

Predicting recovery after urosepsis: a multicentre machine learning study using SOFA-based organ dysfunction measures and microbiology

Ana Djurica

ana.durica.21@ucl.ac.uk

University College London

Stella Champeaux

University College London

Truls E. Bjerklund Johansen

5Institute of Clinical Medicine, University of Oslo, Norway

Florian Wagenlehner

Justus-Liebig University Giessen

Zafer Tandogdu

University College London

Ivana Drobnyak

University College London

Article

Keywords:

Posted Date: October 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10575290/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Predicting recovery after urosepsis: a multicentre machine learning study using SOFA-based organ dysfunction measures and microbiology

Authors

Ana Djurica^{1,2,*}, Stella Champeaux³, Truls E. Bjerklund Johansen^{4,5}, Florian Wagenlehner⁶, Zafer Tandogdu^{7,8,†}, Ivana Drobnjak^{2,†}

Affiliations

1 - Institute of Health Informatics, University College London, London, United Kingdom

2 – Department of Computer Science, University College London, London, United Kingdom

3 - Faculty of Medical Sciences, University College London, London, United Kingdom

4 - Institute of Clinical Medicine, University of Oslo, Oslo, Norway

5 - Institute of Clinical Medicine, University of Aarhus, Aarhus, Denmark

6 - Clinic for Urology, Pediatric Urology and Andrology, University Clinic Giessen, Justus-Liebig University Giessen, Giessen, Germany

7 - Division of Surgery and Interventional Sciences, University College London, London, United Kingdom

8 - Department of Urology, University College London Hospitals NHS Foundation Trust, London, United Kingdom

*Corresponding author. E-mail: ana.durica.21@ucl.ac.uk

†These authors jointly supervised this work.

Background: Urosepsis accounts for up to 40% of adult sepsis cases, yet prognostic tools have focused on mortality rather than recovery. Among survivors, recovery trajectories vary substantially and cannot be anticipated from composite severity scores alone. No validated framework exists for predicting recovery failure after the acute episode.

Methods: Using prospectively collected data from 354 adults with confirmed urosepsis enrolled across 34 European centres in the SERPENS cohort, interpretable machine learning models were developed to predict failure to achieve full recovery at Day 3 and Day 30. Clinical, microbiological, and Sequential Organ Failure Assessment (SOFA) data were analysed using logistic regression and XGBoost models with total versus organ-specific SOFA representations.

Results: Using baseline total SOFA, models achieved strong discrimination for short-term recovery prediction (AUROC 0.87) and good discrimination for long-term prediction (AUROC 0.79). Incorporating longitudinal SOFA measurements improved long-term prediction to AUROC 0.87. Models using organ-specific SOFA representations improved prediction at both horizons, achieving AUROCs of 0.90 (Day 3) and 0.83 (Day 30), with renal, CNS, and respiratory dysfunction emerging as dominant predictors. Microbiological features provided additional prognostic information, and performance remained broadly consistent during cross-centre validation.

Conclusions: This study presents the first recovery-focused prognostic framework for confirmed urosepsis, demonstrating that routinely collected organ dysfunction and microbiological data can predict recovery trajectories beyond the acute phase.

1. Introduction

Sepsis, defined as life-threatening organ dysfunction caused by a dysregulated host response to infection, remains a major global health challenge (1). In 2017, an estimated 48.9 million cases and 11 million sepsis-related deaths were reported worldwide, disproportionately affecting populations in low- and middle-income countries (2). Urosepsis, sepsis arising from a urinary tract infection, accounts for approximately 20-40% of all adult sepsis cases, making it one of the most common sources of sepsis encountered in clinical practice (3). At its most severe, urosepsis progresses to septic shock in 10–30% of cases (4), and disproportionately affects older adults, in whom both incidence and sepsis-related mortality remain elevated even after adjustment for comorbidity and illness severity (5).

Mortality alone does not capture the full burden of sepsis, as many survivors experience persistent organ dysfunction, impaired functional recovery, and reduced quality of life following acute infection (6,7). This burden is particularly relevant in urosepsis, where mortality is comparatively low but morbidity remains substantial. In the prospective SERPENS cohort, 30-day mortality was 2.8%, yet organ dysfunction affected 84% of patients within 30 days and persisted in nearly one-quarter at Day 30 (8). These findings underscore why recovery trajectories, rather than survival alone, represent a clinically meaningful endpoint.

Despite this burden, the clinical trajectory after urosepsis diagnosis remains difficult to anticipate. Clinicians currently lack tools to distinguish patients who will follow divergent recovery trajectories despite similar initial severity, limiting early risk stratification beyond mortality-focused frameworks. Among patients who survive the initial episode and receive standard treatment (antimicrobial therapy, haemodynamic support, and source control) outcomes diverge substantially. In the SERPENS cohort, organ dysfunction persisted in 24% of patients at Day 30, and 5% developed new organ failure during follow-up despite initial stabilisation (8). This heterogeneity is not explained by severity at presentation alone. Routine clinical and laboratory measures available at diagnosis reflect the acute inflammatory insult, but do not capture the evolving host response that ultimately determines whether organ function resolves or persists (9). Clinicians are therefore left without a principled basis for stratifying the intensity of early monitoring, identifying which patients warrant treatment escalation, or planning post-acute care pathways for those at risk of prolonged recovery (1,8,10).

No validated prognostic tool exists for predicting recovery failure, as opposed to mortality, in patients with confirmed urosepsis beyond the initial acute phase. Existing machine learning tools for sepsis have focused primarily on early detection or short-term mortality prediction (predominantly in ICU populations) rather than to inform prognosis among patients with established sepsis (11–13). Moreover, many models rely on retrospective electronic health record data, emphasise static baseline features, and do not explicitly address recovery trajectories or longer-term organ dysfunction. This is a critical gap: the patients who survive urosepsis but fail to fully recover represent a distinct, underserved clinical phenotype whose identification requires a prognostic framework rather than a detection one.

Building on the SERPENS cohort of 354 prospectively recruited and microbiologically confirmed urosepsis cases, we developed and internally validated interpretable machine learning models to predict failure to achieve full recovery at two clinically meaningful horizons: Day 3, capturing early trajectory divergence after initial therapy, and Day 30, reflecting post-acute recovery beyond survival alone. Models were constructed using prospectively collected clinical variables, with a particular focus on Sequential Organ Failure Assessment (SOFA) scores as a measure of organ dysfunction burden, and microbiology. Having established predictive performance using SOFA-based models, we then examined whether this prognostic signal was better captured by the aggregated total SOFA score or by its constituent organ-specific sub-scores, providing insight into the relative importance of global versus organ-specific dysfunction in determining recovery trajectories. Model generalisability was further assessed through cross-centre validation, designating held-out centre subsets

to evaluate discrimination on completely unseen clinical sites across the 34 institutions contributing to the SERPENS cohort.

This recovery-focused prognostic framework represents, to our knowledge, the first machine learning approach applied to a urosepsis cohort that targets organ dysfunction rather than mortality as its primary endpoint. By explicitly modelling two clinically distinct timepoints (early trajectory and longer-term recovery), and by interrogating total SOFA and organ-specific SOFA components as candidate predictors, this approach is designed to support three consequential clinical decisions: stratifying monitoring intensity in the first days after diagnosis, identifying patients who may require treatment escalation, and anticipating post-sepsis care needs in those at risk of incomplete recovery (8,14).

2.Methods

2.1 Study design and cohort

This study used data from the SERPENS cohort (NCT02380170), a prospective, multicentre observational study of microbiologically confirmed urosepsis conducted across 34 secondary and tertiary hospitals in Europe between 2014 and 2018 as part of the Global Prevalence of Infections in Urology (GPIU) project. The present work represents a post-hoc secondary analysis of the SERPENS dataset. Ethical approval for the original study was granted by the Justus Liebig University Ethics Committee, Giessen, Germany (AZ: 77/14). All methods were performed in accordance with the relevant guidelines and regulations. Informed consent was obtained from all patients prior to enrolment by local investigators according to the protocol.

2.2 Study subjects

The SERPENS cohort was designed and recruited prior to widespread adoption of the Sepsis-3 definitions. Enrolment therefore followed the contemporaneous Sepsis-2 criteria: a diagnosis of urosepsis required microbiologically confirmed urinary tract infection, via positive urine culture or blood culture with an established urinary source, together with fulfilment of Systemic Inflammatory Response Syndrome (SIRS) criteria (≥ 2 of 4: abnormal temperature, heart rate, respiratory rate, or white blood cell count, with at least one being temperature or white blood cell count). While Sepsis-3 definitions are now widely adopted, the present analysis reflects the enrolment framework used in the original cohort and should be interpreted in that context.

The original cohort comprised 354 eligible patients. Clinical variables were recorded longitudinally at predefined study timepoints: baseline (Day 0) and follow-up assessments on Days 3, 7, and 30. These scheduled observations defined the temporal structure of the dataset and therefore the short-term (Day 3) and long-term (Day 30) prediction horizons evaluated in the present analysis.

Eligibility was defined separately for each prediction horizon based on availability of the required SOFA assessments and outcome information. For the short-term prediction task, patients were included if a total SOFA score was recorded at both baseline and Day 3. For long-term prediction, inclusion required availability of total SOFA scores at baseline (Day 0), Day 3, Day 7 and Day 30. This approach results in a complete-case analysis conditional on availability of longitudinal data.

2.3 Outcome definitions

The primary outcome was recovery status at Day 3, defined as a binary classification of Full Recovery or Failure to Recover. Full Recovery was defined as survival with complete resolution of organ dysfunction, operationalised as SOFA score of 0 at the outcome timepoint. Failure to Recover was defined as either residual organ dysfunction (total SOFA score >0) or death occurring on or before the outcome timepoint. A secondary long-term outcome was defined using the same framework at Day 30.

In the original SERPENS study, outcomes were classified into four health states (recovery, single organ failure, multiple organ failure, or death) based on SOFA-derived organ failure definitions. In the present modelling analysis, recovery was defined more stringently as a total SOFA score of 0 to capture complete resolution of organ dysfunction rather than the absence of overt organ failure.

Deaths were classified as Failure to Recover events at all timepoints. When death was recorded at a follow-up visit, it was assumed to occur after SOFA assessment on that day, and the most recent available SOFA score prior to death was retained for analysis. To avoid temporal leakage, predictor variables were restricted to measurements available prior to the respective prediction horizon (baseline for the Day 3 model and longitudinal inputs up to Day 7 for the Day 30 model).

2.4 Variable identification and encoding

Domain specialists identified and extracted variables considered clinically relevant to predicting recovery outcomes following a diagnosis of urosepsis. The main groups of extracted variables included patient demographics (age, weight, height, sex at birth

and BMI), urological risk factors (ORENUC classification), comorbidities (Charlson Comorbidity Index), microbiological pathogen results from urine cultures, initial antibiotic treatment, susceptibility of pathogens to antibiotics, and SOFA scores. An abbreviated list of initially considered clinically recommended variables is provided in the results for the short-term model (Table 1), with complete lists in the supplementary for the short- and long-term models.

SOFA scores were recorded at baseline (Day 0) and at follow-up timepoints on Days 3, 7 and 30. Microbiological culture results, including pathogen identification from urine samples, were typically available within the first three days of admission and were therefore informative at either the Day 0 or Day 3 inputs. The input structure and temporal data windows for all four models are summarised in Figure 1. Baseline variables collected at Day 0 included demographics, Charlson Comorbidity Index, ORENUC classification, SOFA score, antibiotic treatment, and urine culture results; SOFA scores were additionally recorded at Days 3, 7, and 30. No variables measured after a given prediction timepoint were used as predictors for that timepoint.

