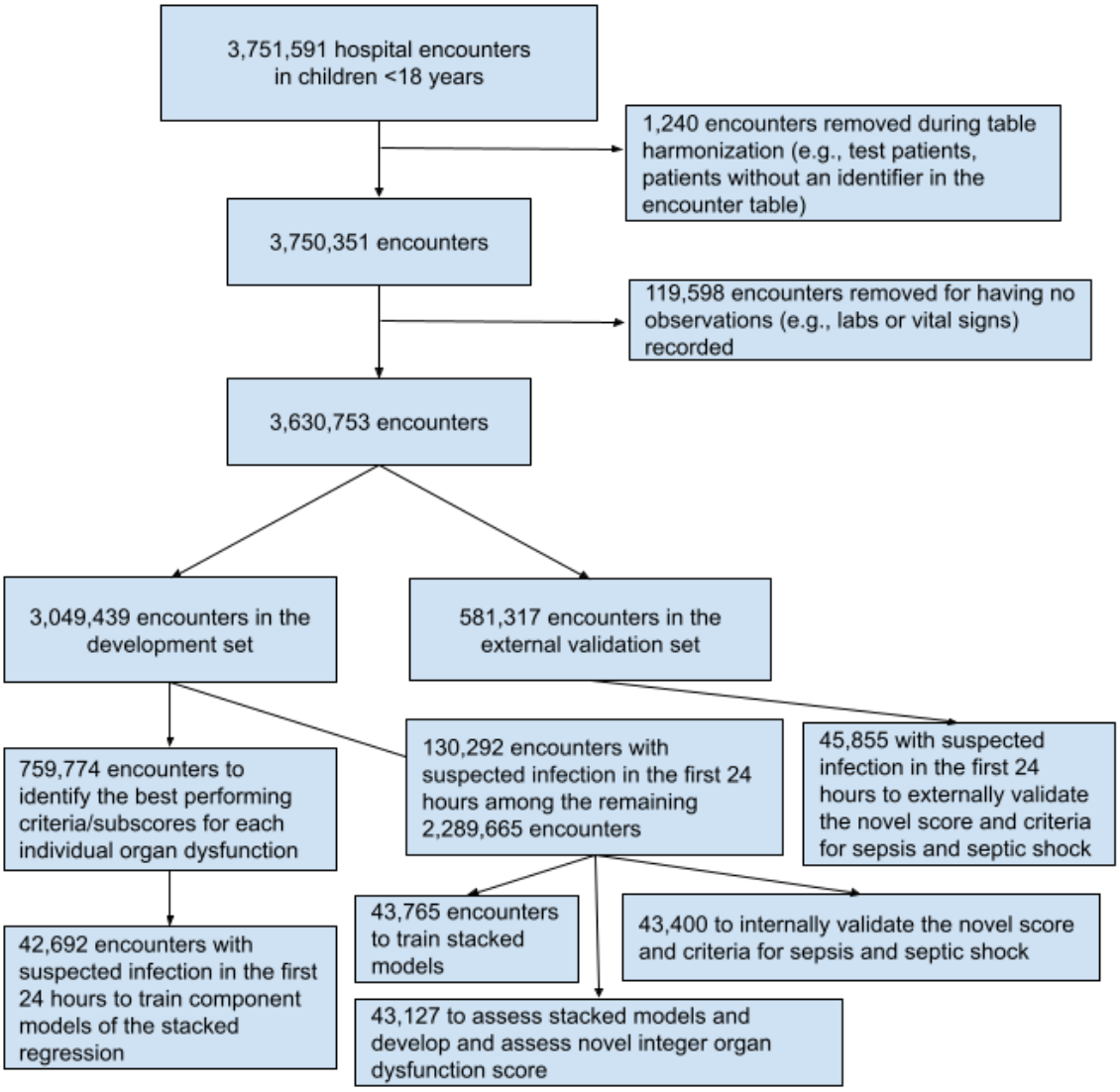
The top row shows illustrative existing organ dysfunction score subcomponents (in this case, the subcomponents that were ultimately integrated into the Phoenix Sepsis Score, but several others were also evaluated). The subcomponents were evaluated for their ability to predict mortality in the first descending step, with AUPRC (area under the precision recall curve) as the primary metric. Among those with suspected infection, stacked regression (a form of model averaging) models were then fit (second descending step) to predict mortality. As before, the primary metric was AUPRC. These models identified sepsis according to the agreed-upon conceptual definition of suspected infection with life-threatening organ dysfunction. An integer-based version of the best-performing sepsis model (the Phoenix Sepsis Score) was then identified, and new binary sepsis and septic shock criteria were chosen as thresholds on the Phoenix Sepsis Score (primary metrics were positive predictive value and sensitivity for this step). Please see Methods and eMethods for additional details . *VIS,* vasoactive-inotrope score; *CV*, cardiovascular; *BP*, blood pressure; *PELOD-2*, Pediatric Logistic Organ Dysfunction, version 2; *Resp*, respiratory; *P/F*, PaO2/FiO2 ratio; *S/F*, SpO2/FiO2 ratio; *pSOFA*, pediatric Sequential Organ Failure Assessment; *DIC*, disseminated intravascular coagulation; *Coag*, coagulation; GCS, Glasgow coma scale; *Neuro*, neurologic, *ODs*, organ dysfunctions.

eFigure 2. Pipeline for data harmonization, data quality, and data analysis (A), and CONSORT-style flow diagram for encounters in the pipeline and the various analyses (B)

A. Pipeline for data harmonization, data quality, and data analysis



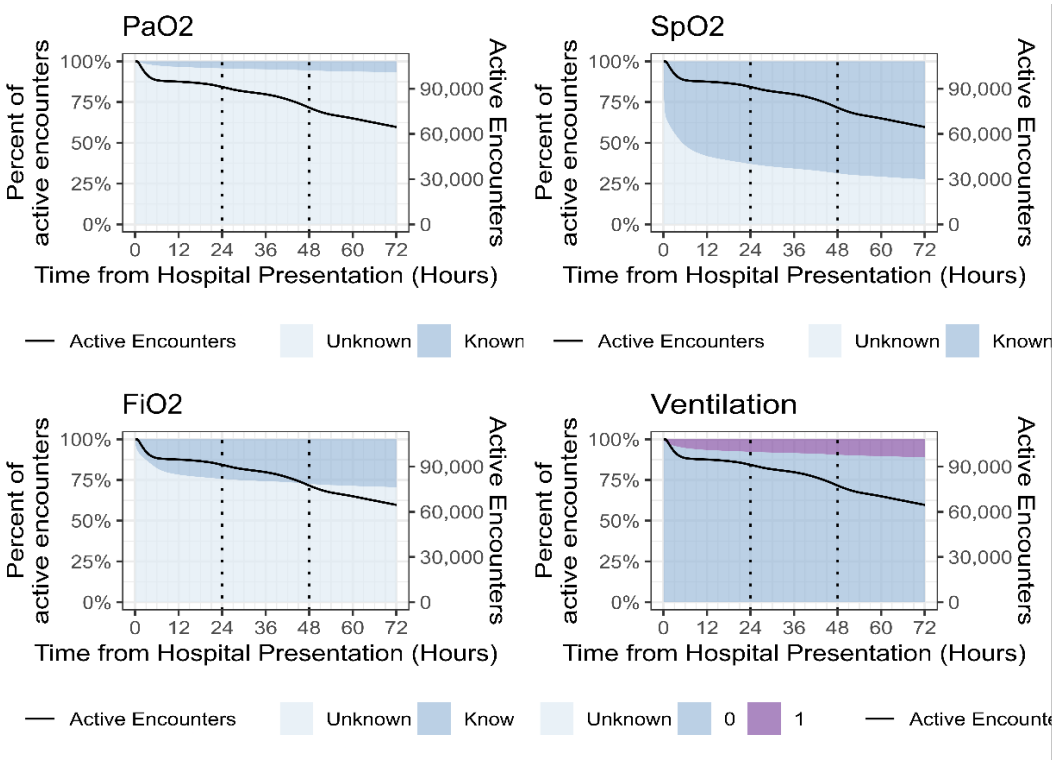
B. CONSORT-style flow diagram for encounters in the various analyses



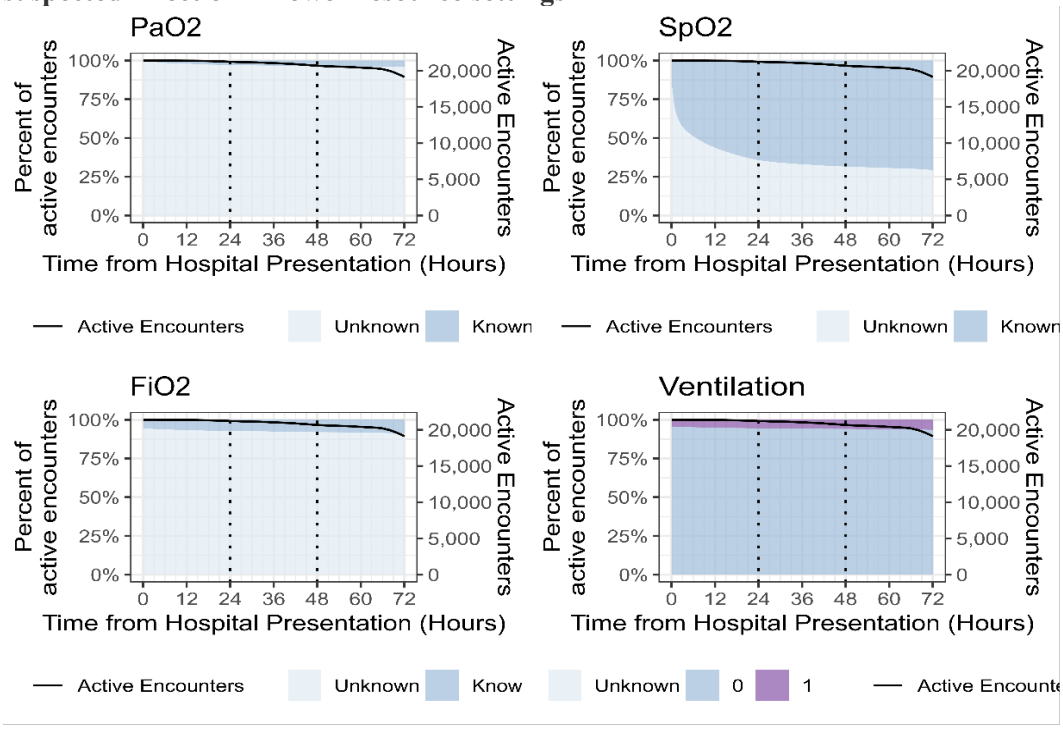
In panel A, Split 4 of the development set represents the 25% holdout internal validation set. Splits 1-3 were used for derivation, tuning, and testing of the best-performing measures and the stacked regression. *HRS*, higher resource setting sites; *LRS*, lower resource setting sites; *LOCF*, last observation carried forward.

eFigure 3A-H. Subscore input availability and missingness among patients with suspected infection in higher resource settings

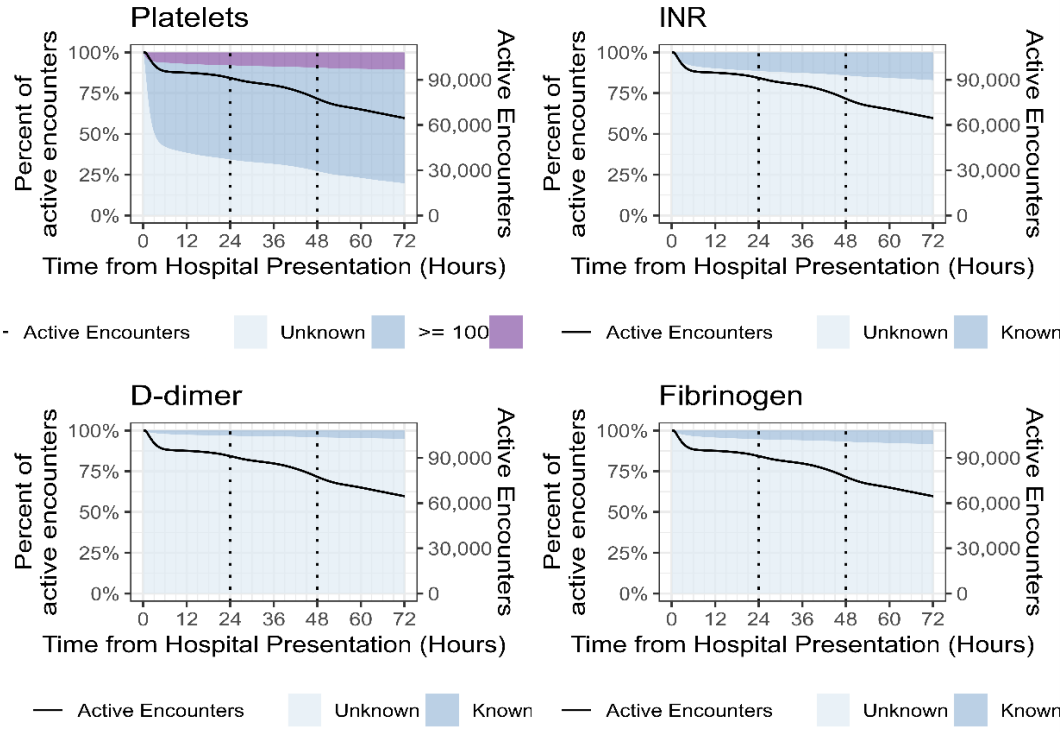
A. Respiratory dysfunction subscore input availability and missingness among patients with suspected infection in higher resource settings



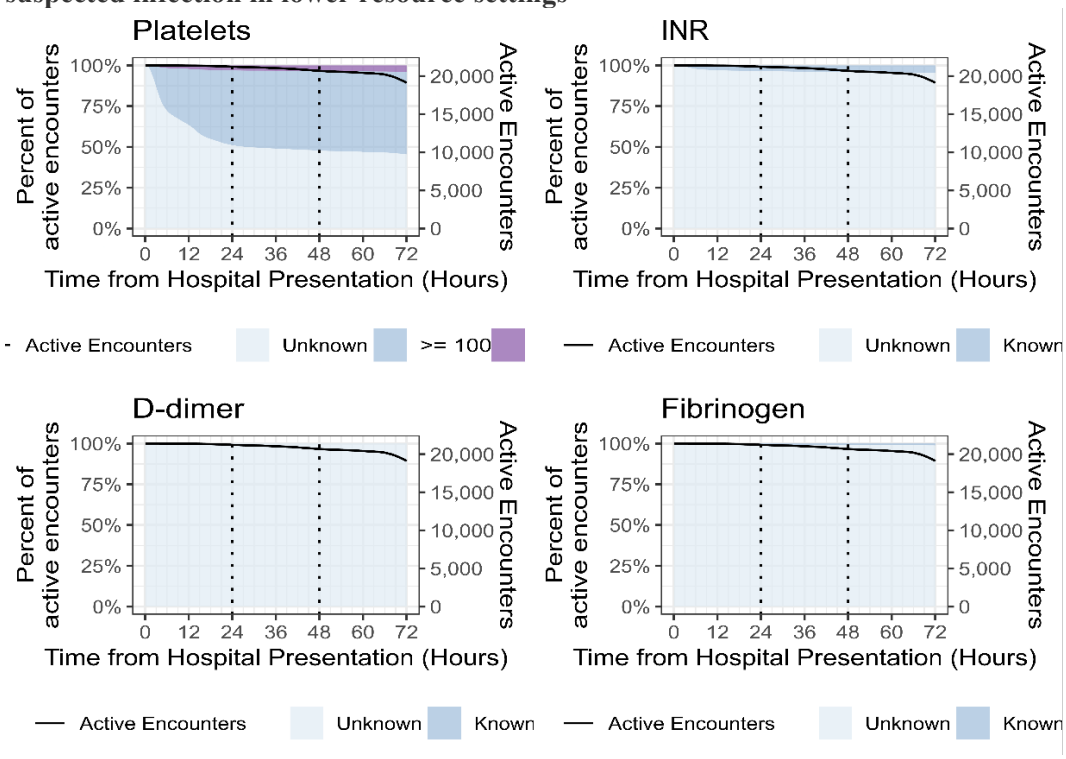
B. Respiratory dysfunction subscore input availability and missingness among patients with suspected infection in lower resource settings



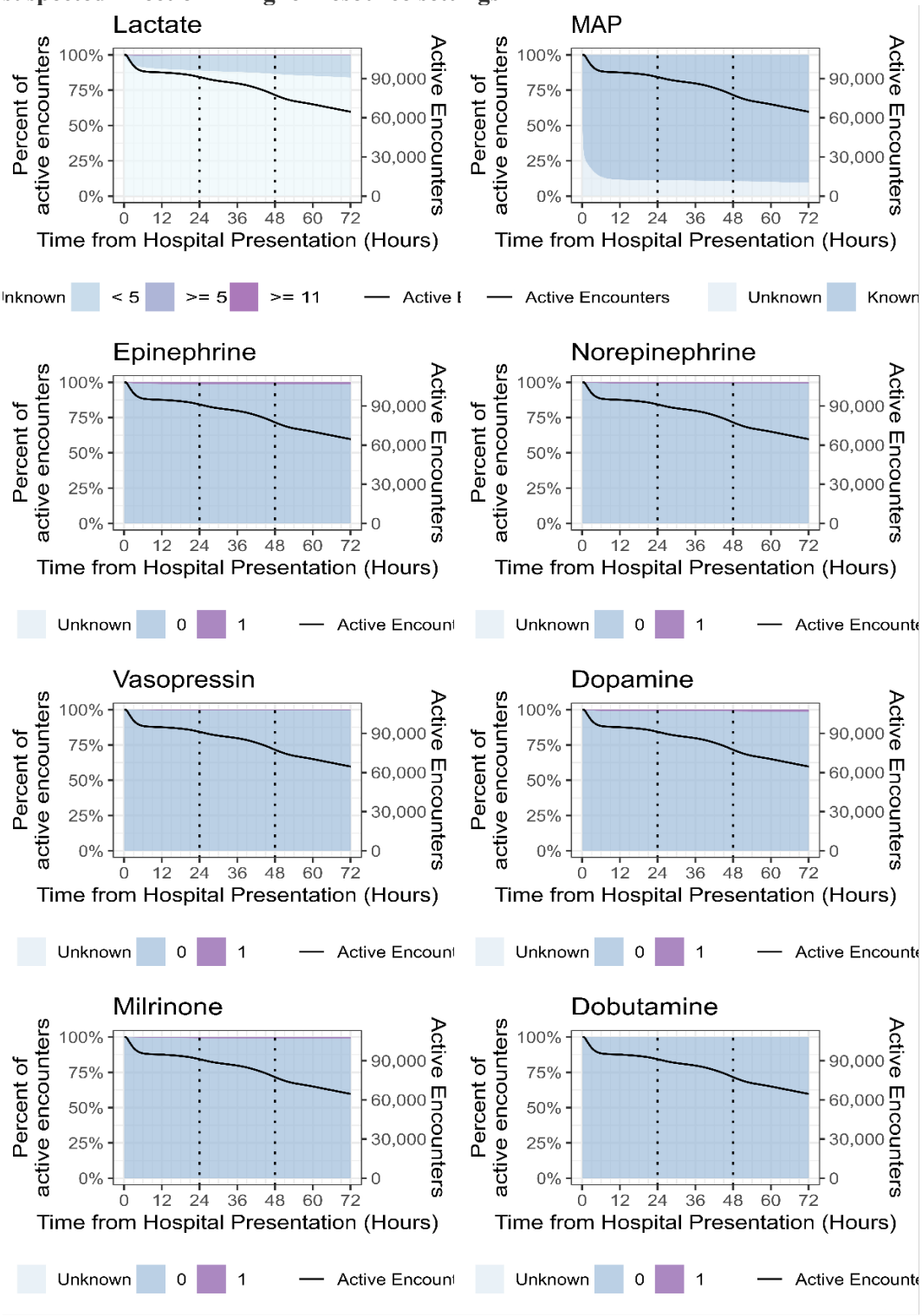
C. Coagulation dysfunction subscore input availability and missingness among patients with suspected infection in higher resource settings



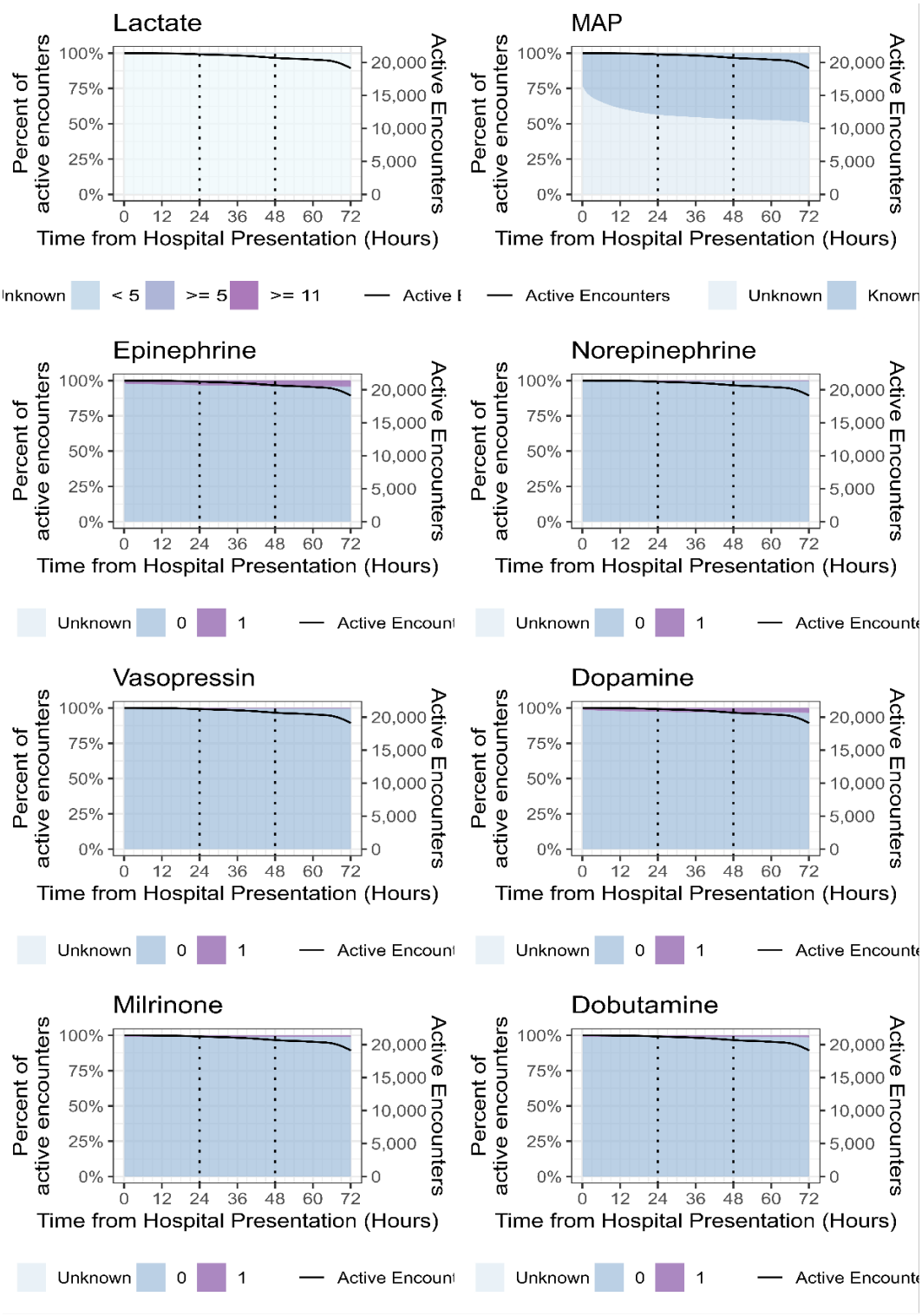
D. Coagulation dysfunction subscore input availability and missingness among patients with suspected infection in lower resource settings



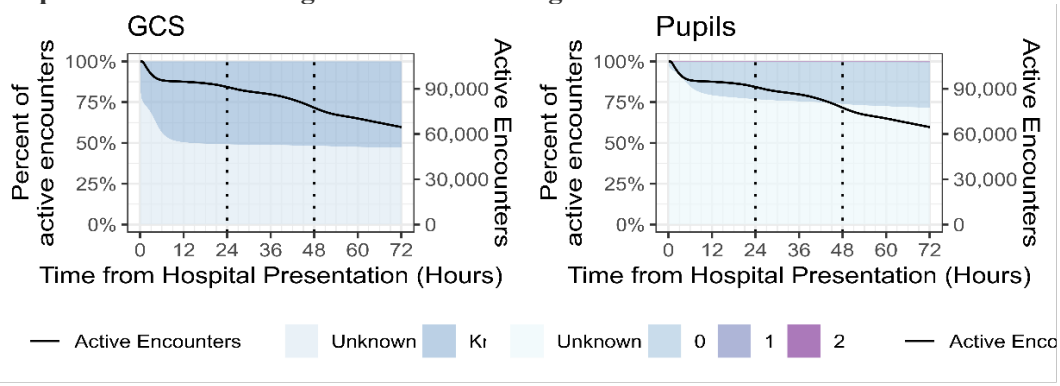
E. Cardiovascular dysfunction subscore input availability and missingness among patients with suspected infection in higher resource settings



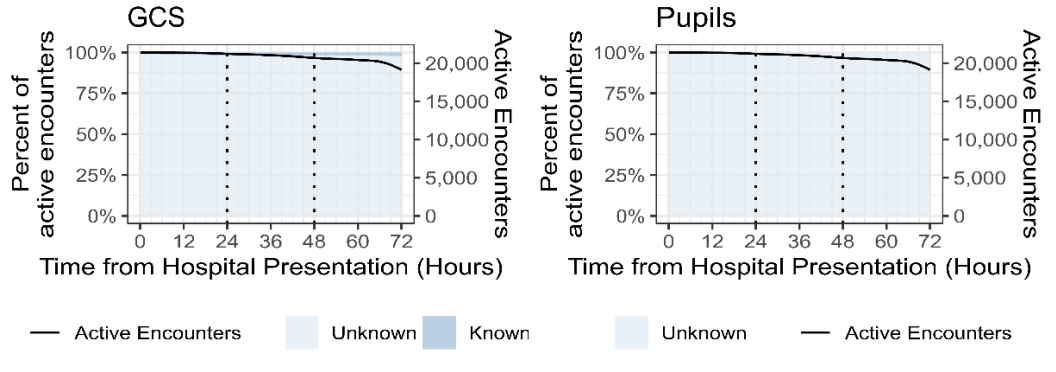
F. Cardiovascular dysfunction subscore input availability and missingness among patients with suspected infection in lower resource settings