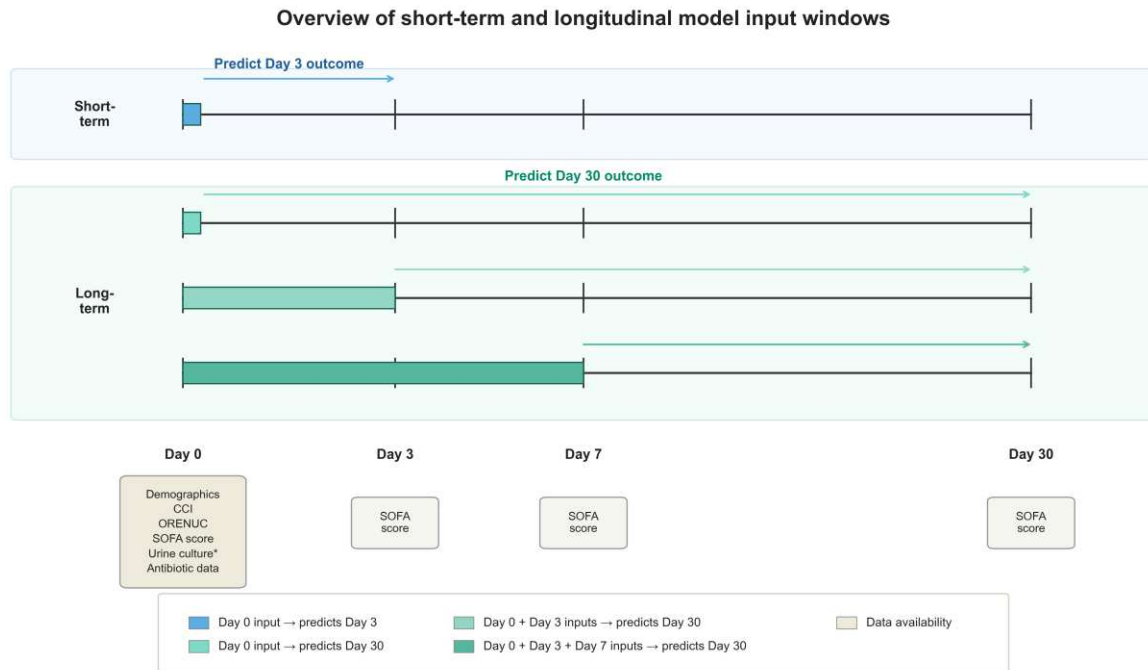


Figure 1. Input structure of four prediction models using progressively expanded temporal data windows. Schematic showing input windows and prediction horizons for short-term (Day 3) in blue and three long-term (Day 30) models in green. The short-term model uses baseline (Day 0) data to predict Day 3 outcome, while three long-term models predict Day 30 outcome using data available up to Day 0, Day 3, or Day 7. Shaded regions indicate the input window for each model, and arrows indicate the corresponding prediction horizon. Clinical variables and SOFA scores were collected at Day 0, Day 3, Day 7, and Day 30. Baseline variables are shown in brown boxes and include demographics, Charlson Comorbidity Index, ORENUC classification, SOFA score, antibiotic treatment and urine culture results; SOFA was additionally recorded at Days 3, 7, and 30.

2.4.1 SOFA scores

SOFA score quantifies the degree of organ dysfunction across six physiological domains — respiratory, coagulation, liver, cardiovascular, central nervous system, and renal function — each scored from 0 (normal function) to 4 (most severe dysfunction), giving a total score ranging from 0 to 24. At each timepoint, organ-specific sub-scores were recorded across all six domains and retained in their original numerical form to preserve severity information and maximise predictive signal. Overall severity was additionally represented as the total SOFA score, calculated as the numerical sum of the six sub-scores, and included as a separate continuous predictor. Models were constructed using either total SOFA or the full set of six organ-specific sub-scores as

mutually exclusive inputs, enabling direct comparison of aggregated and disaggregated representations of organ dysfunction.

2.4.2 Microbiological data (pathogens, antibiotics and susceptibility)

In the original study dataset, urine culture results recorded detected pathogens in two fields (primary and secondary), representing the principal and additional organisms identified within a sample. For the present analysis, pathogens identified across both fields were combined to represent organism presence within a single patient sample. Several organisms were infrequently identified and were therefore grouped into clinically meaningful categories. Six commonly observed organisms were retained as distinct groups: *Escherichia coli*, *Klebsiella* spp., *Enterococcus* spp., *Proteus* spp., *Enterobacter* spp., and *Pseudomonas aeruginosa*. All remaining organisms were grouped into a single “Other” category (Supplementary Table S4).

For modelling, each pathogen group was encoded using two binary variables representing organism presence (yes/no) and culture availability (known/unknown). This approach allowed unavailable culture data to be represented explicitly without conflating missing cultures with true organism absence.

For antibiotic exposure, 19 antibiotic classes administered within the first 24 hours following confirmed urosepsis diagnosis (Day 0) were recorded. To reduce dimensionality while preserving clinically relevant information regarding antimicrobial coverage, antibiotics were grouped according to spectrum activity into broad-spectrum and narrow-spectrum categories (classification provided in Supplementary Table S5). Patients were classified as receiving broad-spectrum only therapy if at least one

broad-spectrum agent was administered and no narrow-spectrum agents were administered, narrow-spectrum therapy if only narrow-spectrum agents were administered, and combined therapy if agents from both categories were present. Classification was mutually exclusive and exhaustive at the patient level. Antibiotic classes containing agents with heterogeneous antimicrobial spectra, including penicillins, were classified at the agent level prior to grouping. This approach captured escalation in antimicrobial coverage while avoiding sparsity associated with modelling individual antibiotic classes separately.

Antimicrobial susceptibility profiles were grouped into four categories reflecting increasing levels of antimicrobial resistance: pan-susceptible, non-multi-drug-resistant (non-MDR), multi-drug-resistant (MDR), and unknown. Pan-susceptible organisms showed no resistance to any tested antimicrobial class. Non-MDR organisms were resistant to one or two antimicrobial classes, whereas MDR organisms were resistant to three or more classes. Cases in which susceptibility testing was unavailable were classified as unknown.

For model development, susceptibility categories were converted into binary indicator variables using standard dummy encoding. Non-MDR was selected as the reference category, such that the model estimated the effect of pan-susceptible, MDR, and unknown susceptibility relative to non-MDR organisms. This approach avoided imposing an artificial numerical ordering between categories while retaining clinically interpretable comparisons between resistance states. MDR classification was based on established international definitions used in the original SERPENS study.

2.4.3 Comorbidity and urological risk factors

Data on 18 individual Charlson Comorbidity Index (CCI) conditions were extracted (full list of variables in Table 1). Only binary indicators reflecting the presence or absence of each individual condition were retained; no composite CCI score was calculated. Both any tumour and metastatic tumour were encoded as independent binary indicators, reflecting their independent recording in the source data.

The ORENUC classification (full list in Supplementary Table S6), comprising six urological risk categories, was recorded for all patients. For model development, the “No known risk factor” category was inverted to generate a binary indicator reflecting the presence of any known urological risk factor.

2.5 Data preprocessing

2.5.1 Missing data

Missingness was negligible for non-microbiological variables these observations were excluded during cross-validation and model training (complete-case approach); imputation was not performed.

SOFA scores were available for all patients at Days 0, 3, and 7 unless death occurred prior to the respective assessments. Observations containing missing organ-specific sub-scores were handled using a complete-case approach and excluded during cross-validation and model training.

For baseline microbiological variables, urine culture data were available at Day 0 for 220 of 233 patients (94.4%) in the long-term cohort and 333 of 349 patients (95.4%) in the short-term cohort. To account for this missingness, an additional binary indicator variable representing urine culture availability at baseline was included in the

modelling framework. Pathogen presence variables were encoded as binary indicators and set to zero when culture results were unavailable. Given the low frequency of missing cultures (<6%), this approach preserved the full cohort while allowing the model to account for culture availability. Urine culture variables were included only at Day 0, as cultures represent baseline microbiological information obtained at presentation.

MDR classification was unavailable for 12.6% of patients in the short-term cohort and 12.9% in the long-term cohort. Missing or unknown MDR status was therefore retained as a separate category during feature encoding, allowing the models to preserve all patients while accounting for potentially informative missingness in resistance profiling. No missing data were present following antibiotic feature encoding, as all patients received at least one recorded antibiotic agent within the first 24 hours after diagnosis.

2.5.2 Data normalisation

Continuous variables, including age and total SOFA scores, were standardised using z-score scaling. Scaling parameters were estimated from the training set and subsequently applied to both training and test sets to ensure consistent feature representation. Binary variables were used without further encoding.

2.5 Feature selection

Feature selection was performed using a structured, domain-informed approach designed to preserve the clinical integrity of SOFA representations while reducing redundancy among non-SOFA variables. Feature selection was performed separately

for each prediction horizon (Day 3 and Day 30) to account for potential differences in the relationship between baseline predictors and short-term versus longer-term recovery outcomes. Given the composite structure of the SOFA score, models were constructed using either the total SOFA score or the full set of six organ-specific sub-scores, but not both simultaneously, to enable direct comparison of aggregated and disaggregated representations of organ dysfunction

For models based on total SOFA score, baseline clinical variables were initially screened using univariate analysis (Mann–Whitney U test for continuous variables and χ^2 test for categorical variables), with variables meeting a threshold of $p \leq 0.05$ retained as candidate predictors. Univariate screening was performed on the full cohort prior to model development. In a small number of variables, extremely low cell counts resulted in highly imbalanced contingency tables, limiting the reliability of statistical tests.

For models incorporating organ-specific SOFA sub-scores, all six organ domains were retained irrespective of univariate association with outcome.

Redundancy analysis was performed to assess collinearity and clinical overlap between candidate predictors. Moderate collinearity was identified between height and sex, and between Charlson moderate/severe renal disease and ORENUC nephropathic risk; in each case, the clinically more specific variable (sex and ORENUC nephropathic risk, respectively) was retained.

2.7 Model development and validation

For the primary analysis, models incorporated the aggregated total SOFA score as a continuous predictor, evaluated separately for short-term (Day 3) and long-term (Day 30) recovery outcomes. For the secondary analysis, models incorporated the six

individual organ-specific SOFA sub-scores as separate predictors, again evaluated at both prediction horizons. The overall model development and validation pipeline is summarised in Figure 2.