G. Neurological dysfunction subscore input availability and missingness among patients with suspected infection in higher resource settings



H. Neurological dysfunction subscore input availability and missingness among patients with suspected infection in lower resource settings



In each of the panels in Figure 3, the black line represents the number of active encounters. For binary variables, 0 represents absence and 1 represents presence of the variable. For other categorical variables, the cutoffs used in the figure are shown under each figure. The dashed lines represent the 24 and 48 hour timepoints. *MAP*, mean arterial pressure, *GCS*, Glasgow coma scale.

eFigure 4. Performance of the individual subscores for each organ system based on the AUPRC and AUROC to predict mortality

Organ System	Criteria (first 24 hours)	Mean AUPRC	AUPRC 95% CI	Mean AUROC	AUROC 95% CI
Cardiovascular	IPSCC	0.017	(0.016, 0.018)	0.773	(0.769, 0.777)
Cardiovascular	**PELOD-2***	0.131	(0.128, 0.135)	0.746	(0.742, 0.750)
Cardiovascular	PODIUM	0.047	(0.045, 0.049)	0.720	(0.716, 0.725)
Cardiovascular	Proulx	0.044	(0.042, 0.046)	0.737	(0.733, 0.741)
Cardiovascular	pSOFA	0.063	(0.061, 0.065)	0.780	(0.776, 0.784)
Cardiovascular	Shock Index	0.012	(0.011, 0.013)	0.673	(0.668, 0.677)
Cardiovascular	Vasoactive inotrope score	0.108	(0.105, 0.111)	0.731	(0.727, 0.735)
Cardiovascular	**Vasoactive medication count***	0.135	(0.132, 0.138)	0.712	(0.708, 0.717)
Heme/Coag	IPSCC	0.022	(0.021, 0.023)	0.668	(0.664, 0.673)
Heme/Coag	PELOD-2	0.018	(0.017, 0.019)	0.658	(0.654, 0.663)
Heme/Coag	PODIUM Coagulation	0.007	(0.007, 0.008)	0.500	(0.495, 0.505)
Heme/Coag	PODIUM Hematologic	0.015	(0.014, 0.016)	0.639	(0.634, 0.643)
Heme/Coag	Proulx	0.017	(0.016, 0.018)	0.640	(0.636, 0.645)
Heme/Coag	pSOFA	0.022	(0.020, 0.023)	0.668	(0.664, 0.673)
Heme/Coag	**DIC Score**	0.131	(0.127, 0.134)	0.765	(0.761, 0.769)
Renal	IPSCC	0.023	(0.021, 0.024)	0.578	(0.573, 0.583)
Renal	PELOD-2	0.018	(0.017, 0.019)	0.646	(0.642, 0.651)
Renal	PODIUM	0.007	(0.006, 0.008)	0.598	(0.593, 0.602)
Renal	Proulx	0.034	(0.032, 0.036)	0.547	(0.543, 0.552)
Renal	**pSOFA***	0.028	(0.027, 0.030)	0.666	(0.662, 0.671)
Respiratory	IPSCC	0.020	(0.019, 0.022)	0.774	(0.770, 0.778)
Respiratory	PELOD-2	0.082	(0.079, 0.084)	0.763	(0.759, 0.767)
Respiratory	PODIUM	0.042	(0.040, 0.044)	0.720	(0.716, 0.725)
Respiratory	Proulx	0.026	(0.025, 0.028)	0.767	(0.763, 0.771)
Respiratory	**pSOFA***	0.050	(0.048, 0.052)	0.777	(0.773, 0.781)

Hepatic	**IPSCC***	0.034	(0.032, 0.036)	0.643	(0.638, 0.647)
Hepatic	PODIUM	0.029	(0.027, 0.031)	0.703	(0.699, 0.708)
Hepatic	Proulx	0.014	(0.013, 0.015)	0.523	(0.519, 0.528)
Hepatic	pSOFA	0.017	(0.015, 0.018)	0.559	(0.554, 0.564)
Neurological	IPSCC	0.035	(0.033, 0.036)	0.712	(0.708, 0.716)
Neurological	**PELOD-2**	0.119	(0.116, 0.122)	0.721	(0.717, 0.726)
Neurological	PODIUM	0.042	(0.040, 0.044)	0.710	(0.706, 0.715)
Neurological	Proulx	0.069	(0.067, 0.071)	0.688	(0.683, 0.692)
Neurological	pSOFA	0.047	(0.045, 0.049)	0.720	(0.715, 0.724)
Immunologic	**PODIUM**	0.011	(0.010, 0.012)	0.627	(0.622, 0.631)
Endocrine	**PODIUM**	0.027	(0.025, 0.028)	0.716	(0.711, 0.720)

Criteria selected are **bolded**. Criteria selected by the Task Force through modified Delphi consensus based on generalizability of the subscore (e.g., input availability in various settings, see Methods), are denoted by an asterisk (*****). Point estimates are mean values based on bootstrap analysis. Confidence intervals (CIs) were determined using Logit transform. *AUPRC*, area under the precision recall curve; *AUROC*, area under the receiver operator characteristic curve.

eTable 3. Cohort characteristics of the development set stratified by infection status

	All encounters	Suspected infection in first 24h
Encounters, No.	3,049,699	172,984
Resource Setting, No. (%)		
Higher Resource Sites	2,953,967 (96.9)	144,379 (83.5)
Lower Resource Sites	95,732 (3.1)	28,605 (16.5)
Age in years, median (IQR)	4.8 (1.6, 10.5)	3.7 (0.9, 9.4)
Sex, No. (%)		
Female	1,427,844 (46.8)	83,909 (48.5)
Male	1,621,784 (53.2)	89,069 (51.5)
Race		
American Indian or Alaskan Native	1,193 (0.0)	130 (0.1)
Asian	117,181 (3.8)	6,852 (4.0)
Black	725,752 (23.8)	30,221 (17.5)
Multiple Races	565,404 (18.5)	29,456 (17.0)
Native Hawaiian or Other Pacific Islander	1,192 (0.0)	136 (0.1)
Other/Unknown	183,835 (6.0)	29,404 (17.0)
White	1,457,527 (47.8)	77,051 (44.5)
Ethnicity, No. (%)		
Hispanic or Latino	899,305 (29.5)	45,155 (26.1)
Major comorbidity, No. (%)		
Malignancy	62,706 (2.1)	14,633 (8.5)
Technology Dependence	138,336 (4.5)	24,962 (14.4)
Transplantation	17,981 (0.6)	4,976 (2.9)
Severe malnutrition	101,432 (5)	17,983 (16)
Comorbidities per PCCC, No. (%)		
No known prior comorbidity	2,680,871 (87.9)	125,366 (72.5)
1 PCCC	165,554 (5.4)	12,556 (7.3)
2 or more PCCCs	203,274 (6.7)	35,062 (20.3)
SIRS, No. (%)	417,417 (13.7)	75,559 (43.7)
Locations visited during encounter, No. (%)		

Presented through the ED	2,740,205 (90.0)	123,599 (71.7)
Had 1 or more OR visit(s)	168,662 (5.5)	23,702 (13.7)
Had 1 or more ICU stays	112,614 (3.7)	30,968 (18.0)
Outcomes, No. (%)		
Death	4,688 (0.2)	2,065 (1.2)
Early Death or ECMO	2,778 (0.1)	1,139 (0.7)

eTable 3 shows site, demographic, care location, comorbidity, and outcome characteristics of the cohort at the 7 development sites, stratified by infection status at 24 hours. For race categories, “Multiple Races” indicates that in the EHR data, the patient’s race was recorded as “multi-racial,” “multiple,”, or “two or more races.” “Unknown/Other” indicates that the patient’s race was recorded in the EHR data as “other,” “unknown,” “not specified,” “information not recorded,” “patient declined,” “patient refused,” “refused,” or as a race category unique to a particular international country or region. For ethnicity categories PCCC is a system to classify pediatric chronic diseases using International Classification of Diseases (ICD) diagnosis and procedure codes and was only assessed in the higher resource sites, where the information was available (percentages for PCCC-related counts are based on higher resource setting encounters).7 The major comorbidities of technology dependence (e.g. requiring a gastrostomy, a tracheostomy, a central line, etc.), malignancy, and transplantation were defined in the PCCC system. Severe malnutrition was defined as based on <3 standard deviations below the mean based on weight-for-age standards from the World Health Organization and assessed in all sites.8 Early Death is defined as death in <72 hours from the beginning of the encounter. Systemic inflammatory response syndrome (SIRS) is calculated using temperature, white blood cell count, heart rate, and respiratory rate, with higher values reflecting more inflammation. SIRS criteria are met when two or more values are above the threshold for age, including at least temperature or white blood cell count. See Supplemental Methods for additional details. IQR, interquartile range; PCCC, pediatric complex chronic conditions; SIRS, systemic inflammatory response syndrome; ED, emergency department; OR, operating room; ICU, intensive care unit (locations not mutually exclusive). ECMO, extracorporeal membrane oxygenation.

eTable 4. Cohort characteristics of the development set stratified by infection status and site

	Higher Resource sites 1-5		Lower Resource site 1		Lower Resource site 2	
	All encounters	Confirmed or suspected infection in first 24h	All encounters	Confirmed or suspected infection in first 24h	All encounters	Confirmed or suspected infection in first 24h
Encounters	2,953,965	144,379	46,075	8,595	49,657	20,010
Age in years (IQR)	4.9 (1.7, 10.5)	4.9 (1.5, 10.4)	1.8 (0, 13.9)	0.2 (0, 2.4)	0.8 (0.42, 1.33)	0.8 (0.4, 1.5)
Severe malnutrition	74,250 (4)	4,266 (5)	4,154 (27)	1,498 (34)	23,028 (54)	12,219 (65)
No known prior comorbidity*	2,585,139 (87.5)	96,761 (67.0)	-	-	-	-
SIRS	384,672 (13.0)	62,112 (43.0)	18,531 (40.2)	3,784 (44.0)	14,214 (28.6)	9,663 (48.3)
Outcomes						
Death	3,036 (0.1)	1,049 (0.7)	764 (1.7)	356 (4.1)	888 (1.8)	660 (3.3)
Early Death or ECMO	2,004 (0.1)	621 (0.4)	240 (0.5)	105 (1.2)	534 (1.1)	413 (2.1)

*Lower resource sites 1 and 2 had limited information to ascertain pediatric complex chronic conditions. *IQR*, interquartile range; *SIRS*, systemic inflammatory response syndrome; *ECMO*, extracorporeal membrane oxygenation.

eTable 5. Cohort characteristics of the external validation set stratified by infection status and site

	Higher Resource site 6		Lower Resource site 3		Lower Resource site 4	
	All encounters	**Confirmed or suspected infection in first 24h**	**All encounters**	**Confirmed or suspected infection in first 24h**	**All encounters**	**Confirmed or suspected infection in first 24h**
Encounters	535,940	33,020	34,275	10,830	11,102	2,005
Age in years, (IQR)	5 (1.6, 10.7)	3.8 (1, 9.5)	1.1 (0, 3.8)	1.1 (0, 3.3)	0.6 (0, 3)	0.6 (0, 3)
Severe malnutrition*	7,447 (2)	1,405 (6)	8,209 (27)	2,012 (21)	-	-
No known prior comorbidity**	477,883 (89.2)	22,603 (68.5)	-	-	-	-
SIRS	94,825 (17.7)	15,865 (48.0)	15,345 (44.8)	4,945 (45.7)	2,202 (19.8)	626 (31.2)
Outcomes						
Death	448 (0.1)	212 (0.6)	2,977 (8.7)	241 (2.2)	537 (4.8)	87 (4.3)
Early Death or ECMO	366 (0.1)	188 (0.6)	1,882 (5.5)	124 (1.1)	188 (1.7)	37 (1.8)

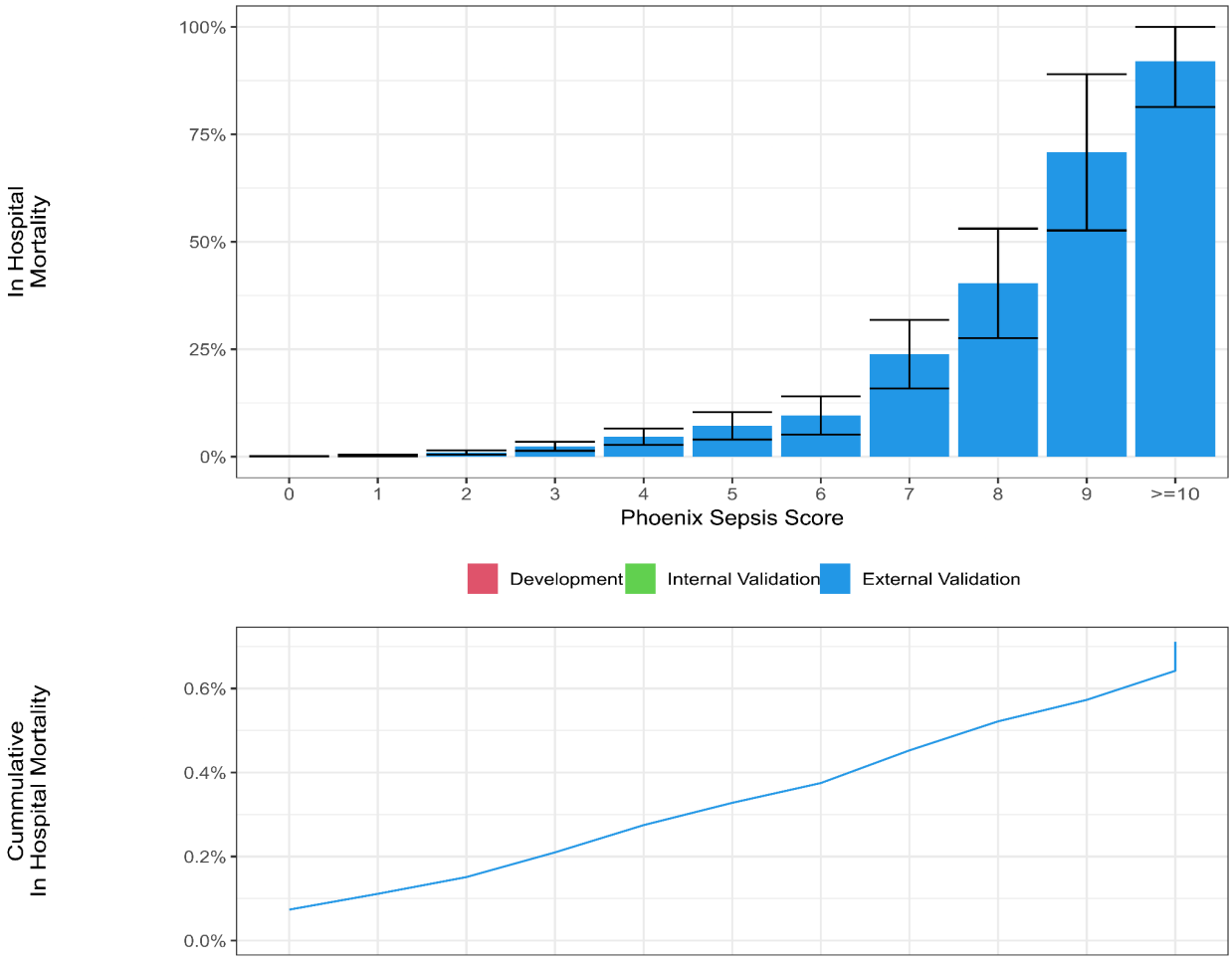
*Lower resource site 4 had no weight information. **Lower resource sites 3 and 4 had limited information to ascertain pediatric complex chronic conditions. *IQR*, interquartile range; *SIRS*, systemic inflammatory response syndrome; *ECMO*, extracorporeal membrane oxygenation.

eTable 6. Stacked regression coefficients of the 8-organ system ridge regression model and the 4-organ system LASSO model

Stacked penalized regression models	Component model coefficients
Ridge	
Intercept	-5.12
Vasoactive Count	5.38
PELOD-2 Cardiovascular	2.54
PELOD-2 Neurological	5.98
pSOFA Respiratory	5.05
DIC Score	3.77
IPSCC Hepatic	5.04
pSOFA Renal	6.72
PODIUM Endocrine	5.15
PODIUM Immunologic	4.03
LASSO	
Intercept	-5.08
Vasoactive Count	7.48
PELOD-2 Cardiovascular	0.4
PELOD-2 Neurological	10.79
pSOFA Respiratory	4.83
DIC Score	5.61

Note that because these are stacked regression coefficients of the component models, there is not necessarily a linear translation between the weight of the coefficients and the relative importance of the component models.

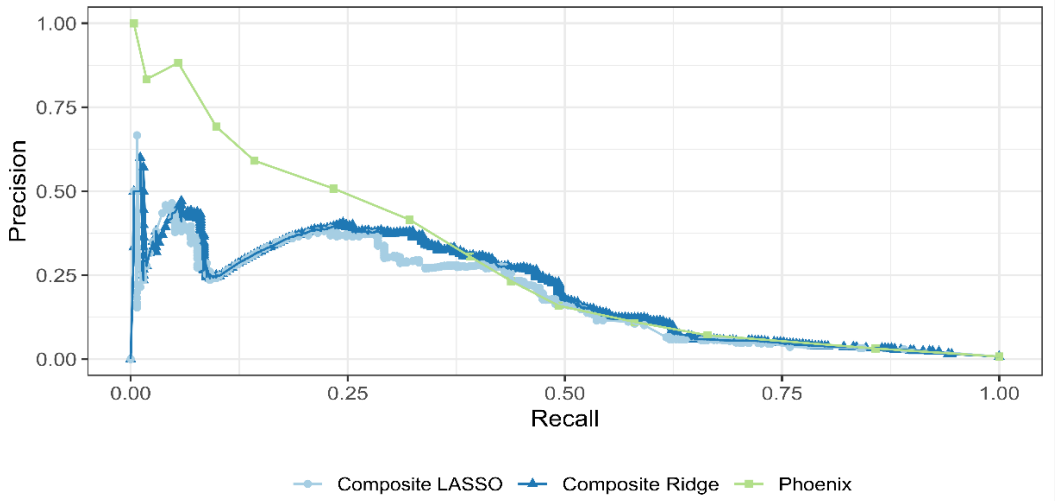
eFigure 5. In-hospital mortality associated with the Phoenix Sepsis Score in patients with suspected infection in the first 24 hours at higher resource site 6 (the geographic external validation set)



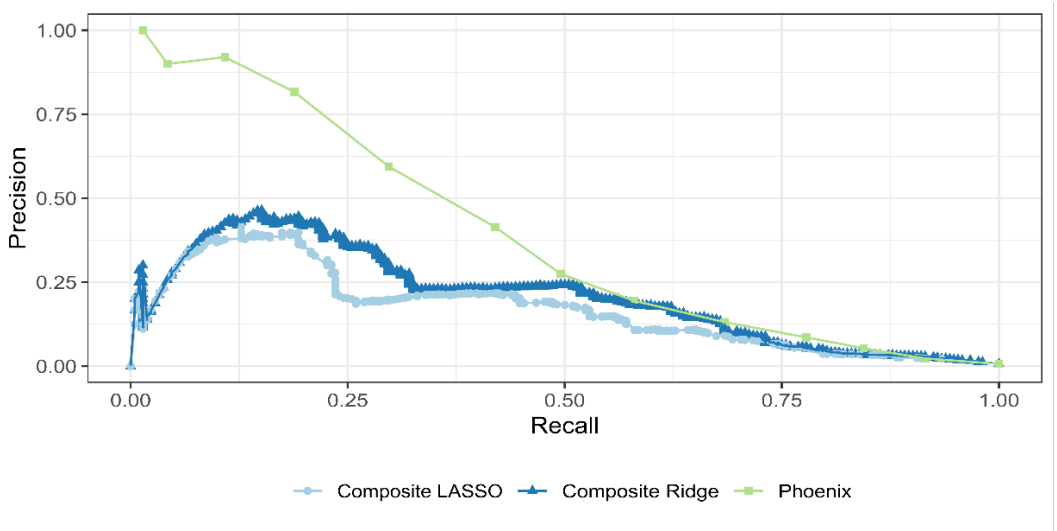
eFigure 5 shows the calibration of the Phoenix Sepsis Score in higher resource site 6 (the higher resource external validation set). For patients with suspected infection who have each possible integer value (lower x-axis) of the Phoenix Sepsis Score in the first 24 hours of the encounter, the y-axis shows mortality (blue bar graph). Binomial confidence intervals for the mortality point estimate in each group are also shown. The middle of each panel shows cumulative mortality across Phoenix Sepsis Score categories. The number of encounters "at risk" and mortality counts in each group are shown across the bottom of that plot.

eFigure 6A-J. AUPRC and AUROC curves for the four-organ system model

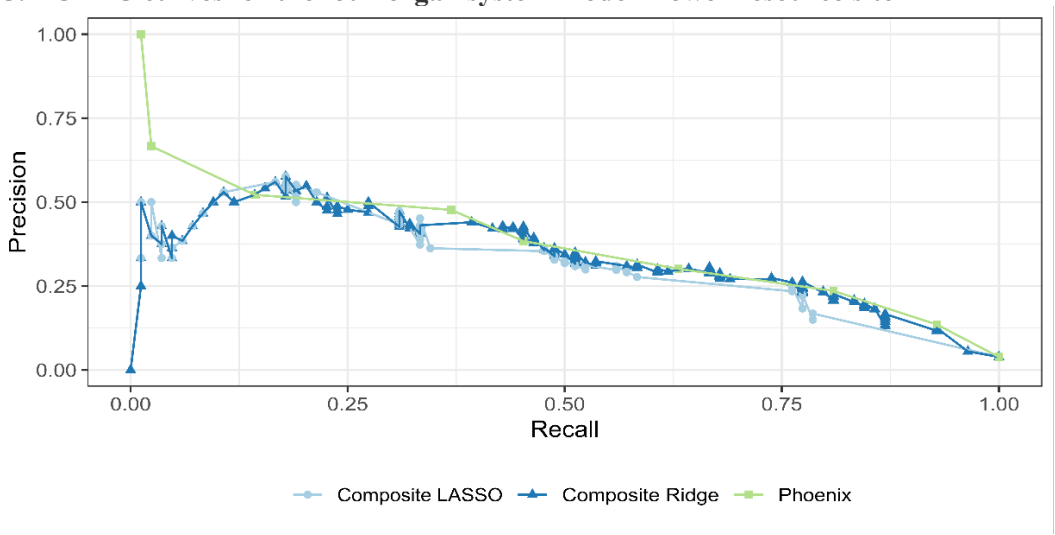
A. AUPRC curves for the four-organ system model - Higher resource sites 1-5 in the derivation and internal validation set



B. AUPRC curves for the four-organ system model - Higher Resource site 6 in the external validation set

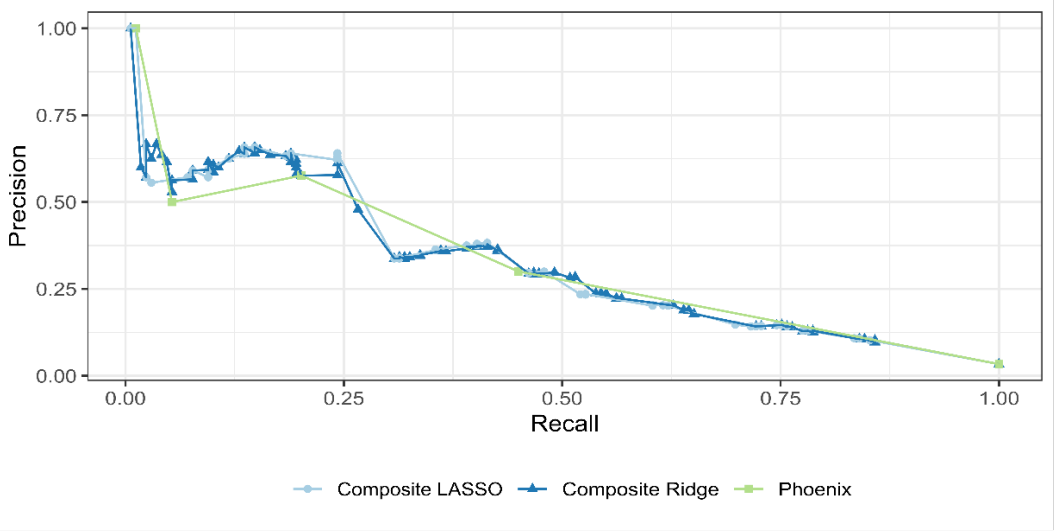


C. AUPRC curves for the four-organ system model - lower resource site 1



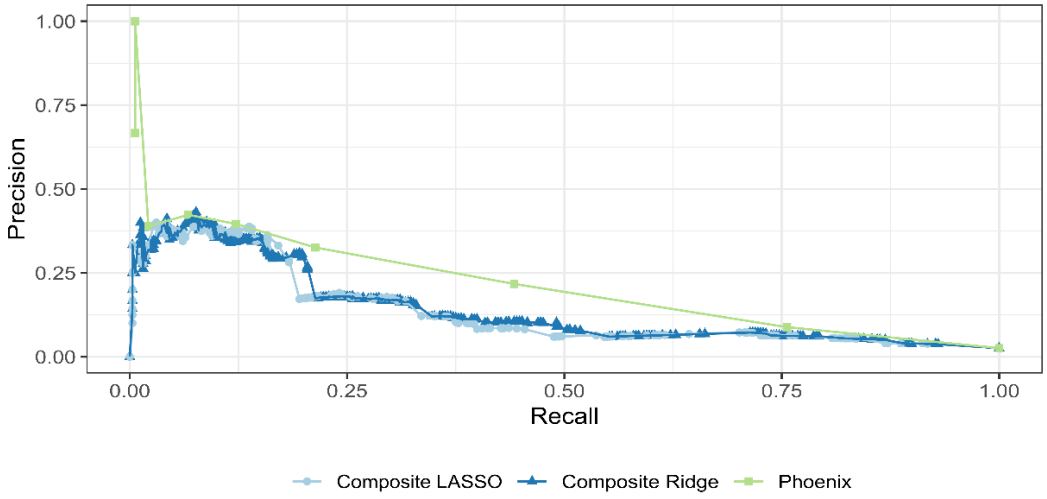
Shown separate from lower resource site 2 due to data availability, see text.