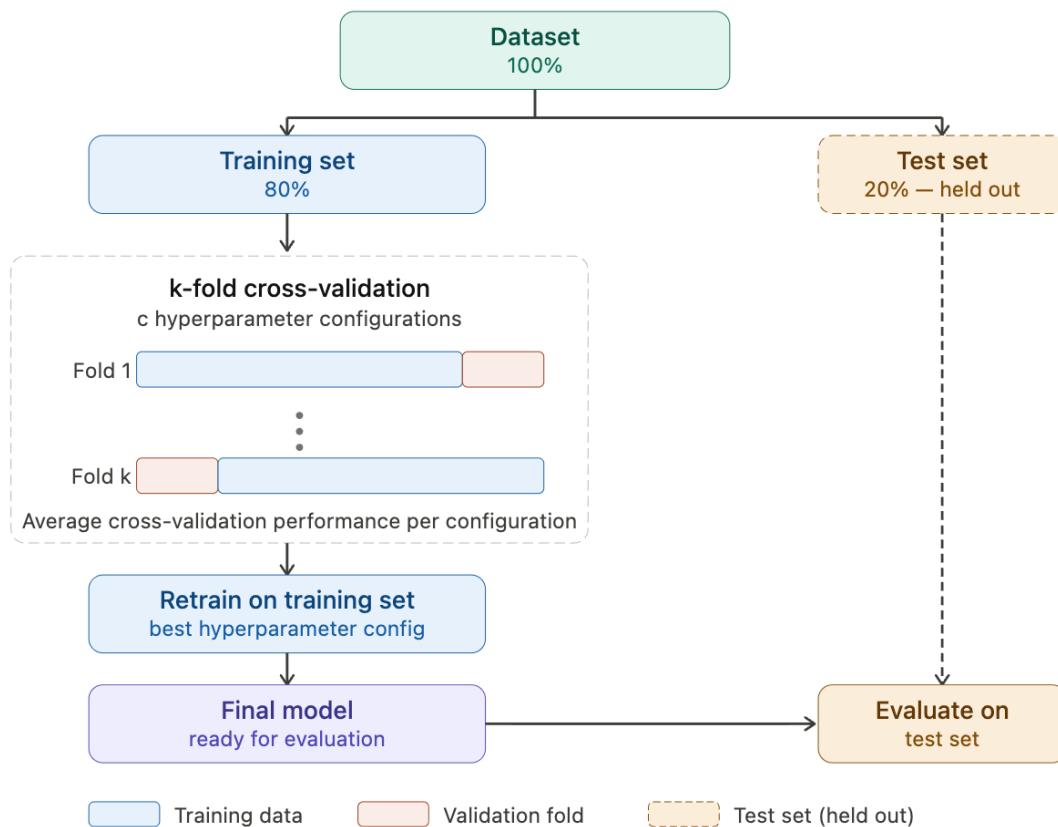


Figure 2. Model development and validation pipeline. Schematic of the data splitting, model training, and evaluation procedure. The dataset was divided into a training set (80%) and a held-out test set (20%). Within the training set, hyperparameter configurations were evaluated using k-fold cross-validation (dashed box), in which each fold was used once as a validation set (orange) and the remaining data for training (blue), and performance was averaged across folds. The best-performing configuration was then used to retrain the model on the full training set, and final performance was evaluated on the held-out test set (dashed outline), which remained unseen during model development.

Logistic regression was included as an interpretable linear baseline, enabling estimation of the direction and magnitude of associations between predictors and recovery outcomes. Extreme gradient boosting (XGBoost) was selected as a non-linear modelling approach capable of capturing complex interactions and non-linear relationships across the demographic, clinical, and microbiological predictor space.

The dataset was split into training (80%) and test (20%) sets using a fixed random seed to ensure reproducibility. Model development and hyperparameter optimisation were performed exclusively on the training set. Hyperparameter tuning was conducted using grid search with stratified 4-fold cross-validation, with performance assessed using the mean area under the receiver operating characteristic curve (AUROC) across folds. Stratification was applied during both train–test splitting and cross-validation to preserve class balance. For logistic regression, regularisation type and strength were optimised; for XGBoost, tree depth, learning rate, and column subsampling parameters were explored (full parameter ranges provided in Supplementary Table S7).

The optimal model configuration was retrained on the full training set and evaluated on the held-out test set. Model discrimination was assessed using receiver operating characteristic (ROC) curves and AUROC as the primary performance metric. Confidence intervals for AUROC were estimated using bootstrap resampling of the test set. Given the moderate cohort size and limited number of minority-class events, performance estimates may be subject to variability; confidence intervals were therefore reported to reflect uncertainty in model discrimination. Confusion matrices were used to summarise classification performance at the selected operating point.

To interpret model predictions and identify key drivers of recovery outcomes, SHAP (SHapley Additive exPlanations) values were computed for the XGBoost models using TreeSHAP and evaluated on the held-out test set. SHAP values were summarised using beeswarm plots to illustrate the relative importance and direction of predictor contributions.

2.8 Cross-centre validation

Cross-centre validation was performed to assess model generalisability to unseen clinical sites. The SERPENS cohort was collected across 34 European centres with unequal patient contributions per site. For each prediction horizon, a subset of centres was designated as the held-out test set, with their combined patient count approximating 20% of the total cohort; no patients from these centres were included in model retraining. This process was repeated five times per prediction horizon using different centre subsets, with repetitions not required to be mutually exclusive. For each fold, XGBoost models were retrained from scratch on the remaining centres using stratified cross-validation for hyperparameter optimisation, and discrimination was evaluated on the held-out centre patients. Discrimination performance was summarised as the mean AUROC across the five folds.

Results

3.1 Cohort characteristics

Of 354 patients with microbiologically confirmed urosepsis enrolled in the SERPENS cohort, 349 were included in the short-term (Day 3) analysis and 233 in the long-term (Day 30) analysis (Figure 3).

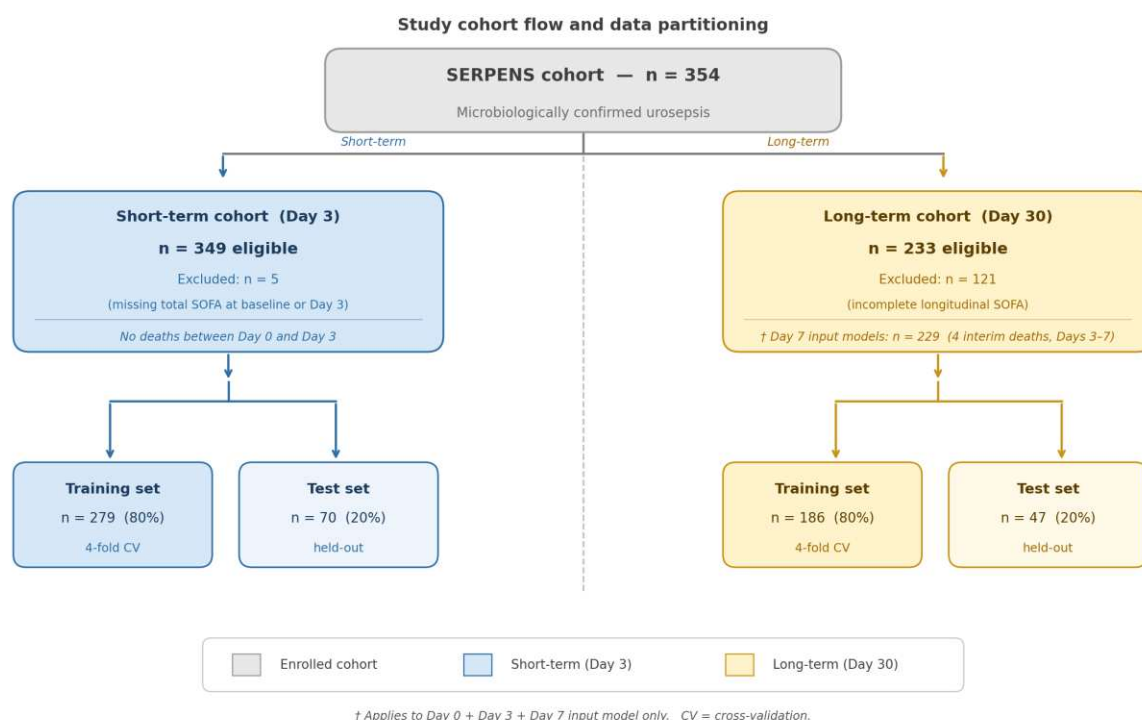


Figure 3. Study cohort flow and data partitioning for short-term and long-term prediction models. Flow diagram of cohort selection and data partitioning for short-term (Day 3; blue) and long-term (Day 30; yellow) analyses. The full SERPENS cohort ($n = 354$) is shown at the top and splits into short-term (left) and long-term (right) cohorts. For the short-term analysis, 349 patients were eligible after exclusion of those with missing baseline or Day 3 SOFA data. For the long-term analysis, 233 patients were eligible after exclusion of those with incomplete longitudinal SOFA data; patients who died between Days 3 and 7 were additionally excluded from models incorporating Day 7 inputs ($n = 4$). Each cohort is further divided into training (80%) and held-out test (20%) sets, as indicated by the lower boxes.

Baseline characteristics of the short-term cohort are summarised in Table 1, stratified by recovery status. Missing and unknown categories are reported in Table S1 in the supplementary. For descriptive analyses and presentation of baseline characteristics, organ-specific SOFA sub-scores were additionally dichotomised (0 vs ≥ 1) to reflect the presence or absence of organ dysfunction and to enhance clinical interpretability. Corresponding numerical representations are provided in Table S2. To improve readability in the main manuscript, only seven representative CCI variables are presented in Table 1, while the complete set of 18 CCI comorbidity variables for the short-term cohort is provided in Table S1. The characteristics for the long-term cohort are provided in Table S3.

Table 1. Baseline characteristics of the short-term cohort stratified by Day 3 recovery outcome. Demographic, clinical, microbiological, and treatment characteristics are shown for the overall population and by recovery status (full recovery versus failure to recover). Continuous variables are presented as median (IQR) and categorical variables as n (%). Organ-specific SOFA sub-scores are dichotomised (0 versus ≥ 1) for descriptive purposes only; numerical values used in modelling are provided in Supplementary Table S2. Seven representative Charlson Comorbidity Index variables are shown; the complete set of 18 variables is reported in Supplementary Table S1. P-values were calculated using Mann–Whitney U tests for continuous variables and chi-squared tests for categorical variables. Abbreviations: CCI, Charlson Comorbidity Index; MDR, multidrug resistance; SOFA, Sequential Organ Failure Assessment.