D. AUPRC curves for the four-organ system model - lower resource site 2

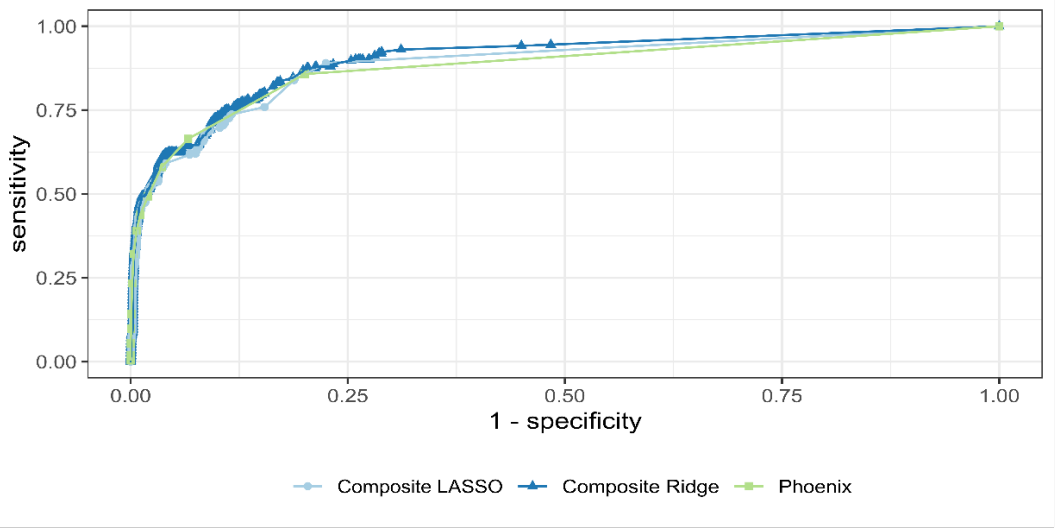


Shown separate from lower resource site 1 due to data availability, see text.

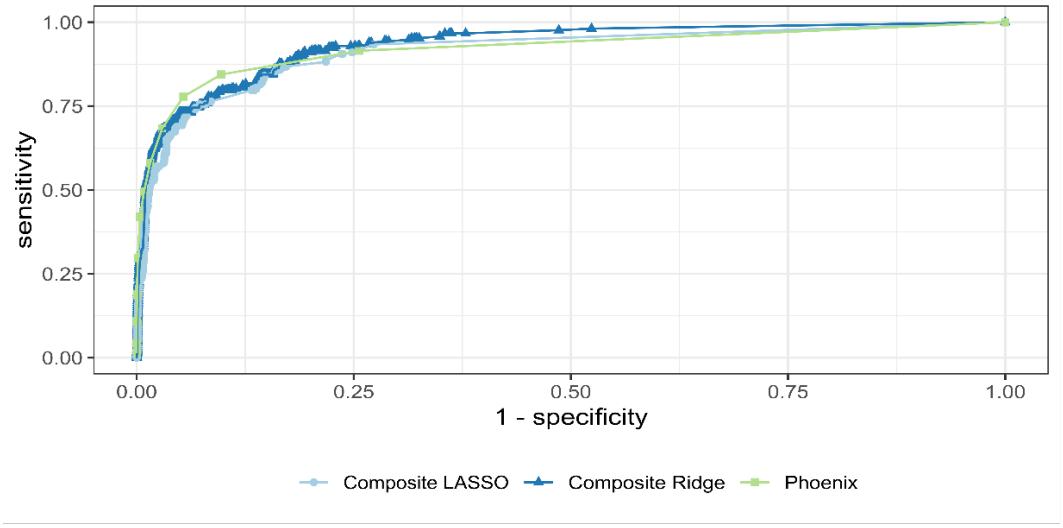
E. AUPRC curves for the four-organ system model - lower resource sites 3-4 (external validation set)



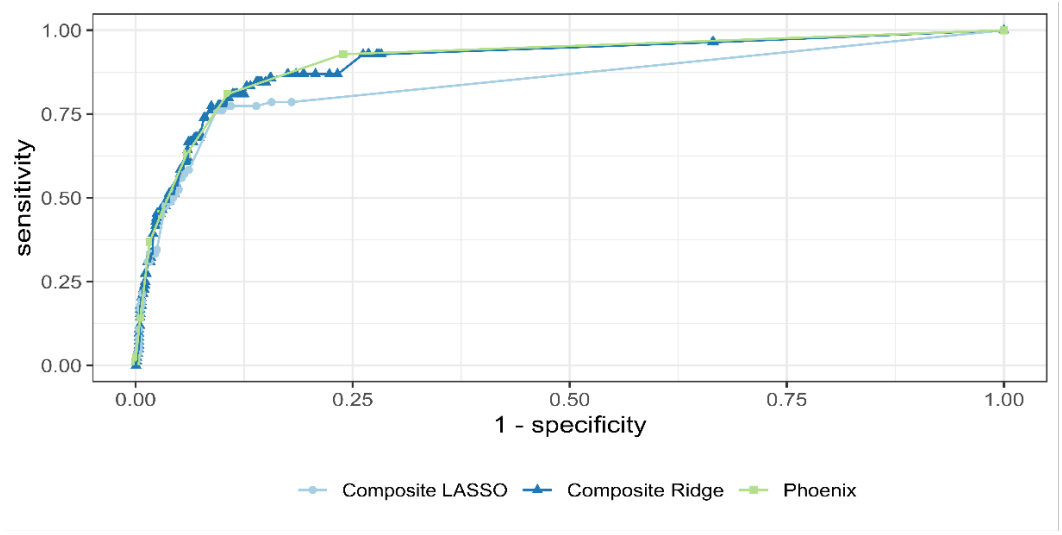
F. AUROC curves for the four-organ system model - higher resource sites 1-5 (derivation and internal validation)



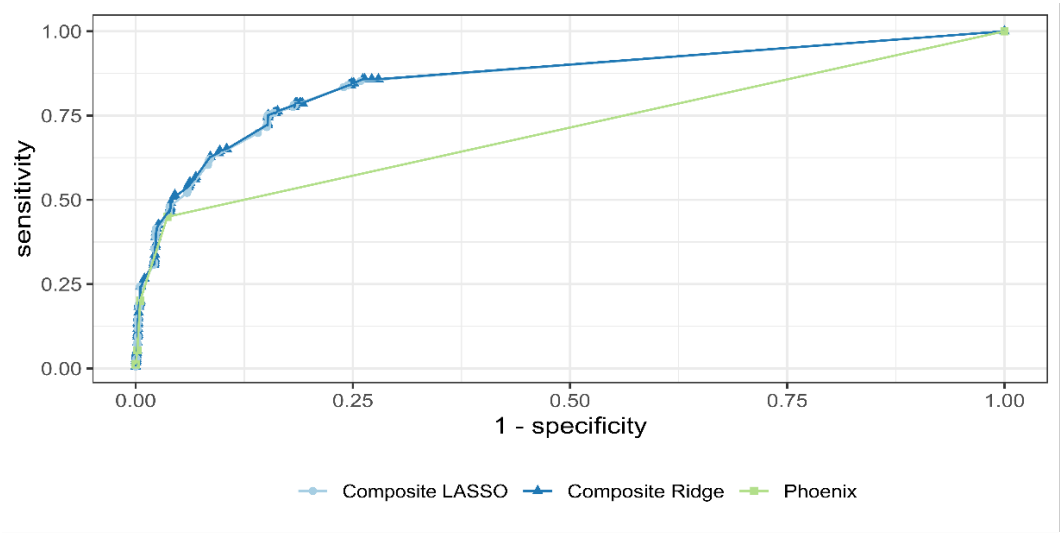
G. AUROC curves for the four-organ system model - higher resource site 6 (external validation)



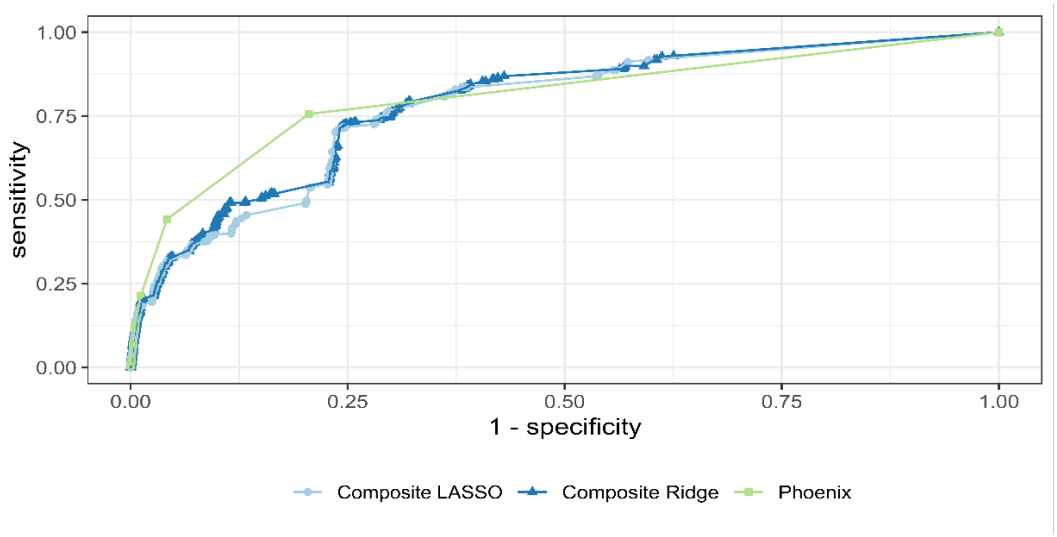
H. AUROC curves for the four-organ system model - lower resource site 1 (derivation and internal validation)



I. AUROC curves for the four-organ system model - lower resource site 2 (derivation and internal validation)



J. AUROC curves for the four-organ system model - lower resource sites 3-4 (external validation)



In each panel in Figure 6, lower resource sites 1 and 2 are shown separately due to data availability (see text). *AUPRC*, area under the precision-recall curve; *AUROC*, area under the receiver operating characteristic curve; *composite LASSO*, 4-organ system stacked model; *composite Ridge*, 8-organ system stacked model; *Phoenix*, Phoenix Sepsis Score.

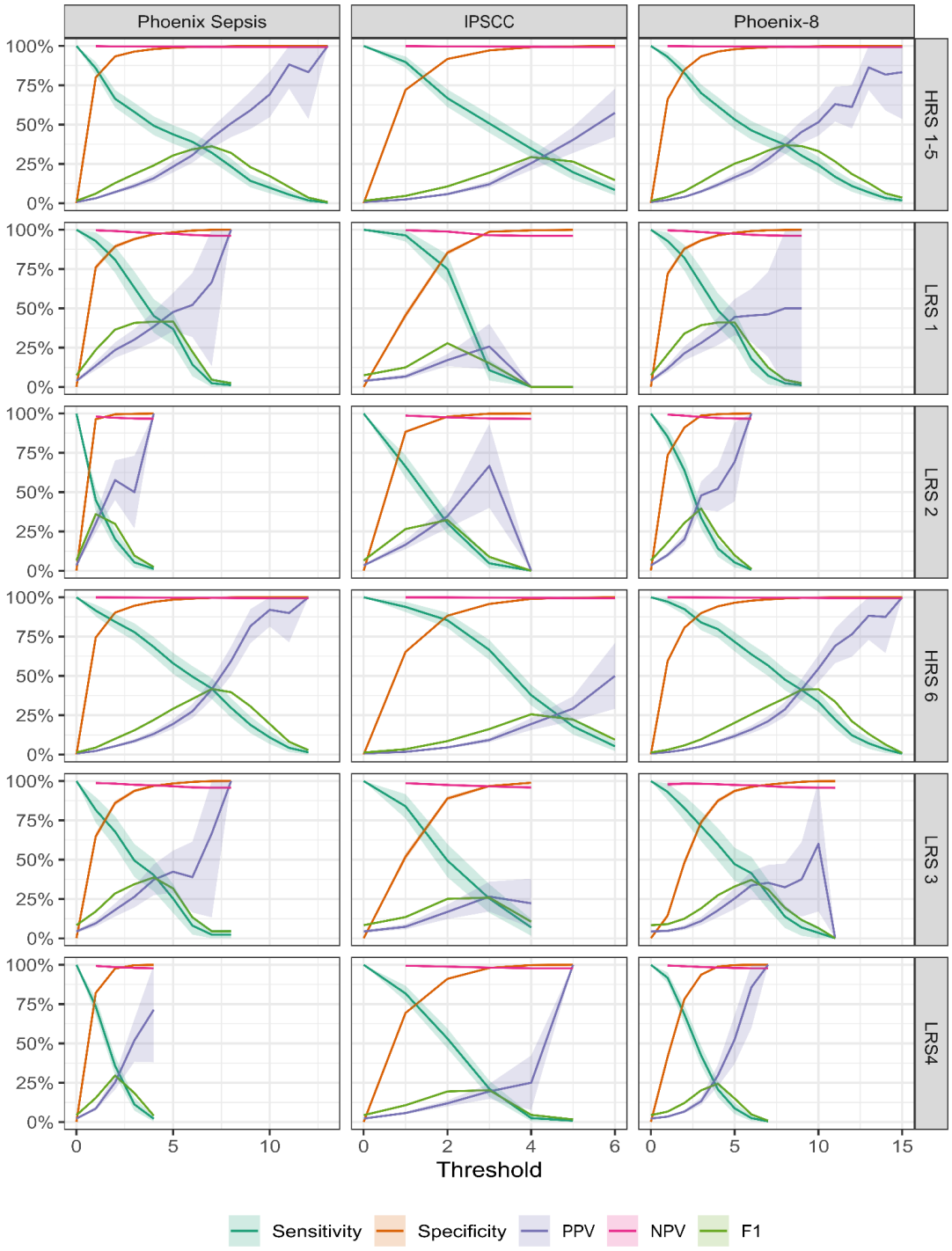
eFigure 7. Performance of the Phoenix Sepsis Score and organ dysfunction scores to predict early death or extracorporeal membrane oxygenation