Variable	Total (n = 349)	Full Recovery (n = 138)	Failure to Recover (n = 211)	p-value
Numerical, Median (IQR)				
Age at diagnosis (years)	65.08 (51.02–73.97)	58.80 (43.35–70.32)	66.65 (54.25–75.94)	<0.001
Weight (kg)	75.00 (65.00–85.00)	75.00 (62.00–83.00)	75.00 (67.00–85.00)	0.294
Height (m)	1.69 (1.62–1.75)	1.68 (1.60–1.74)	1.70 (1.64–1.76)	0.020
BMI (kg/m²)	25.52 (23.45–28.81)	25.54 (23.34–28.70)	25.51 (23.65–29.10)	0.761
SOFA total day 0	2.00 (1.00–3.00)	1.00 (0.00–2.00)	3.00 (2.00–5.00)	<0.001
Categorical, N (%)				
Individual SOFA component scores on day 0 (0 vs ≥1)				
Respiration				
SOFA = 0	259 (74.2%)	125 (90.6%)	134 (63.5%)	<0.001
SOFA ≥ 1	84 (24.1%)	9 (6.5%)	75 (35.5%)	
Coagulation				
SOFA = 0	248 (71.1%)	115 (83.3%)	133 (63.0%)	<0.001
SOFA ≥ 1	101 (28.9%)	23 (16.7%)	78 (37.0%)	
Liver				
SOFA = 0	296 (84.8%)	122 (88.4%)	174 (82.5%)	0.038
SOFA ≥ 1	51 (14.6%)	14 (10.1%)	37 (17.5%)	
Cardiovascular				
SOFA = 0	282 (80.8%)	127 (92.0%)	155 (73.5%)	<0.001
SOFA ≥ 1	67 (19.2%)	11 (8.0%)	56 (26.5%)	
Central nervous system (CNS)				
SOFA = 0	242 (69.3%)	122 (88.4%)	120 (56.9%)	<0.001
SOFA ≥ 1	102 (29.2%)	13 (9.4%)	89 (42.2%)	
Renal				
SOFA = 0	152 (43.6%)	97 (70.3%)	55 (26.1%)	<0.001
SOFA ≥ 1	197 (56.4%)	41 (29.7%)	156 (73.9%)	
Sex				
Female	157 (45.0%)	73 (52.9%)	84 (39.8%)	0.022
Male	192 (55.0%)	65 (47.1%)	127 (60.2%)	
ORENUC				
ORENUC: No known risk				
Yes	29 (8.3%)	17 (12.3%)	12 (5.7%)	0.046
No	320 (91.7%)	121 (87.7%)	199 (94.3%)	
ORENUC: recurrent UTI risk				
Yes	67 (19.2%)	24 (17.4%)	43 (20.4%)	0.580
No	282 (80.8%)	114 (82.6%)	168 (79.6%)	
ORENUC: extra-urogenital risk				
Yes	87 (24.9%)	30 (21.7%)	57 (27.0%)	0.324
No	262 (75.1%)	108 (78.3%)	154 (73.0%)	
ORENUC: nephropathic risk				
Yes	27 (7.7%)	4 (2.9%)	23 (10.9%)	0.011
No	322 (92.3%)	134 (97.1%)	188 (89.1%)	
ORENUC: urological resolvable risk				
Yes	244 (69.9%)	89 (64.5%)	155 (73.5%)	0.096
No	105 (30.1%)	49 (35.5%)	56 (26.5%)	
ORENUC: permanent catheter				
Yes	68 (19.5%)	15 (10.9%)	53 (25.1%)	0.002
No	281 (80.5%)	123 (89.1%)	158 (74.9%)	
Part of Charlson Comorbidity Index (CCI)				
Moderate/severe renal disease				
Yes	34 (9.7%)	5 (3.6%)	29 (13.7%)	0.003
No	315 (90.3%)	133 (96.4%)	182 (86.3%)	
Any tumour				
Yes	65 (18.6%)	11 (8.0%)	54 (25.6%)	<0.001
No	284 (81.4%)	127 (92.0%)	157 (74.4%)	
Metastatic tumour				
Yes	14 (4.0%)	3 (2.2%)	11 (5.2%)	0.256

No	335 (96.0%)	135 (97.8%)	200 (94.8%)	
Peripheral vascular disease				
Yes	38 (10.9%)	16 (11.6%)	22 (10.4%)	0.868
No	311 (89.1%)	122 (88.4%)	189 (89.6%)	
Chronic pulmonary disease				
Yes	34 (9.7%)	10 (7.2%)	24 (11.4%)	0.277
No	315 (90.3%)	128 (92.8%)	187 (88.6%)	
Congestive heart failure				
Yes	24 (6.9%)	6 (4.3%)	18 (8.5%)	0.196
No	325 (93.1%)	132 (95.7%)	193 (91.5%)	
Diabetes				
Yes	85 (24.4%)	31 (22.5%)	54 (25.6%)	0.590
No	264 (75.6%)	107 (77.5%)	157 (74.4%)	
Urine culture				
E. coli				
Yes	178 (51.0%)	84 (60.9%)	94 (44.5%)	0.007
No	155 (44.4%)	47 (34.1%)	108 (51.2%)	
Unknown	16 (4.6%)	7 (5.1%)	9 (4.3%)	
Klebsiella sp.				
Yes	55 (15.8%)	17 (12.3%)	38 (18.0%)	0.353
No	278 (79.7%)	114 (82.6%)	164 (77.7%)	
Enterococcus sp.				
Yes	44 (12.6%)	14 (10.1%)	30 (14.2%)	0.516
No	289 (82.8%)	117 (84.8%)	172 (81.5%)	
Proteus sp.				
Yes	22 (6.3%)	7 (5.1%)	15 (7.1%)	0.711
No	311 (89.1%)	124 (89.9%)	187 (88.6%)	
Enterobacter sp.				
Yes	25 (7.2%)	7 (5.1%)	18 (8.5%)	0.454
No	308 (88.3%)	124 (89.9%)	184 (87.2%)	
Pseudomonas aeruginosa				
Yes	13 (3.7%)	3 (2.2%)	10 (4.7%)	0.445
No	320 (91.7%)	128 (92.8%)	192 (91.0%)	
Other pathogen				
Yes	36 (10.3%)	12 (8.7%)	24 (11.4%)	0.693
No	297 (85.1%)	119 (86.2%)	178 (84.4%)	
Antibiotic susceptibility status				
MDR	88 (25.2%)	22 (15.9%)	66 (31.3%)	0.0027
Not MDR	97 (27.8%)	36 (26.1%)	61 (28.9%)	
Pan susceptible	120 (34.4%)	60 (43.5%)	60 (28.4%)	
Unknown	44 (12.6%)	20 (14.5%)	24 (11.4%)	
Antibiotic spectrum at day 0				
Broad only	219 (62.8%)	83 (60.1%)	136 (64.5%)	0.685
Both	58 (16.6%)	24 (17.4%)	34 (16.1%)	
Narrow only	71 (20.3%)	31 (22.5%)	40 (19.9%)	

In the short-term cohort, 211 patients (60.5%) did not achieve full recovery by Day 3, consistent with the sustained organ dysfunction observed in the early post-diagnosis period in the original SERPENS cohort (8). The long-term distribution was more balanced, with 134 patients (57.5%) recovering and 99 (42.5%) not recovering by Day 30. The higher non-recovery rate compared with SERPENS (26.9%) reflects both the more stringent outcome definition applied here, requiring complete resolution of organ dysfunction (SOFA = 0) rather than absence of overt organ failure, and restriction to the 233 patients with complete longitudinal SOFA data.

At baseline, organ dysfunction was common across multiple systems, with respiratory, cardiovascular, central nervous system, and renal involvement frequently observed. Across both timepoints, a higher prevalence of organ dysfunction was consistently observed among patients who did not achieve full recovery. Notably, central nervous system and renal dysfunction demonstrated the most pronounced and consistent differences between outcome groups at both Day 3 and Day 30 and anticipated their later dominance in model-based analyses. Older age, comorbid malignancy, permanent urinary catheterisation, and nephropathic urological risk were also more prevalent among non-recovering patients.

The candidate predictors included in model development for each prediction horizon are summarised in Table 2, reflecting differences in variable availability and association across short- and long-term outcomes.

Table 2. Candidate predictors retained after feature selection for short-term (Day 3) and long-term (Day 30) recovery prediction models. Variables were selected using univariate screening ($p \leq 0.05$) followed by assessment of redundancy and collinearity. A tick (✓) indicates that a variable was retained as a candidate predictor for the corresponding model. Total SOFA score and organ-specific sub-scores were treated as mutually exclusive inputs and are shown separately to indicate their selection status. Variables excluded due to redundancy or collinearity are not shown; the complete list of evaluated variables is provided in the Supplementary Material. Abbreviations: CCI, Charlson Comorbidity Index; CNS, central nervous system; ORENUC, urological risk classification; SOFA, Sequential Organ Failure Assessment.

Predictor	Short-term (Day 3)	Long-term (Day 30)
Demographics		
Patient age	✓	✓
Sex	✓	
SOFA scores at Day 0		
SOFA total at 0	✓	✓
SOFA respiration at day 0	✓	✓
SOFA coagulation at day 0	✓	✓
SOFA liver at day 0	✓	✓
SOFA cardiovascular at day 0	✓	✓
SOFA CNS at day 0	✓	✓
SOFA renal at day 0	✓	✓
Urological risk (ORENUC)		

ORENUC: no known risk	✓	
ORENUC: nephropathic risk	✓	✓
ORENUC: permanent catheter	✓	
Comorbidities (CCI)		
CCI: any tumour	✓	✓
CCI: metastatic tumour		✓
Microbiology		
E. Coli presence in urine culture	✓	✓
Antibiotic susceptibility status	✓	✓

3.2 Primary analysis: Total SOFA models

3.2.1 Model performance at Day 3 and Day 30

Using baseline (Day 0) total SOFA scores, both models demonstrated good discrimination for recovery prediction at short- and long-term horizons (Table 3).

Table 3. Model discrimination for short-term and long-term recovery prediction using total SOFA score representation. Discriminative performance of logistic regression and XGBoost models for short-term (Day 3) and long-term (Day 30) recovery prediction using total SOFA score inputs. Values are reported as AUROC with 95% confidence intervals. Best-performing models at each prediction horizon are shown in bold.

Short-term 0 → 3	0.87 [0.78-0.94]	0.87[0.76-0.96]
Long-term 0 → 30	0.76 [0.61-0.89]	0.79 [0.65-0.91]

XGBoost consistently achieved the highest AUROC across both prediction tasks, although performance differences between models were modest. Discrimination was substantially stronger for short-term recovery prediction (Day 3) than for long-term recovery prediction (Day 30), indicating that baseline illness severity was more predictive of early clinical trajectory than longer-term recovery outcome.

ROC curves for all models are shown in Figure 4. Confidence bands derived from bootstrap resampling illustrate the uncertainty in discrimination estimates across classification thresholds. Corresponding confusion matrices at selected probability thresholds are presented in Supplementary Figure S1.