Area Under the Precision Recall Curve (AUPRC)							
	Phoenix sepsis	**IPSCC**	**Phoenix-8**	**PELOD-2**	**pSOFA**	**Proulx**	**PODIUM**
Internal validation set							
Higher Resource Sites 1-5	**0.48**	0.20	0.42	0.44	0.35	0.39	0.29
	(0.47, 0.48)	(0.20, 0.20)	(0.42, 0.43)	(0.44, 0.45)	(0.35, 0.36)	(0.38, 0.39)	(0.29, 0.30)
Lower Resource Site 1	**0.14**	0.07	0.13	0.07	0.10	0.18	0.07
	(0.13, 0.16)	(0.06, 0.08)	(0.12, 0.15)	(0.06, 0.08)	(0.09, 0.11)	(0.16, 0.20)	(0.06, 0.08)
Lower Resource Site 2	0.30	0.27	**0.31**	0.13	0.35	0.19	0.20
	(0.29, 0.31)	(0.26, 0.28)	(0.29, 0.32)	(0.12, 0.14)	(0.33, 0.36)	(0.18, 0.20)	(0.19, 0.21)
External validation set							
Higher Resource Site 6	**0.31**	0.13	0.30	0.28	0.20	0.19	0.21
	(0.30, 0.31)	(0.12, 0.13)	(0.30, 0.31)	(0.28, 0.29)	(0.19, 0.20)	(0.18, 0.19)	(0.21, 0.22)
Lower Resource Site 3	**0.23**	0.10	0.21	0.16	0.09	0.09	0.11
	(0.21, 0.25)	(0.09, 0.12)	(0.19, 0.23)	(0.14, 0.17)	(0.08, 0.10)	(0.07, 0.10)	(0.10, 0.13)
Lower Resource Site 4	**0.18**	0.09	0.15	0.08	0.13	0.09	0.07
	(0.18, 0.19)	(0.09, 0.10)	(0.15, 0.16)	(0.08, 0.09)	(0.12, 0.14)	(0.09, 0.10)	(0.07, 0.08)
All Sites, Internal and External Validation Sets	**0.26**	0.12	0.24	0.22	0.20	0.20	0.17
	(0.26, 0.26)	(0.12, 0.13)	(0.24, 0.25)	(0.22, 0.22)	(0.20, 0.20)	(0.20, 0.21)	(0.17, 0.17)
Area Under the Receiver Operator Characteristic Curve (AUROC)							
Internal validation set							
Higher Resource Sites 1-5	0.96	0.95	**0.97**	0.95	0.97	0.95	0.95
	(0.96, 0.96)	(0.95, 0.96)	(0.97, 0.97)	(0.95, 0.95)	(0.97, 0.97)	(0.95, 0.95)	(0.95, 0.95)
Lower Resource Site 1	**0.87**	0.83	0.86	0.82	0.84	0.89	0.73
	(0.85, 0.88)	(0.81, 0.84)	(0.85, 0.88)	(0.81, 0.84)	(0.83, 0.86)	(0.88, 0.90)	(0.71, 0.75)
Lower Resource Site 2	0.79	0.85	**0.90**	0.82	0.89	0.79	0.81
	(0.77, 0.80)	(0.84, 0.86)	(0.89, 0.91)	(0.81, 0.83)	(0.88, 0.90)	(0.78, 0.80)	(0.80, 0.82)
External validation set							
Higher Resource Site 6	0.96	0.93	**0.96**	0.94	0.96	0.93	0.94
	(0.96, 0.96)	(0.93, 0.93)	(0.96, 0.97)	(0.94, 0.94)	(0.96, 0.96)	(0.93, 0.94)	(0.94, 0.95)
Lower Resource Site 3	**0.94**	0.85	0.92	0.82	0.83	0.79	0.86
	(0.93, 0.95)	(0.84, 0.87)	(0.91, 0.93)	(0.80, 0.83)	(0.82, 0.85)	(0.77, 0.81)	(0.85, 0.88)
Lower Resource Site 4	0.87	0.86	0.85	0.79	**0.88**	0.79	0.79
	(0.87, 0.88)	(0.86, 0.87)	(0.84, 0.86)	(0.78, 0.80)	(0.87, 0.88)	(0.78, 0.79)	(0.79, 0.80)
All Sites, Internal and External Validation Sets	0.89	0.89	**0.92**	0.87	**0.92**	0.87	0.89
	(0.89, 0.90)	(0.89, 0.89)	(0.92, 0.92)	(0.86, 0.87)	(0.92, 0.92)	(0.87, 0.87)	(0.89, 0.89)

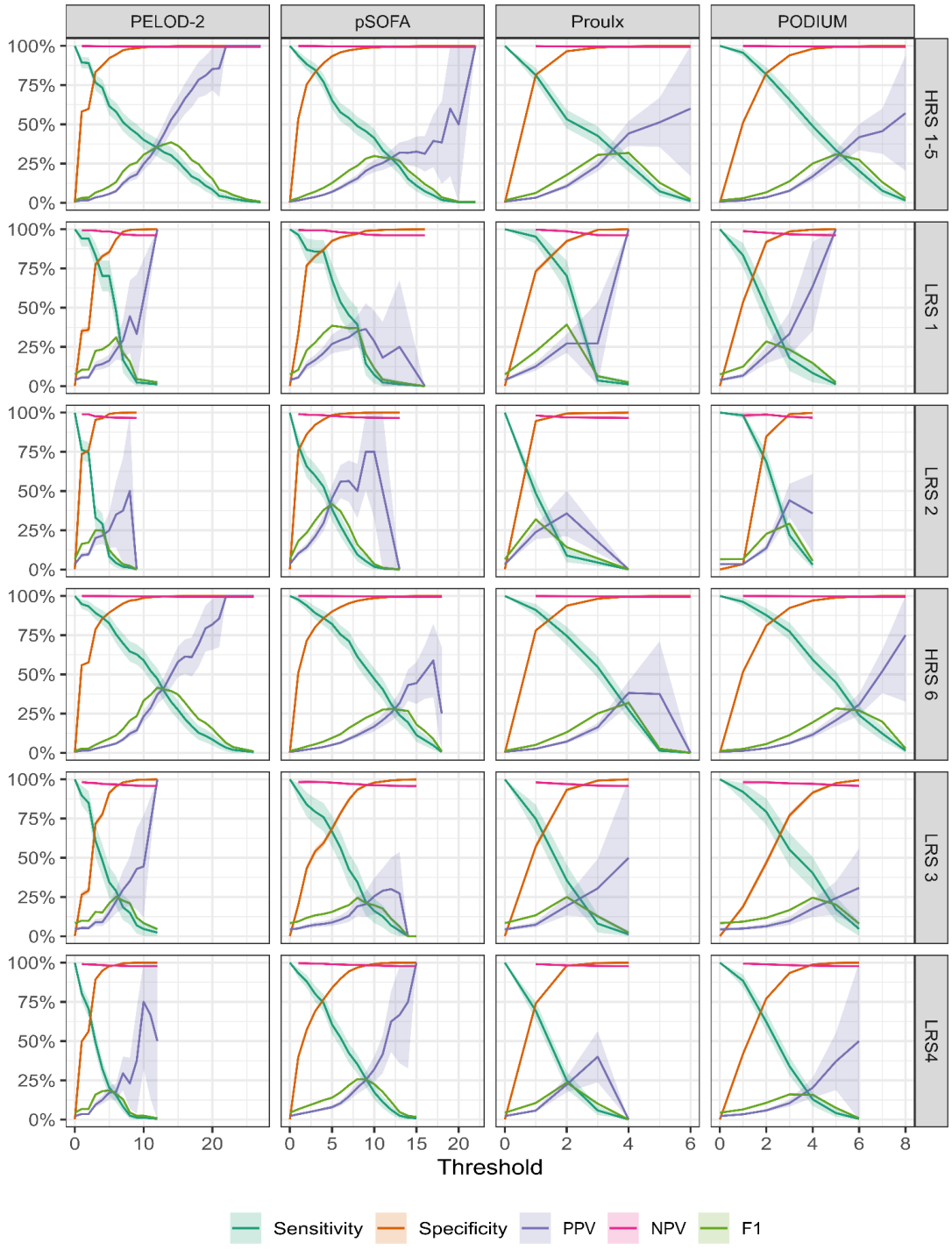
eFigure 7 shows the performance of the Phoenix Sepsis Score (across the entire range from 0 to 13 points) across sites and in comparison to other validated pediatric organ dysfunction scores and criteria to predict early (first 72 hours of the encounter) mortality or extracorporeal membrane oxygenation (ECMO) in patients with suspected infection in the first 24 hours. All organ dysfunctions are evaluated across their respective full ranges, with higher scores indicating more organ dysfunction burden. The scores for IPSCC, Proulx, and PODIUM are based on the count of organ dysfunctions. More information about these scores is provided in the Methods and eTable 2. Bolded values indicate the best-performing score for the respective dataset and performance measure. The performance is presented as both quantitative AUPRC (top) and AUROC (bottom), with 95% confidence intervals (calculated using Logit transform and are shown below each point estimate of performance), as well as visually using a color heatmap. Shading indicates highest (darkest) to lowest (lightest) in each row, blue for AUPRC and yellow for AUROC. AUPRC is the area under a curve drawn with sensitivity (also referred to as "recall") and positive predictive value (also referred to as "precision"), across all potential thresholds for the points in the scores. AUPRC is a more reliable classifier performance metric than AUROC when the classes are imbalanced, typically when mortality is very low (as in this study). AUROC is the area under a curve drawn with false positive rate on the x-axis and true positive rate on y-axis, again across all potential thresholds for the points in the scores. In this study, it is an indicator of how well a classifier can rank encounters with respect to mortality risk. IPSCC, International Pediatric Sepsis Consensus Conference, PELOD-2, Pediatric Logistic Organ Dysfunction, version 2; pSOFA, pediatric Sequential Organ Failure Assessment; PODIUM, Pediatric Organ Dysfunction Information Update Mandate.

eFigure 8A-B. Performance of the Phoenix Sepsis Score and other sepsis scores (A) and other organ dysfunction scores (B) to predict mortality across all thresholds

A. Performance of the Phoenix Sepsis Score and other sepsis scores to predict mortality across all thresholds



B. Performance of organ dysfunction scores to predict mortality across all thresholds



eFigures 8A-B show the mortality prediction performance across all possible thresholds of the Phoenix Sepsis Score in comparison to other sepsis scores as well as pediatric organ dysfunction scores and criteria in patients with suspected infection in the first 24 hours of an encounter. Each score is evaluated across its entire range, which is shown on the "Threshold" x-axis. All potential thresholds or "cut points" are assessed. The performance is presented as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 measure (the harmonic mean of PPV and sensitivity), with 95% confidence intervals. IPSCC, International Pediatric Sepsis Consensus Conference, PELOD-2, Pediatric Logistic Organ Dysfunction, version 2; pSOFA, pediatric Sequential Organ Failure Assessment; PODIUM, Pediatric Organ Dysfunction Information Update Mandate. The 95% confidence intervals were calculated using Logit transform. IPSCC, Proulx, and PODIUM are based on the count of organ dysfunctions.

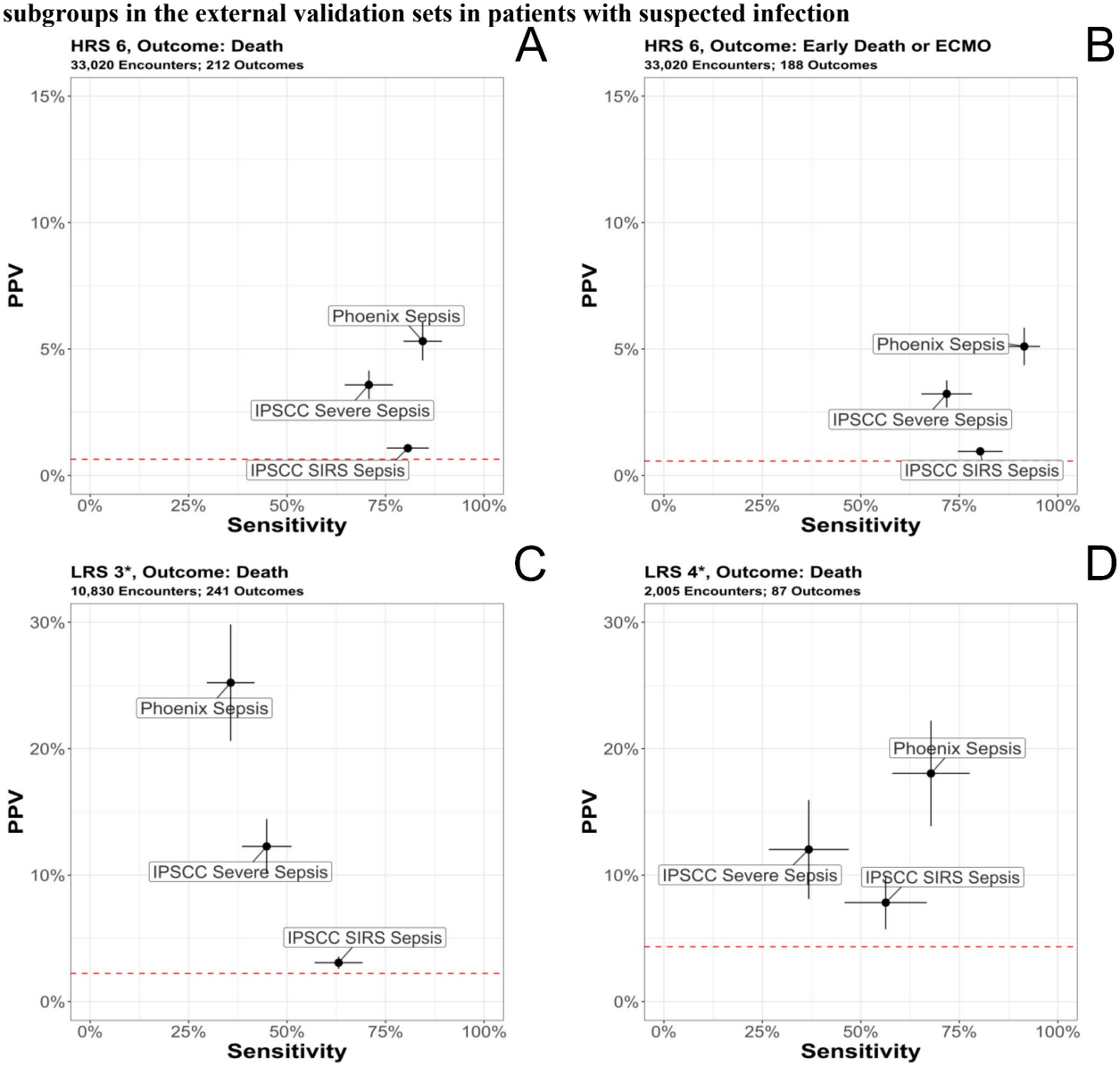
eFigure 9. The Phoenix-8 organ dysfunction score

	0 points	1 point	2 points	3 points
Respiratory (0-3 points)	P/F ≥400 and S/F ≥292	P/F <400 on any respiratory support or S/F <292 on any respiratory support	P/F 101-200 and IMV or S/F 149-220 and IMV	P/F <100 and IMV or S/F <148 and IMV
Cardiovascular (0-6 points)		1 point each (up to 3 points) for:	2 points each (up to 6 points) for:	
	• No vasoactive medications • Lactate <5 mmol/L	• 1 vasoactive medications • Lactate 5-10.9 mmol/L	• ≥2 vasoactive medications • Lactate ≥11 mmol/L	
Age-based	• MAP (mmHg)	• MAP (mmHg)	• MAP (mmHg)	
<1 month	>30	17-30	<17	
1 to 11 months	>38	25-38	<25	
1 to <2 years	>43	31-43	<31	
2 to <5 years	>44	32-44	<32	
5 to <12 years	>48	36-48	<36	
12 to 17 years	>51	38-51	<38	
Coagulation (0-2 points)	• Platelets ≥100 K/μL • INR ≤1.3 • D-Dimer ≤2 mg/L FEU • Fibrinogen ≥100 mg/dL	1 point each (max. 2 points) for: • Platelets <100 K/μL • INR >1.3 • D-Dimer >2 mg/L FEU • Fibrinogen <100 mg/dL		
Neurologic (0-2 points)	• GCS >10 • Pupils reactive	GCS≤10	Fixed pupils	
Endocrine (0-1 point)	Blood glucose 50 to 150 mg/dL	Blood glucose <50 or >150 mg/dL		
Immunologic (0-1 point)	• ANC >500 and • ALC >1000 cells/mm3	• ANC <500 and/or • ALC <1000 cells/mm3		

Renal				
(0-1 point) **Age-based** **<1 month** **1 to 11 months** **1 to <2 years** **2 to <5 years** **5 to <12 years** **12 to 17 years**	 Creatinine (mg/dL) <0.8 <0.3 <0.4 <0.6 <0.7 <1.0	 Creatinine (mg/dL) ≥0.8 ≥0.3 ≥0.4 ≥0.6 ≥0.7 ≥1.0		
Hepatic **(0-1 point)**	• Total bilirubin <4mg/dL and • ALT≤102 IU/L	• Total bilirubin ≥4mg/dL and/or • ALT>102 IU/L		

P/F, PaO_2/FiO_2 ratio; *S/F*, SpO_2/FiO_2 ratio; *IMV*, invasive mechanical ventilation; *mo.*, month(s); *MAP*, mean arterial pressure; INR, international normalized ratio of prothrombin time; GCS, Glasgow coma scale score; *ANC*, absolute neutrophil count; *ALC*, absolute lymphocyte count; *ALT*, alanine aminotransferase. **Notes for use:** The score may be calculated in the absence of some variables (e.g., even if lactate level is not measured and vasoactive medications are not used, a cardiovascular score can still be ascertained using blood pressure). Subscores with no variables measured are assigned a score of 0. It is expected that laboratory tests and other measurements will be obtained at the discretion of the medical team based on clinical judgment. Unmeasured variables contribute no points to the score. The respiratory dysfunction of 1 point can be assessed in any patient on oxygen, high-flow, non-invasive positive pressure, or IMV respiratory support, and includes P/F <200 and S/F <220 in children who are not on IMV. S/F ratio is only calculated if SpO2 is 97% or less. Use measured MAP preferentially (invasive arterial if available or non-invasive oscillometric), and if measured MAP is not available, a calculated MAP (1/3*systolic + 2/3*diastolic) may be used as an alternative. Lactate can be arterial or venous. Lactate reference range is 0.5-2.2 mmol/L. Vasoactive medications include any dose of epinephrine, norepinephrine, dopamine, dobutamine, milrinone, and/or vasopressin (for shock).

eFigure 10. Comparison of the sensitivity and positive predictive value of the novel Phoenix Sepsis Criteria with the current IPSCC Sepsis and Severe Sepsis criteria across outcomes and patient subgroups in the external validation sets in patients with suspected infection



eFigure 10 shows the positive predictive value (PPV, or precision) and sensitivity for the Phoenix sepsis criteria compared to the 2005 International Pediatric Sepsis Consensus Conference (IPSCC) criteria for sepsis in children with suspected infection. The Phoenix sepsis criteria was based on achieving ≥2 points in the Phoenix Sepsis Score among patients with suspected infection in the first 24 hours of an encounter. The IPSCC sepsis and severe sepsis criteria were based on the systemic inflammatory response syndrome (SIRS) and IPSCC-based organ dysfunction among patients with suspected infection in the first 24 hours of an encounter. The baseline rate of the outcome in each group (death and early death or extracorporeal membrane oxygenation [ECMO]) is shown as a horizontal dashed red line. Confidence intervals for each component (sensitivity, PPV) are shown as bands from each point in the plane representing that component (e.g., confidence intervals for PPV are parallel to the y-axis). When a confidence band is not visible, that means that it is narrow enough to be completely hidden by the point. These figures are similar to AUPRCs except at a single threshold (e.g., yes/no binary sepsis criteria met in patients with ≥2 points in the Phoenix Sepsis Score) instead of across the entire range of possible points in the curve (e.g., 0-13 points of the Phoenix Sepsis Score, which is shown in Figure 2). Better performing criteria on these figures will be closer to the top right corner of the figure. There is always a tradeoff between sensitivity and PPV for the different outcomes, with more sensitive criteria usually having lower PPV, and more specific criteria usually having higher

PPV and lower sensitivity. Criteria that are close to the baseline outcome rate (horizontal dashed red line) have poor predictive value. The comparison is shown for higher resource setting (HRS) site 6 (the HRS external validation site) with encounter mortality (A) and death in the first 72 hours or use of ECMO (B) as the outcomes (or prediction targets). Panels C and D show the same comparison at lower resource setting (LRS) sites 3-4 (the LRS external validation sites). *At LRS sites 3 and 4, some of the Phoenix Sepsis Score and IPSCC data inputs (e.g., invasive mechanical ventilation, Glasgow Coma Scale score) are not recorded, even when they are performed, thus the assessment of the criteria performance at those sites is limited.

eTable 7. Diagnostic performance measures of the sepsis and septic shock criteria in the entire development set (derivation and internal validation)

Subset	Criteria	TN	FN	FP	TP	% Baseline Mortality	% Met criteria	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
Outcome = Death											
HRS 1-5	Phoenix Sepsis	133813	323	9517	726	0.7	7.1	69 (66, 72)	93 (93, 93)	7 (7, 8)	100 (100, 100)
HRS 1-5	IPSCC SIRS Sepsis	81998	269	61332	780	0.7	43.0	74 (72, 77)	57 (57, 57)	1 (1, 1)	100 (100, 100)
HRS 1-5	IPSCC Severe Sepsis	131076	433	12254	616	0.7	8.9	59 (56, 62)	91 (91, 92)	5 (4, 5)	100 (100, 100)
LRS 1	Phoenix Sepsis	7230	67	1009	289	4.3	15.1	81 (77, 85)	88 (87, 88)	22 (20, 25)	99 (99, 99)
LRS 1	IPSCC SIRS Sepsis	4649	162	3590	194	4.3	44.0	54 (49, 60)	56 (55, 57)	5 (4, 6)	97 (96, 97)
LRS 1	IPSCC Severe Sepsis	7035	181	1204	175	4.3	16.0	49 (44, 54)	85 (85, 86)	13 (11, 14)	97 (97, 98)
LRS 2	Phoenix Sepsis	19251	508	99	152	3.4	1.3	23 (20, 26)	99 (99, 100)	61 (55, 67)	97 (97, 98)
LRS 2	IPSCC SIRS Sepsis	10198	149	9152	511	3.4	48.3	77 (74, 81)	53 (52, 53)	5 (5, 6)	99 (98, 99)
LRS 2	IPSCC Severe Sepsis	18808	429	542	231	3.4	3.9	35 (31, 39)	97 (97, 97)	30 (27, 33)	98 (98, 98)
Outcome = Early Death or ECMO											
HRS 1-5	Phoenix Sepsis	134074	62	9684	559	0.4	7.1	90 (88, 92)	93 (93, 93)	5 (5, 6)	100 (100, 100)
HRS 1-5	IPSCC SIRS Sepsis	82160	107	61598	514	0.4	43.0	83 (80, 86)	57 (57, 57)	1 (1, 1)	100 (100, 100)
HRS 1-5	IPSCC Severe Sepsis	131338	171	12420	450	0.4	8.9	72 (69, 76)	91 (91, 92)	3 (3, 4)	100 (100, 100)
LRS 1	Phoenix Sepsis	7283	14	1207	91	1.2	15.1	87 (80, 93)	86 (85, 87)	7 (6, 8)	100 (100, 100)
LRS 1	IPSCC SIRS Sepsis	4771	40	3719	65	1.2	44.0	62 (53, 71)	56 (55, 57)	2 (1, 2)	99 (99, 99)
LRS 1	IPSCC Severe Sepsis	7172	44	1318	61	1.2	16.0	58 (49, 68)	84 (84, 85)	4 (3, 6)	99 (99, 100)
LRS 2	Phoenix Sepsis	19472	287	125	126	2.1	1.3	31 (26, 35)	99 (99, 99)	50 (44, 56)	99 (98, 99)
LRS 2	IPSCC SIRS Sepsis	10265	82	9332	331	2.1	48.3	80 (76, 84)	52 (52, 53)	3 (3, 4)	99 (99, 99)
LRS 2	IPSCC Severe Sepsis	19002	235	595	178	2.1	3.9	43 (38, 48)	97 (97, 97)	23 (20, 26)	99 (99, 99)
Criteria = Septic Shock											
HRS 1-5	Phoenix Septic Shock	138421	456	4909	593	0.7	3.8	57 (54, 60)	97 (96, 97)	11 (10, 12)	100 (100, 100)
HRS 1-5	IPSCC Septic Shock	135838	548	7492	501	0.7	5.5	48 (45, 51)	95 (95, 95)	6 (6, 7)	100 (100, 100)
LRS 1	Phoenix Septic Shock	7499	86	740	270	4.3	11.8	76 (71, 80)	91 (90, 92)	27 (24, 29)	99 (99, 99)
LRS 1	IPSCC Septic Shock	7211	189	1028	167	4.3	13.9	47 (42, 52)	88 (87, 88)	14 (12, 16)	97 (97, 98)
LRS 2	Phoenix Septic Shock	19252	508	98	152	3.4	1.2	23 (20, 26)	99 (99, 100)	61 (55, 67)	97 (97, 98)
LRS 2	IPSCC Septic Shock	18834	439	516	221	3.4	3.7	33 (30, 37)	97 (97, 98)	30 (27, 33)	98 (98, 98)

TN, true negative*; FN,* false negative*; FP,* false positive*; TP,* true positive*; HRS*, higher resource settings sites; *LRS*, lower resource setting sites; *CI*, confidence interval; *IPSCC*, International Pediatric Sepsis Consensus Conference criteria; *SIRS*, systemic inflammatory response syndrome; *ECMO*, extracorporeal membrane oxygenation.