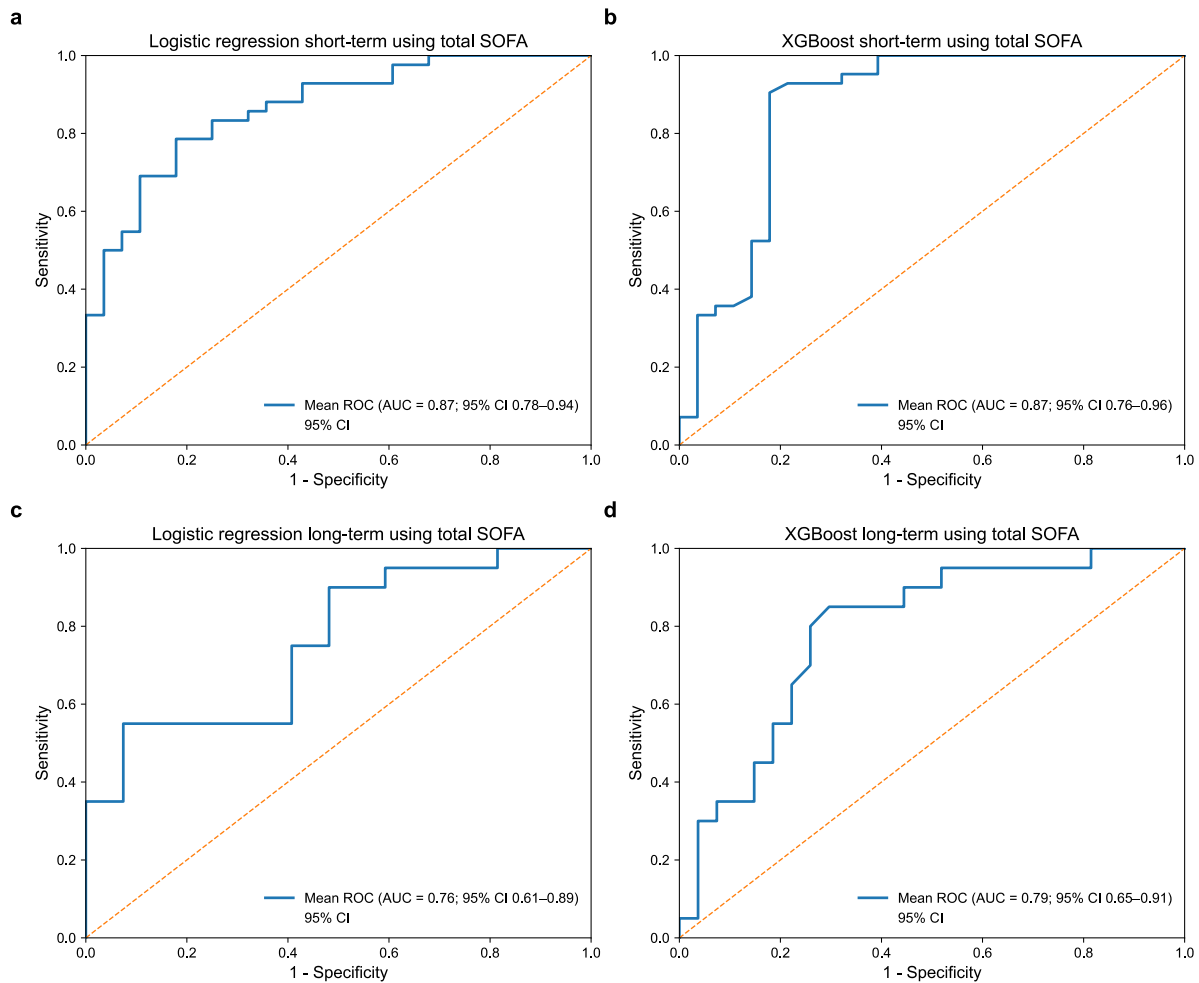


Figure 4. Receiver operating characteristic (ROC) curves for short-term and long-term recovery prediction using total SOFA score. Receiver operating characteristic (ROC) curves for logistic regression (a, c) and XGBoost (b, d) models. Panels a and b show short-term (Day 3) prediction, and panels c and d show long-term (Day 30) prediction. Solid lines indicate mean ROC curves. The diagonal dashed line indicates chance-level discrimination. AUROC values with corresponding confidence intervals are shown within each panel.

3.2.2 Impact of longitudinal data on model performance

Incorporating longitudinal SOFA measurements improved model discrimination for long-term (Day 30) recovery prediction across both modelling approaches (Figure 5).

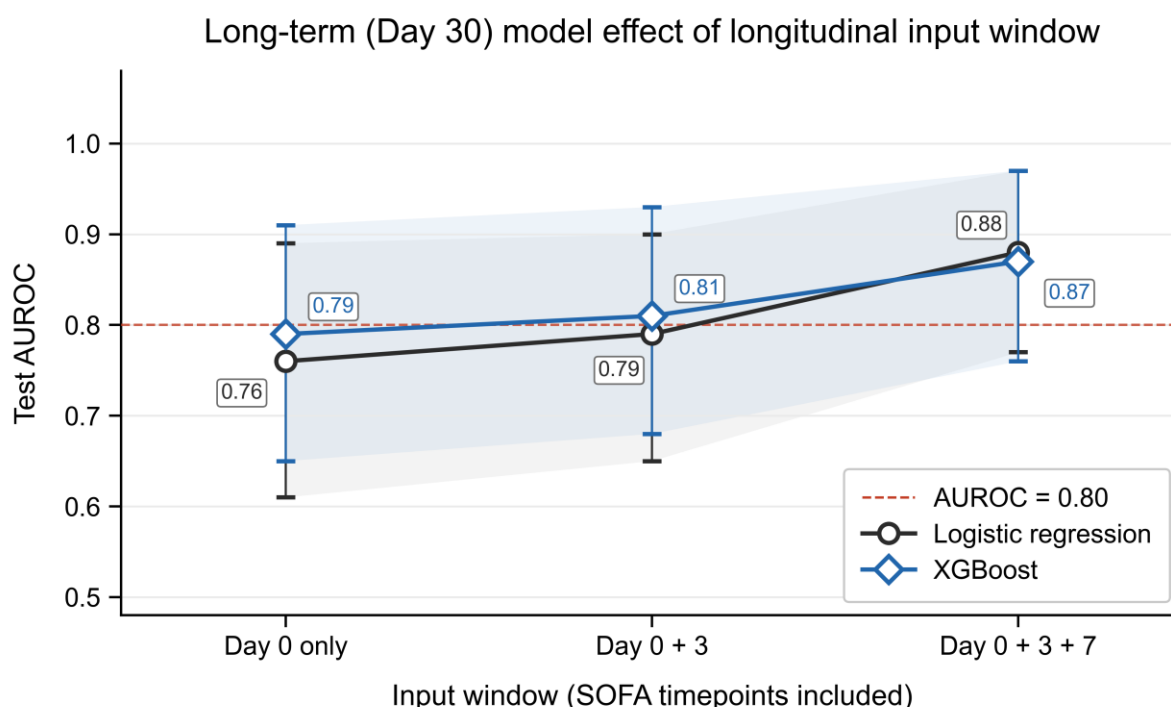


Figure 5. Effect of longitudinal input window on long-term (Day 30) recovery prediction. Test set AUROC for logistic regression (black circles) and XGBoost (blue diamonds) models across three input windows: Day 0 only, Day 0 + Day 3, and Day 0 + Day 3 + Day 7. All models use total SOFA score inputs. Points indicate mean AUROC, with error bars and shaded regions representing 95% confidence intervals from bootstrap resampling of the held-out test set. The red dashed line indicates a reference threshold of AUROC = 0.80. AUROC, area under the receiver operating characteristic curve; SOFA, Sequential Organ Failure Assessment.

Using baseline data alone (Day 0), performance was modest, with AUROC of 0.76 for logistic regression and 0.79 for XGBoost. Inclusion of early follow-up data (Day 0 + Day 3) resulted in improved discrimination for both models, with further gains observed when full longitudinal information (Day 0 + Day 3 + Day 7) was incorporated. At this stage, AUROC reached 0.88 for logistic regression and 0.87 for XGBoost.

Improvements were more pronounced for logistic regression, resulting in convergence of model performance as additional timepoints were included. The incorporation of longitudinal SOFA trajectories captures evolving disease dynamics, reducing reliance

on model complexity and highlighting the importance of temporal information in recovery prediction.

3.2.3 Model interpretation (SHAP analysis)

To characterise the contribution of individual predictors to model output, SHAP (SHapley Additive exPlanations) values were computed for the XGBoost models at both prediction horizons (Figure 6). SHAP analysis was performed for the XGBoost models, as these represent the highest-performing and non-linear modelling approach. SHAP values quantify the contribution of each feature to an individual prediction, with positive values indicating increased probability of failure to achieve full recovery and negative values indicating decreased probability.

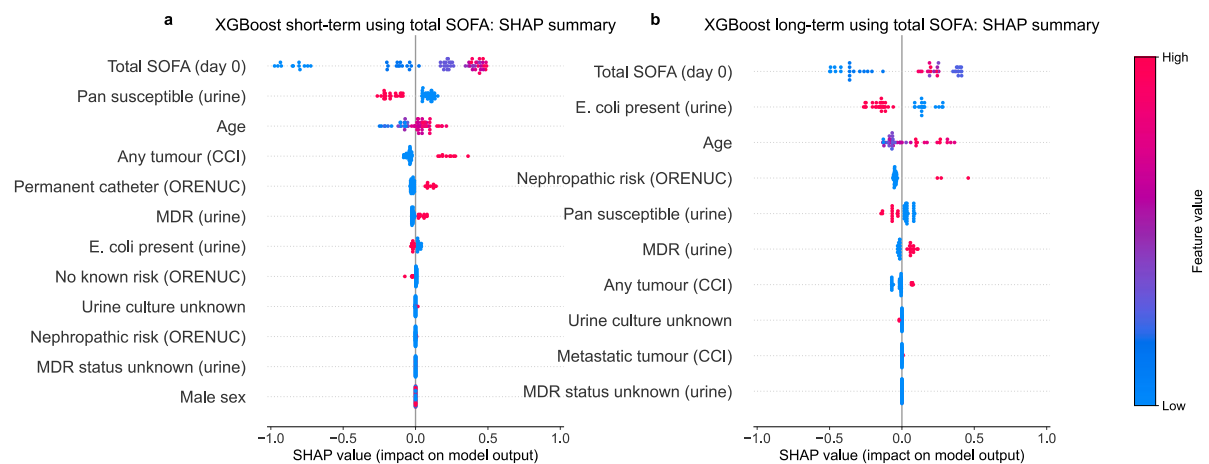


Figure 6. SHAP summary plots for XGBoost models predicting short-term and long-term recovery. SHAP (SHapley Additive exPlanations) summary plots for XGBoost models predicting short-term (Day 3; a) and long-term (Day 30; b) outcomes. Each point represents an individual patient. The x-axis shows the SHAP value, indicating the direction and magnitude of each feature's contribution to the model prediction, with positive values corresponding to an increased likelihood of failure to recover. Points are coloured by feature value (red, high; blue, low). Features are ordered by mean absolute SHAP value.

Across both prediction horizons, total SOFA score at baseline emerged as the dominant contributor to model predictions, with higher scores consistently associated with increased likelihood of failure to achieve full recovery.

In the short-term model (Day 3), pan-susceptible susceptibility status represented the second strongest contributor after total SOFA score and was associated with favourable recovery, whereas MDR organisms contributed towards adverse outcomes. Increasing age, comorbid malignancy, and permanent urinary catheterisation were also among the contributors associated with poorer recovery. *Escherichia coli* in urine and absence of known ORENUC risk factors showed smaller protective associations. Nephropathic ORENUC risk, unknown urine culture status, unknown MDR status, and sex contributed minimally to prediction.

In the long-term model (Day 30), *Escherichia coli* in urine was the strongest microbiological contributor after total SOFA score and was associated with favourable recovery. Increasing age and nephropathic ORENUC risk were also prominent contributors associated with adverse outcomes. Pan-susceptible organisms favoured recovery, whereas MDR status and comorbid malignancy contributed towards poorer outcomes, although with smaller overall magnitude than in the short-term model. Unknown urine culture status, unknown MDR status, and metastatic malignancy contributed minimally to model output.

3.3 Secondary analysis: Organ-specific SOFA models

3.3.1 Model performance compared with total SOFA

While total SOFA captures overall illness severity, it does not preserve information about the distribution of dysfunction across organ systems, which may be critical for determining recovery trajectories. Building on the baseline performance achieved using total SOFA, we next evaluated whether disaggregating organ dysfunction into

its constituent components reveals prognostic structure not captured by the composite SOFA score (Table 4).

Table 4. Predictive performance using total versus organ-specific SOFA representations. AUROC values for logistic regression and XGBoost models using either total SOFA score or organ-specific SOFA sub-scores as predictors for short-term (Day 3) and long-term (Day 30) outcomes. Values are presented as mean AUROC with 95% confidence intervals derived from cross-validation. Best-performing models within each prediction horizon are shown in bold.

SOFA combination	Logistic Regression	XG Boost
Total SOFA	0.87 [0.78, 0.94]	0.87 [0.76, 0.96]
All 6 organs	0.89 [0.81, 0.96]	0.90 [0.83, 0.96]
SOFA combination	Logistic Regression	XG Boost
Total SOFA	0.76 [0.61, 0.89]	0.79 [0.65, 0.91]
All 6 organs	0.81 [0.67, 0.92]	0.83 [0.71, 0.94]

Across both prediction horizons and modelling approaches, models incorporating organ-specific SOFA sub-scores demonstrated consistently higher discrimination than those based on total SOFA alone. In the short-term setting (Day 3), AUROC increased from 0.87 to 0.89 for logistic regression and from 0.87 to 0.90 for XGBoost. In the long-term setting (Day 30), AUROC improved from 0.76 to 0.81 for logistic regression and from 0.79 to 0.83 for XGBoost.