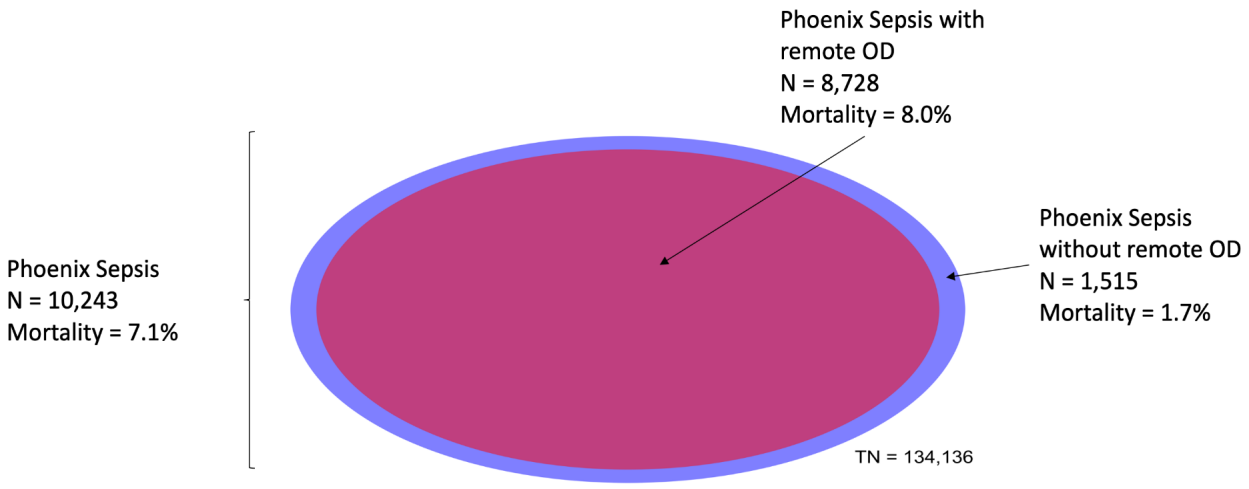
eTable 8. Diagnostic performance measures of the Phoenix sepsis criteria across sensitivity analyses in the entire development set (derivation and internal validation)

Subset	TN	FN	FP	TP	% Baseline Mortality	% Phoenix sepsis	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
Age groups in HRS 1-5										
< 1 month	5521	20	654	106	2.0	12.1	84 (78, 91)	89 (89, 90)	14 (11, 16)	100 (99, 100)
1 to 11 months	19184	48	1440	171	1.1	7.7	78 (73, 84)	93 (93, 93)	11 (9, 12)	100 (100, 100)
12 to 23 months	14863	42	1021	57	0.6	6.7	58 (48, 67)	94 (93, 94)	5 (4, 7)	100 (100, 100)
24 to 59 months	27989	53	1656	114	0.6	5.9	68 (61, 75)	94 (94, 95)	6 (5, 8)	100 (100, 100)
60 to 143 months	39901	81	2378	134	0.5	5.9	62 (56, 69)	94 (94, 95)	5 (4, 6)	100 (100, 100)
144 months or older	26355	79	2368	144	0.8	8.7	65 (58, 71)	92 (91, 92)	6 (5, 7)	100 (100, 100)
Age groups in LRS 1										
< 1 month	2896	44	607	190	6.7	21.3	81 (76, 86)	83 (81, 84)	24 (21, 27)	99 (98, 99)
1 to 11 months	1752	6	198	42	2.5	12.0	88 (78, 97)	90 (89, 91)	18 (13, 22)	100 (99, 100)
12 to 23 months	517	3	42	13	2.9	9.6	81 (62, 100)	92 (90, 95)	24 (12, 35)	99 (99, 100)
24 to 59 months	605	3	37	10	2.0	7.2	77 (54, 100)	94 (92, 96)	21 (10, 33)	100 (99, 100)
60 to 143 months	654	6	57	14	2.8	9.7	70 (50, 90)	92 (90, 94)	20 (10, 29)	99 (98, 100)
144 months or older	806	5	68	20	2.9	9.8	80 (64, 96)	92 (90, 94)	23 (14, 31)	99 (99, 100)
Age groups in LRS 2										
< 1 month	381	13	7	6	4.9	3.2	32 (11, 52)	98 (97, 100)	46 (19, 73)	97 (95, 98)
1 to 11 months	11694	375	36	95	4.0	1.1	20 (17, 24)	100 (100, 100)	73 (65, 80)	97 (97, 97)
12 to 23 months	3534	64	11	29	2.6	1.1	31 (22, 41)	100 (100, 100)	73 (59, 86)	98 (98, 99)
24 to 59 months	2255	34	10	16	2.2	1.1	32 (19, 45)	100 (99, 100)	62 (43, 80)	99 (98, 99)
60 to 143 months	827	14	12	5	2.3	2.0	26 (7, 46)	99 (98, 99)	29 (8, 51)	98 (97, 99)
144 months or older	560	8	23	1	1.5	4.1	11 (0, 32)	96 (94, 98)	4 0, 12)	99 (98, 100)
Other subsets in HRS 1-5										
Excluding OR cases	113173	191	6771	542	0.6	6.1	74 (71, 77)	94 (94, 94)	7 (7, 8)	100 (100, 100)
ICU encounters	14994	251	7748	692	4.1	35.6	73 (71, 76)	66 (65, 67)	8 (8, 9)	98 (98, 99)
No known prior comorbidity	91086	99	5118	458	0.6	5.8	82 (79, 85)	95 (95, 95)	8 (7, 9)	100 (100, 100)

TN, true negative*; FN,* false negative*; FP,* false positive*; TP,* true positive*; HRS*, higher resource settings sites; *LRS*, lower resource setting sites; *CI*, confidence interval; *OR*, operating room; *ICU*, intensive care unit; *PCCC*, pediatric complex chronic conditions.

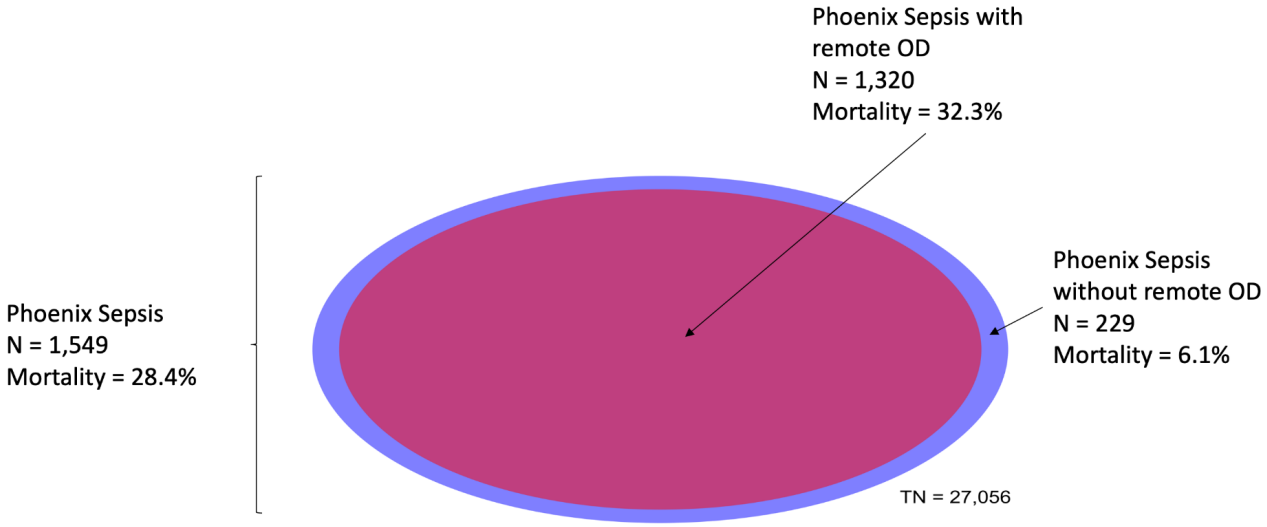
eFigure 11A-B. Venn diagram of sepsis with remote organ dysfunction in the development set

A. Venn diagram of sepsis with remote organ dysfunction for higher resource sites in the development set



OD, Organ Dysfunction; *TN*, true negatives.

B. Venn diagram of sepsis with remote organ dysfunction for lower resource sites in the development set



OD, Organ Dysfunction; *TN*, true negatives.

eAppendix 2. Clinical vignettes with calculation of the Phoenix Sepsis Score and the Phoenix Sepsis Criteria

A previously healthy 3-year-old girl presents to an emergency department in Lima, Peru, with a temperature of 39ºC, tachycardia, and irritability. Blood pressure with an oscillometric device is 67/32 mmHg (mean arterial pressure of 43 mmHg). She is given fluid resuscitation per local best practice guidelines, is started on broad spectrum antibiotics, and blood and urine cultures are sent. After an hour, she becomes hypotensive again and she is started on a norepinephrine drip. A complete blood count reveals leukocytosis, mild anemia, and a platelet count of 95 K/μL.

Phoenix Sepsis Score: 0 respiratory points (no hypoxemia or respiratory support) + 2 cardiovascular points (1 for low mean arterial pressure for age, 1 for use of a vasoactive medication) + 1 coagulation points (for low platelet count) + 0 neurologic points (irritability would result in a Glasgow Coma Scale of approximately 14) = 3 points.
Phoenix Sepsis Criteria: The patient has suspected infection, ≥2 points of the Phoenix Sepsis Score, and ≥1 cardiovascular points, so she meets criteria for **septic shock**.

A 6-year-old boy with a history of prematurity presents with respiratory distress to his pediatrician's office in Tucson, Arizona. He is noted to have a temperature of 38.7ºC, tachypnea, crackles in the left lower quadrant on chest auscultation, and an oxygen saturation of 89% on room air. He is started on supplemental oxygen and is transported to the local emergency department via ambulance. In the emergency department, a chest X-ray shows a consolidation in the left lower lobe and hazy bilateral lung opacities, so he is started on antibiotics for a suspected bacterial pneumonia. His respiratory status worsens, and he is started on non-invasive positive pressure ventilation. While awaiting to be admitted, his level of consciousness deteriorates rapidly: with nailbed pressure he only opens his eyes briefly, moans in pain, and withdraws his hand (Glasgow Coma Scale: 2 for eye response + 2 for verbal response + 4 for motor response = 8). He is intubated using rapid sequence induction and placed on a conventional ventilator. During this time, his lowest mean arterial pressure using a non-invasive oscillometric device is 52 mmHg and he receives a fluid bolus. He is then transferred to the pediatric intensive care unit where he requires a high positive end expiratory pressure and an FiO_2 of 0.45 to achieve an oxygen saturation of 92% (S/F ratio: 204). Complete blood count and lactate level reveal a platelet count of 120 K/μL and a serum lactate of 2.9 mmol/L. Given his platelet count below the normal reference range, a coagulation panel is sent, which reveals an INR of 1.7, a D-Dimer of 4.4 mg/L, and a fibrinogen of 120 mg/dL.

Phoenix Sepsis Score: 2 respiratory points (for an S/F ratio <292 on invasive mechanical ventilator) + 0 cardiovascular points (mean arterial pressure >48 mmHg and Lactate level <5 mmol/L) + 2 coagulation points (for high INR and D-Dimer) + 1 neurologic point (Glasgow Coma Scale ≤10) = 5 points.
Phoenix Sepsis Criteria: The patient has a suspected infection, ≥2 points of the Phoenix Sepsis Score, and 0 cardiovascular points, so he meets criteria for **sepsis**.

eReferences

1. Wolpert DH. Stacked generalization. *Neural Netw*. 1992;5(2):241-259.

2. Breiman L. Stacked regressions. *Mach Learn*. 1996;24(1):49-64.

3. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63.

4. Zeng X, Yu G, Lu Y, et al. PIC, a paediatric-specific intensive care database. *Sci Data*. 2020;7(1):14.

5. Tange O. *GNU Parallel 2018*. Lulu.com; 2018.

6. Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)*. 2016;4(1):1244.

7. Reilly PL, Simpson DA, Sprod R, Thomas L. Assessing the conscious level in infants and young children: a paediatric version of the Glasgow Coma Scale. *Childs Nerv Syst*. 1988;4(1):30-33.

8. Martin B, DeWitt PE, Scott HF, Parker S, Bennett TD. Machine Learning Approach to Predicting Absence of Serious Bacterial Infection at PICU Admission. *Hosp Pediatr*. 2022;12(6):590-603.

9. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*. 2016;315(8):762-774.

10. Raith EP, Udy AA, Bailey M, et al. Prognostic Accuracy of the SOFA Score, SIRS Criteria, and qSOFA Score for In-Hospital Mortality Among Adults With Suspected Infection Admitted to the Intensive Care Unit. *JAMA*. 2017;317(3):290-300.

11. Fitzgerald JC, Basu RK, Fuhrman DY, et al. Renal Dysfunction Criteria in Critically Ill Children: The PODIUM Consensus Conference. *Pediatrics*. 2022;149(1 Suppl 1):S66-S73.

12. Weiss SL, Peters MJ, Alhazzani W, et al. Surviving Sepsis Campaign International Guidelines for the Management of Septic Shock and Sepsis-Associated Organ Dysfunction in Children. *Pediatr Crit Care Med*. 2020;21(2):e52-e106.

13. Clarke B. Comparing Bayes model averaging and stacking when model approximation error cannot be ignored. *J Mach Learn Res*. 2003;4:683-712.