These findings show that aggregating organ dysfunction into a single composite score obscures clinically relevant prognostic structure, whereas organ-level representations reveal heterogeneity in recovery trajectories. Notably, the greater relative improvement observed for logistic regression compared with XGBoost suggests that organ-level representations may partially compensate for the limited capacity of linear models to capture complex interactions.

3.3.2 Organ-level importance (SHAP analysis)

To delineate which organ systems underpinned the observed performance gains with organ-specific SOFA representation, SHAP values were examined for XGBoost models incorporating individual SOFA sub-scores (Figure 7).

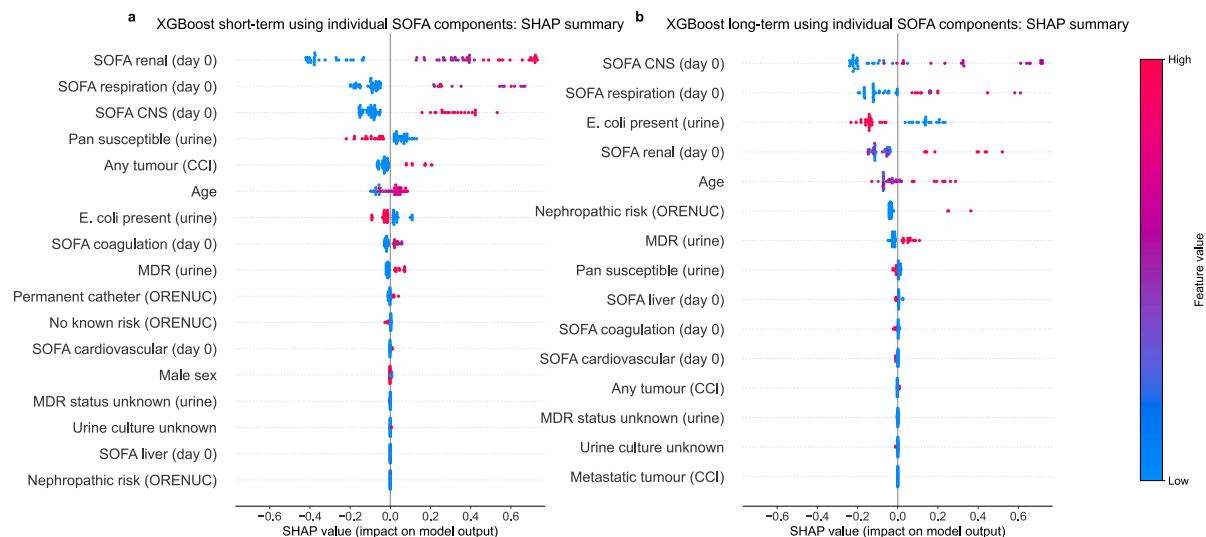


Figure 7. SHAP summary plots for XGBoost short-term and long-term models using organ-specific SOFA sub-scores. SHAP (SHapley Additive exPlanations) summary plots for XGBoost models trained on organ-specific SOFA sub-scores for short-term (Day 3; a) and long-term (Day 30; b) prediction. Each point represents an individual patient. The x-axis shows the SHAP value, indicating the direction and magnitude of each feature's contribution to the model prediction, with positive values corresponding to an increased likelihood of failure to recover. Points are coloured by feature value (red, high; blue, low). Features are ordered by mean absolute SHAP value.

In the short-term model (Day 3), renal, respiratory, and central nervous system (CNS) dysfunction were the dominant contributors, with higher SOFA scores consistently associated with adverse outcomes. Similarly to the total sofa models, pan-susceptible status was the strongest microbiological contributor and was associated with favourable recovery. Comorbid malignancy, increasing age, permanent catheterisation, and MDR status were associated with failure to achieve full recovery. The presence of *Escherichia coli* in urine and absence of known urological risk factors demonstrated smaller protective associations. The coagulation SOFA component contributed weakly towards adverse outcomes, whereas liver and cardiovascular

SOFA scores, nephropathic risk, unknown MDR status, sex at birth, and absence of urine culture data contributed minimally to prediction.

In the long-term model (Day 30), CNS and respiratory dysfunction were the dominant contributors. The presence of *Escherichia coli* in urine emerged as the strongest microbiological predictor of favourable recovery, followed by renal dysfunction, increasing age, and nephropathic risk, which were associated with adverse outcomes. MDR status contributed towards poorer recovery, whereas pan-susceptible organisms were associated with favourable outcomes, both with moderate overall contribution. Liver, coagulation, and cardiovascular SOFA components showed weak contributions, while comorbid malignancy, metastatic malignancy, unknown MDR status, and unknown urine culture status contributed minimally to model output.

3.3.3 Individual-level interpretation of model predictions

To illustrate how model predictions arise from clinically meaningful features, we examined correctly classified test set cases with high prediction confidence in the short-term and long-term models using total SOFA scores. Figure 8 represent examples of how individual variables contribute to predictions at the patient level in waterfall plots for short-term and long-term models.

XGBoost using total SOFA: case studies

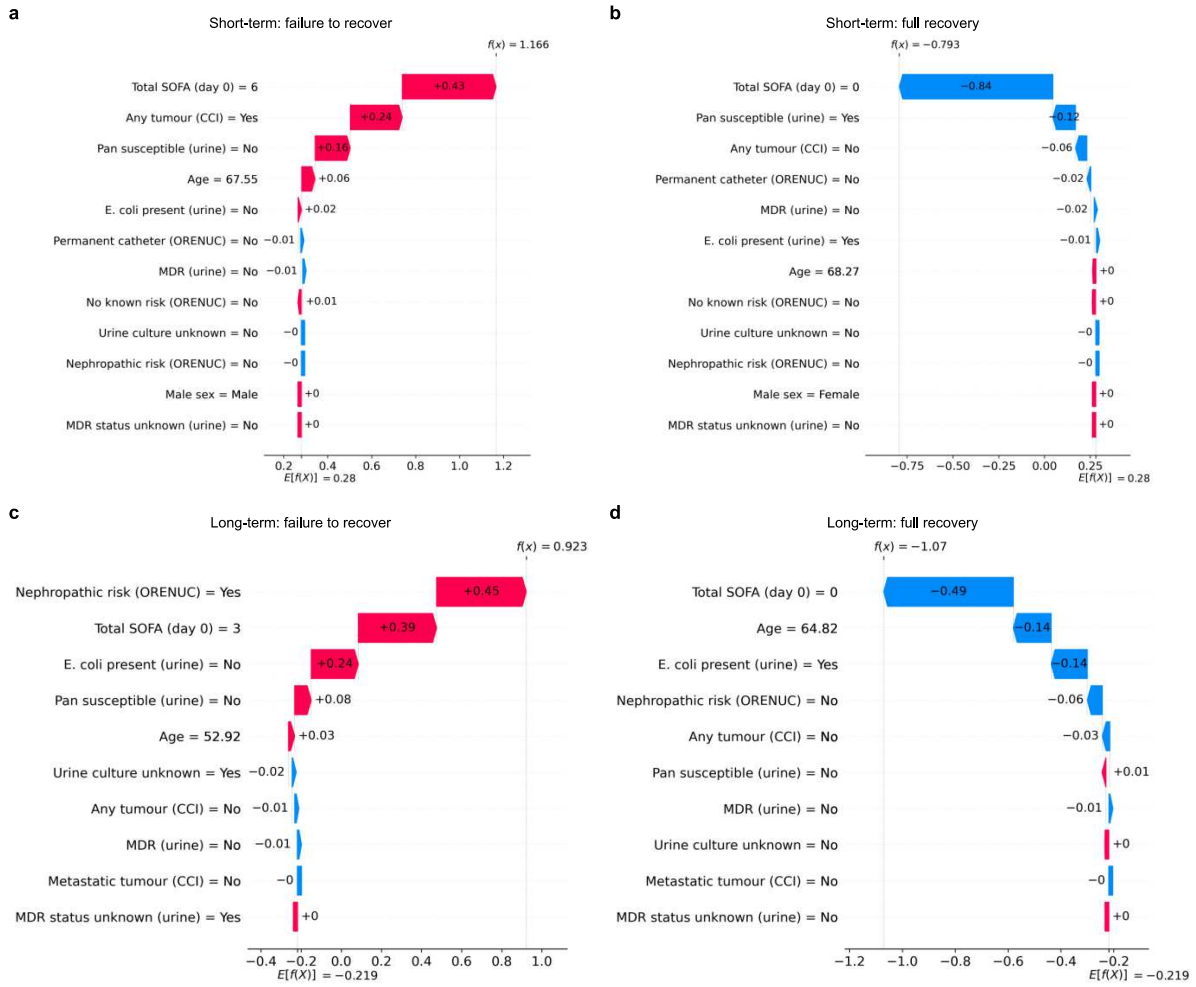


Figure 8. Individual-level feature contributions to model predictions for representative test cases. Waterfall plots showing feature contributions for representative correctly classified test set cases. Panels a and b show short-term (Day 3) predictions, and panels c and d show long-term (Day 30) predictions; a and c correspond to failure to recover, and b and d to full recovery. Bars indicate the contribution of each feature to the final prediction ($f(x)$) relative to the baseline ($E[f(x)]$). Positive values (red) increase the probability of failure to recover, and negative values (blue) decrease it.

In the short-term model (Figure 8a,b), predictions were driven primarily by baseline illness severity and microbiological susceptibility. A patient classified as Failure to Recover (Figure 8a) demonstrated strong adverse contributions from elevated baseline SOFA score, comorbid malignancy, absence of pan-susceptibility, increasing age, and absence of *Escherichia coli* in urine. In contrast, a patient classified as Full Recovery (Figure 8b) showed a strong favourable contribution from a baseline SOFA score of 0, together with pan-susceptible pathogen status, absence of malignancy,

absence of permanent catheterisation, absence of MDR resistance, and the presence of *Escherichia coli* in urine.

In the long-term model (Figure 8c,d), predictions reflected a broader combination of host vulnerability and microbiological features. A patient classified as Failure to Recover (Figure 8c) demonstrated adverse contributions from nephropathic risk, elevated baseline SOFA score, absence of *Escherichia coli* in urine, and absence of pan-susceptibility. Conversely, a patient classified as Full Recovery (Figure 8d) showed favourable contributions from low baseline SOFA score, the presence of *Escherichia coli* in urine, absence of nephropathic risk, and absence of malignancy.

To further examine how model predictions are constructed, we applied the same case-based decomposition to models incorporating organ-specific SOFA sub-scores (Figure 9) to evaluate the contributions of individual organ systems.

XGBoost using individual SOFA: case studies

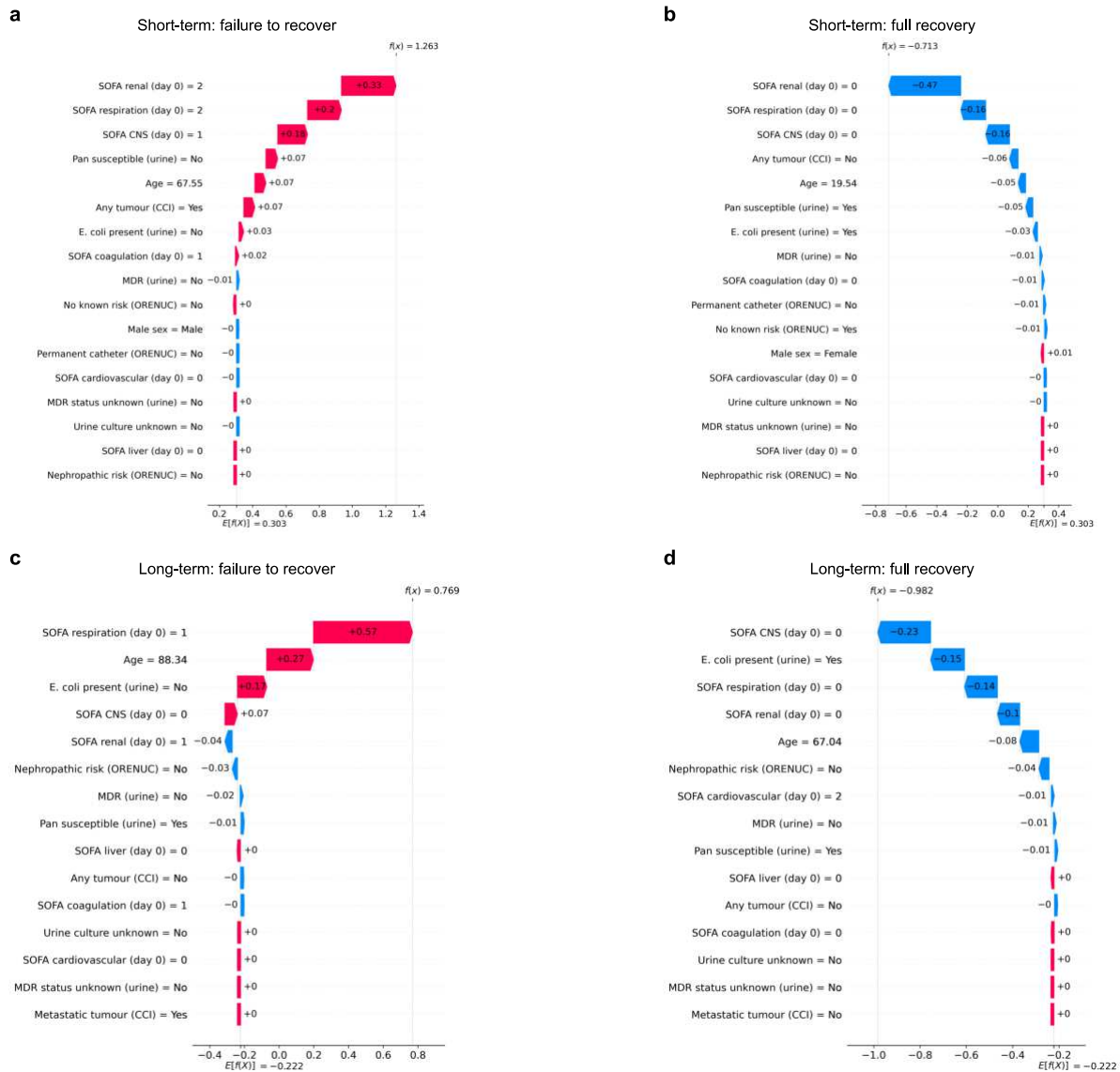


Figure 9. Individual-level feature contributions to model predictions using organ-specific SOFA sub-scores. Waterfall plots showing feature contributions for representative correctly classified test set cases using models incorporating organ-specific SOFA sub-scores. Panels a and b show short-term (Day 3) predictions, and panels c and d show long-term (Day 30) predictions; a and c correspond to failure to recover, and b and d to full recovery. Bars indicate the contribution of each feature to the final prediction ($f(x)$) relative to the baseline ($E[f(x)]$). Positive values (red) increase the probability of failure to recover, and negative values (blue) decrease it.

In the short-term model (Figure 9a,b), predictions were primarily driven by dysfunction within individual organ systems. A patient classified as Failure to Recover (Figure 9a) demonstrated strong adverse contributions from renal, respiratory, and central nervous system (CNS) dysfunction, together with absence of pan-susceptibility, increasing age, comorbid malignancy, and absence of *Escherichia coli* in urine. In

contrast, a patient classified as Full Recovery (Figure 9b) showed favourable contributions from low renal, respiratory, and CNS SOFA scores, alongside absence of malignancy, younger age, pan-susceptible pathogen status, and the presence of *Escherichia coli* in urine.

In the long-term model (Figure 9c,d), predictions reflected a broader combination of organ dysfunction and host-related risk factors. A patient classified as Failure to Recover (Figure 9c) demonstrated adverse contributions from respiratory dysfunction, advanced age, absence of *Escherichia coli* in urine, CNS dysfunction, renal dysfunction, and nephropathic risk. Conversely, a patient classified as Full Recovery (Figure 9d) showed favourable contributions from low CNS, respiratory, and renal dysfunction, presence of *Escherichia coli* in urine, and absence of nephropathic risk.

Across these examples, decomposition into organ-specific SOFA components demonstrated how dysfunction within renal, respiratory, and CNS domains contributed disproportionately to model predictions while preserving the broader recovery patterns observed using total SOFA scores.

3.4 Cross-centre validation

To assess model discrimination on completely unseen clinical sites, cross-centre validation was performed by designating a subset of centres as a held-out test set across five folds; results are shown in Figure 10.

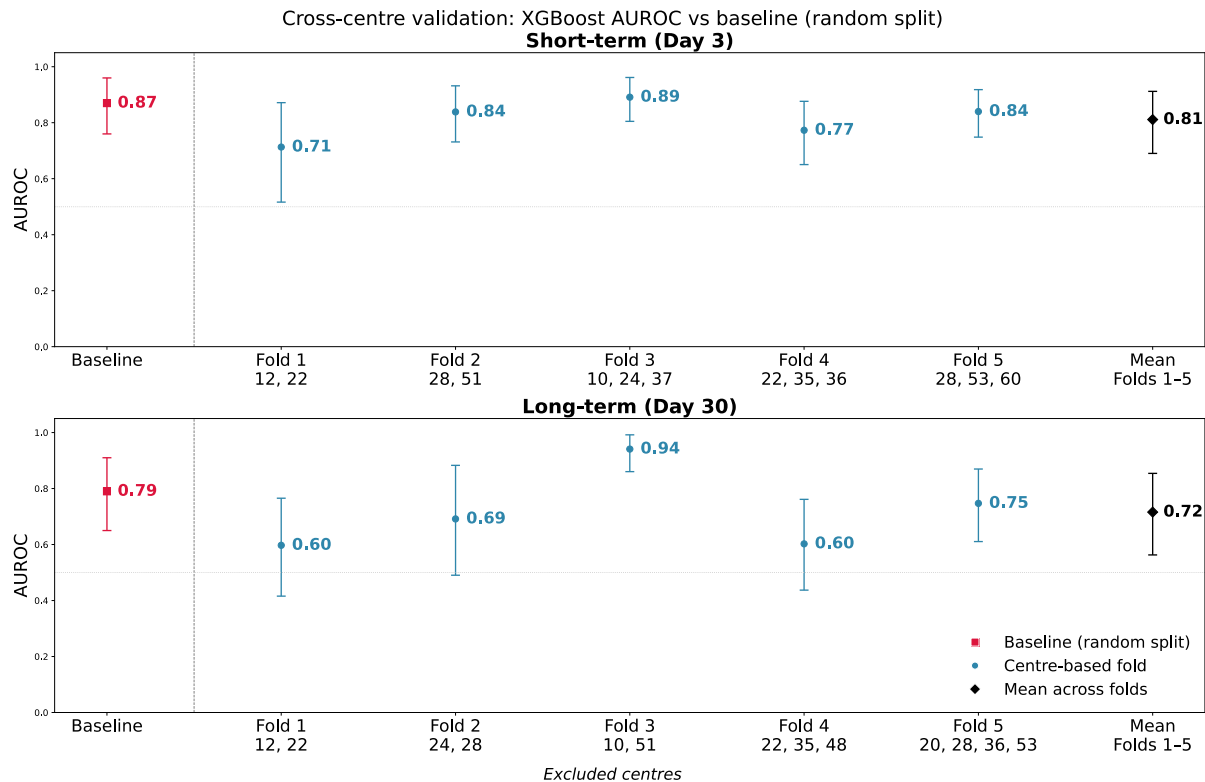


Figure 10. Cross-centre validation performance for short-term (Day 3) and long-term (Day 30) recovery prediction. XGBoost AUROC for the random split baseline (red square), each of five centre-based held-out folds (blue circles), and the mean AUROC across folds (black diamond) for short-term (left) and long-term (right) prediction horizons. Excluded centres for each fold are indicated on the x-axis. Error bars represent 95% confidence intervals derived from bootstrap resampling. AUROC, area under the receiver operating characteristic curve.

Cross-centre validation demonstrated broadly maintained discrimination across both prediction horizons. For short-term recovery prediction, mean AUROC dropped from a baseline of 0.87 to 0.81 across five folds, a reduction of 7% relative to baseline. For long-term recovery prediction, mean AUROC dropped from 0.79 to 0.72, a reduction of 9% relative to baseline. Fold-level AUROC ranged from 0.71 to 0.89 for short-term and from 0.60 to 0.94 for long-term prediction, with greater variability observed across folds for the long-term horizon.

Discussion

Recovery following urosepsis is not adequately captured by a single severity score at presentation but instead reflects a dynamic, organ-specific process that evolves over

time. Across both prediction horizons, decomposing total SOFA into its constituent organ sub-scores consistently improved discrimination, with renal, CNS, and respiratory dysfunction emerging as the dominant predictors, with clear shifts in their relative importance between Day 3 and Day 30. Early outcomes were driven by acute physiological derangement, whereas longer-term outcomes reflected residual dysfunction and underlying vulnerability. These findings support a trajectory-based model of prognosis that more closely reflects clinical recovery patterns.

Current urosepsis management prioritises acute intervention but provides limited guidance for longer-term risk stratification (15). While prompt antimicrobial therapy, source control, and organ support remain central to care (1), they do not distinguish patients who will recover from those with persistent organ dysfunction. This distinction is clinically important, as illustrated by the SERPENS cohort, in which mortality was low, but organ dysfunction affected most patients and persisted in a substantial proportion at 30 days (8). Prior analysis identified baseline respiratory failure, comorbidity burden, and multidrug-resistant pathogens as predictors of 30-day outcomes (6), but did not account for time-dependent organ-level contributions. The present findings extend this framework by demonstrating how organ-specific dysfunction refines prognostic assessment beyond the acute phase.

Total SOFA at presentation retains prognostic value but can also be extended beyond its original role as a descriptive severity score. Although SOFA was developed to characterise organ dysfunction rather than direct management decisions (1), serial assessment inherently reflects evolving physiological response and treatment trajectory. In its original validation, SOFA achieved AUROC values of 0.74 for in-

hospital mortality prediction in ICU patients and 0.79 in ward settings (1) — benchmarks against which recovery-focused discrimination in the present analysis compares favourably, while targeting a clinically distinct endpoint. The observed improvement in long-term prediction with incorporation of serial SOFA measurements suggests that longitudinal organ dysfunction profiling, particularly across Days 0, 3, and 7, may provide a useful framework for dynamic clinical phenotyping in urosepsis. Combined with baseline host-risk stratification (ORENUC classification) and microbiological features, this approach may support more individualised monitoring of recovery trajectory at the point of care.

Prognostic performance is driven more by the composition of SOFA than by its total score alone. Replacing total SOFA with organ sub-scores improved model discrimination, consistent with recognised limitations of equal weighting across organ systems in Sepsis-3 (1). The present data demonstrate that prognostic signal in urosepsis is concentrated within a subset of domains, with renal dysfunction dominating early prediction and CNS dysfunction becoming more prominent over longer horizons. This implies that two patients with identical total SOFA scores may have materially different risk profiles depending on the distribution of organ involvement. In clinical terms, this supports a more granular interpretation of SOFA in clinical practice.

The organ-specific patterns observed in this study are consistent with established features of sepsis pathophysiology and clinical course. Renal dysfunction as the dominant early predictor aligns with established data on sepsis-associated acute kidney injury (SA-AKI), which affects 30–51% of sepsis patients and frequently

reaches Stage 2–3 severity in the ICU (16). Critically, SA-AKI follows a tripartite trajectory of early recovery, late recovery, and non-recovery (16), that closely mirrors the divergent outcome pathways captured at Day 3 and Day 30 in the present analysis. CNS dysfunction emerged as the strongest predictor of longer-term outcome, consistent with evidence that sepsis-associated encephalopathy is associated with prolonged cognitive and functional impairment among survivors (9). Respiratory dysfunction contributed predominantly to early risk, in keeping with the often-reversible nature of acute respiratory compromise with appropriate supportive care. These findings reinforce the biological plausibility of time-dependent, organ-specific contributions to recovery.

Short- and long-term outcomes reflect distinct but related phases of recovery shaped by both acute dysfunction and baseline vulnerability. At Day 3, outcomes are largely determined by initial organ dysfunction, pathogen, and baseline patient factors, reflecting the acute host response to infection. This is consistent with frameworks such as ORENUC, which capture host-related risk in urological infection (17). In particular, permanent catheter status was associated with adverse short-term outcomes, likely reflecting the combined effects of chronic urinary tract instrumentation, increased management complexity, and underlying host vulnerability. In contrast, long-term recovery at Day 30 was more strongly influenced by nephropathic risk, suggesting that chronic structural or functional renal vulnerability becomes increasingly important during later phases of recovery. These findings support the clinical relevance of ORENUC phenotyping beyond initial infection stratification and suggest that different components of urological risk may contribute differentially across recovery timepoints.

Microbiological features provided prognostic information beyond host and organ-level factors. *Escherichia coli* was consistently associated with more favourable recovery across both prediction horizons, whereas increasing antimicrobial resistance shifted predictions towards failure to achieve full recovery. This pattern was most apparent in the short-term model, where susceptibility status was among the strongest contributors after baseline organ dysfunction. Clinically, this likely reflects the greater predictability of empirical antimicrobial coverage in pan-susceptible infections, contrasted with the delayed or suboptimal effective therapy often associated with multidrug-resistant organisms. These findings are consistent with prior SERPENS analyses linking multidrug resistance to adverse outcomes (8) , and support incorporating culture and susceptibility data into prognostic reassessment once microbiological results become available.

Cross-centre validation demonstrated that model discrimination was broadly maintained when applied to completely unseen clinical sites, supporting the generalisability of the prognostic framework beyond the centres on which it was trained. For short-term prediction, mean AUROC of 0.81 across five centre-based folds represents a modest reduction from the random-split baseline of 0.87, consistent with the expected performance drop when models encounter institutional variation in case-mix, clinical practice, and data collection. For long-term prediction, greater fold-level variability (AUROC 0.60–0.94) reflects the smaller long-term cohort and unequal distribution of outcome events across centre subsets, meaning that some held-out folds contained very few minority-class cases, producing unstable performance estimates. The mean AUROC of 0.72 for long-term cross-centre prediction should therefore be interpreted cautiously. Nevertheless, the fact that discrimination remained

meaningful across the majority of folds, and that the SERPENS cohort spans 34 geographically diverse European institutions, provides substantially stronger evidence of generalisability than is typically available from single-centre internal validation studies.

Several limitations should be considered when interpreting these findings. The cohort size was moderate, particularly for long-term modelling ($n = 233$), and univariate screening prior to model splitting may introduce modest optimistic bias. However, this approach was adopted to preserve directly comparable feature sets across modelling strategies, as fully nested feature selection would likely have resulted in differing variable subsets between models and reduced interpretability of comparative analyses. This trade-off reflects a common practical challenge in clinical machine learning studies involving moderate-sized datasets and high-dimensional clinical variables. The cohort was enrolled under earlier sepsis definitions, and while the variables and SOFA-based outcome definition are compatible with contemporary Sepsis-3 criteria, differences in patient selection may exist (1).

These findings point towards actionable changes in how information already in front of clinicians might be better used. First, the SOFA sub-score composition at admission should be considered explicitly rather than treated as background detail behind the total and interpreted alongside established clinical risk frameworks such as ORENUC, which capture baseline urological and systemic vulnerability. Second, SOFA trajectory over the first week should be interpreted prognostically alongside guiding treatment: patients who have not cleared their renal dysfunction by Day 3 or their CNS dysfunction by Day 7 may warrant consideration for closer follow-up, discharge

planning, and post-acute review. Third, urine culture results, specifically whether the pathogen is *E. coli* or a pan-susceptible organism, should be incorporated into recovery prognosis alongside organ dysfunction data, supporting a more integrated view of the host–pathogen interaction that governs longer-term outcome. Together, these approaches support a more integrated and longitudinal view of urosepsis prognosis.

Recovery after urosepsis is not a single outcome state but a dynamic, organ-specific trajectory that evolves over time. Prognostic information is not uniformly distributed across organ systems but concentrated within specific domains, and its relevance shifts across early and later phases of recovery. This temporal and organ-level heterogeneity cannot be captured by aggregate severity scores alone. These findings support a shift from static, score-based assessment towards a longitudinal and decomposed interpretation of organ dysfunction, using routinely collected data to inform prognosis beyond survival. More broadly, this framework reframes urosepsis prognosis as a trajectory rather than an event, with potential applicability across infectious syndromes where recovery, rather than mortality, defines outcome.

Data availability

The data used in this study were collected as part of the SERPENS study (NCT02380170). Ethical approval for the study was granted by the Justus Liebig University Ethics Committee, Giessen, Germany (AZ: 77/14). Due to the sensitive nature of patient-level clinical data, the dataset is not publicly available. Requests for access can be directed to the corresponding author, and will be referred to the SERPENS study investigators for consideration.

Code availability

Custom code developed for this analysis is available from the corresponding authors upon reasonable request.

References

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*. 2016 Feb 23;315(8):801–10. doi:10.1001/jama.2016.0287
2. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet*. 2020 Jan 18;395(10219):200–11. doi:10.1016/S0140-6736(19)32989-7 PubMed PMID: 31954465.
3. Porat A, Bhutta BS, Kesler S. Urosepsis. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 [cited 2026 Jan 19]. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK482344/> PubMed PMID: 29493969.
4. Wagenlehner FME, Weidner W, Naber KG. Optimal management of urosepsis from the urological perspective. *Int J Antimicrob Agents*. 2007 Nov 1;30(5):390–7. doi:10.1016/j.ijantimicag.2007.06.027
5. Peach BC, Garvan GJ, Garvan CS, Cimiotti JP. Risk Factors for Urosepsis in Older Adults. *Gerontol Geriatr Med*. 2016 Apr 6;2:2333721416638980. doi:10.1177/2333721416638980 PubMed PMID: 28138493; PubMed Central PMCID: PMC5119864.
6. Skov CS, Østervang C, Brabrand M, Lassen AT, Nielsen DS. How do sepsis survivors experience life after sepsis? A Danish qualitative study exploring factors of importance [Internet]. 2024 Feb 1. doi:10.1136/bmjopen-2023-081558
7. Mostel Z, Perl A, Marck M, Mehdi SF, Lowell B, Bathija S, et al. Post-sepsis syndrome – an evolving entity that afflicts survivors of sepsis. *Mol Med*. 2019 Dec 31;26(1):6. doi:10.1186/s10020-019-0132-z
8. Tandogdu Z, Koves B, Ristovski S, Balci MBC, Rennesund K, Gravas S, et al. Urosepsis 30-day mortality, morbidity, and their risk factors: SERPENS study, a prospective, observational multi-center study. *World J Urol*. 2024 May 10;42(1):314. doi:10.1007/s00345-024-04979-2
9. Meyer NJ, Prescott HC. Sepsis and Septic Shock. *N Engl J Med*. 2024 Dec 4;391(22):2133–46. doi:10.1056/NEJMra2403213
10. Martin-Loeches I, Singer M, Leone M. Sepsis: key insights, future directions, and immediate goals. A review and expert opinion. *Intensive Care Med*. 2024 Dec 1;50(12):2043–9. doi:10.1007/s00134-024-07694-z

11. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med*. 2015 Aug 5;7(299):299ra122-299ra122. doi:10.1126/scitranslmed.aab3719
12. Moor M, Bennett N, Plečko D, Horn M, Rieck B, Meinshausen N, et al. Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *eClinicalMedicine*. 2023 Aug 1;62. doi:10.1016/j.eclinm.2023.102124 PubMed PMID: 37588623.
13. Haas R, McGill S. Artificial Intelligence for the Prediction of Sepsis in Adults. *Can J Health Technol*. 2022 Mar 15;2. doi:10.51731/cjht.2022.283
14. Prescott HC, Iwashyna TJ, Blackwood B, Calandra T, Chlan LL, Choong K, et al. Understanding and Enhancing Sepsis Survivorship. Priorities for Research and Practice. *Am J Respir Crit Care Med*. 2019 Oct 15;200(8):972–81. doi:10.1164/rccm.201812-2383CP
15. Wagenlehner FME. Management of the Urosepsis Syndrome. In: Bjerklund Johansen TE, Cai T, editors. *Guide to Antibiotics in Urology* [Internet]. Cham: Springer International Publishing; 2024 [cited 2026 Apr 3]. p. 153–8. Available from: https://doi.org/10.1007/978-3-319-92366-6_12 doi:10.1007/978-3-319-92366-6_12
16. Xu N, Zhang L. Sepsis-associated Acute Kidney Injury: Epidemiology, Pathophysiology, and Management Strategies in Critical Care. *Nat Cell Sci*. 2025 Mar 30;3(1):54–61. doi:10.61474/ncs.2024.00011
17. Bjerklund Johansen TE, Tandogdu Z, Cai T, Naber KG. Classification of Urinary Tract Infections: A Contemporary Guide for Clinical Practice and Research. In: Bjerklund Johansen TE, Cai T, editors. *Guide to Antibiotics in Urology* [Internet]. Cham: Springer International Publishing; 2024 [cited 2026 Apr 3]. p. 53–67. Available from: https://doi.org/10.1007/978-3-319-92366-6_5 doi:10.1007/978-3-319-92366-6_5

Acknowledgements

We thank the investigators of the SERPENS study for their contribution to data collection; a full list of SERPENS investigators is provided in Tandogdu et al. (8). We also thank Adamos Hadjivasiliou and Bane Trajkovic for discussions and input during the initial stages of this project.

Funding

A.Dj. discloses support for the research of this work from UKRI [grant EP/S021612/1] through the UKRI Centre for Doctoral Training in AI-enabled Healthcare Systems at UCL. All other authors declare no relevant funding.

Author contributions

A.Dj. performed the data preprocessing and analysis, developed the predictive models, conducted the model interpretation analyses, and contributed to manuscript writing. S.C. contributed to the interpretation of results, structured the manuscript, and wrote the manuscript through multiple rounds of revision. T.E.B.J. and F.W. provided clinical expertise and critically revised the manuscript. Z.T. provided supervision, including contributing the data, study concept, and clinical expertise. I.D. provided supervision, including contributing the interpretation of the analyses and modelling, structuring the manuscript, and expertise in AI/machine learning methodology. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Files

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

- [Djuricasupplementarymaterial.docx](#)