A Appendix

A.1 Data distribution by splits

	Counts					
Task	Split	Patients	ICU Episodes	Timesteps	Labels	
					Positive	Negative
Physiological Decompensation	CV-1	5125	6215	528425	10283	518142
	CV-2	5129	6134	507892	10821	497071
	CV-3	5141	6264	511289	10426	500863
	CV-4	5102	6297	527853	11020	516833
	Test	3683	4463	367533	6931	360602
In-Hospital Mortality	CV-1	2929	3382	162063	441	2941
	CV-2	2917	3331	159566	466	2865
	CV-3	2888	3356	160732	439	2917
	CV-4	2936	3410	163284	477	2933
	Test	2119	2453	117500	283	2170
Length of Stay	CV-1	5151	6245	532403	Refer to Table S2	
	CV-2	5145	6154	510227		
	CV-3	5160	6286	514147		
	CV-4	5117	6314	530331		
	Test	3698	4483	369350		

Table S1. Data distribution by splits. For the physiological decompensation and length of stay tasks, timesteps are taken as samples as the predictions are made every hourly timesteps, while for the in-hospital mortality task, ICU episodes are taken as samples as the predictions are made at a fixed timestep. Here, *CV* refers to the training *Cross-Validation* Folds.

A.2 Class distribution for length of stay

Class Label	Class Description (Days)	CV-1	CV-2	CV-3	CV-4	Test
0	<1	131913	129634	131693	133186	95439
1	1 - 2	85311	83558	84065	85818	61372
2	2 - 3	56353	54074	54007	54780	38858
3	3 - 4	39416	37605	38106	38054	27142
4	4 - 5	29384	27982	28760	28573	20171
5	5 - 6	22830	22384	22360	22626	15878
6	6 - 7	18816	18612	18626	18582	12940
7	7 - 8	15925	15583	15697	15863	10953
8	8 - 14	62655	58512	59905	60611	40856
9	>14	69800	62283	60928	72238	45741
Total		532403	510227	514147	530331	369350

Table S2. Class distribution for Length of Stay

A.3 Algorithm hyperparameters

Classifier	Hyperparameters
Random Forest	num of estimators=300, criterion="gini", max depth=None, min samples split=2, min samples leaf=1
LSTM	epochs=30, hidden size=128, batch size=8, num of layers=1, patience=10, dropout rate=0, learning rate=1e-4, weight decay=0.0

Table S3. Hyperparameters for classifiers

A.4 Shapley Values

Shapley values come from game theory and are used to estimate the impact of a feature on a system's output. Feature impact is defined as the variation in the output of the model when the feature is observed versus when it is unknown.

Shapley values belong to a category of methods denominated additive. In particular, the additivity is formulated as

$$f(x) = \varphi_0(f,x) + \sum_{i=1}^{M} \varphi_i(f,x)$$

where $f(x)$ is the prediction made by the model, x are the features fed to the model, M is the number of features, φ_i is the Shapley value of the i-th feature, and $\varphi_0 = E[f(x)]$ is the expected value of the model over the training dataset. Also, this assumption ensures the values correctly reflect the difference between the expected model output and the output for a particular prediction.

The Shapley value of a feature is computed via

$$\begin{aligned} \varphi_i(f,x) &= \sum_{S \subseteq S_{all}/\{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \\ &= \sum_{S \subseteq S_{all}/\{i\}} \frac{1}{(M \text{ choose } |S|)(M-|S|)} [f_x(S \cup \{i\}) - f_x(S)] \end{aligned} \tag{1}$$

where S is a subset of all M input features, and $f_x(S) = E[f(x)|x_s]$ with x_s in a subset of the input features with only those belonging to S present.

In this study we used the SHAP library [13] and its optimisation for tree-based classifiers to compute the Shapley values.

A.5 Significance tests

ML Classification Model	Base Model	Secondary Models			
		S	S + NCR	S + CB	S + Ours
Random Forest	S	-	1	13.25	0
	S + NCR	99	-	94.82	26.13
	S + CB	86.75	5.18	-	1.07
	S + Ours	100	73.87	98.93	-
LSTM	S	-	0	0	0
	S + NCR	100	-	100	0
	S + CB	100	0	-	0
	S + Ours	100	100	100	-

(a) In-Hospital Mortality

ML Classification Model	Base Model	Secondary Models			
		S	S + NCR	S + CB	S + Ours
Random Forest	S	-	81.4	69	0
	S + NCR	18.6	-	32.4	0
	S + CB	31	67.6	-	0
	S + Ours	100	100	100	-
LSTM	S	-	0	0	0
	S + NCR	100	-	73	0
	S + CB	100	27	-	0
	S + Ours	100	100	100	-

(b) Physiological Decompensation

ML Classification Model	Base Model	Secondary Models			
		S	S + NCR	S + CB	S + Ours
Random Forest	S	-	22.1	100	0
	S + NCR	77.9	-	100	0
	S + CB	0	0	-	0
	S + Ours	100	100	100	-
LSTM	S	-	0	0	0
	S + NCR	100	-	100	0
	S + CB	100	0	-	0
	S + Ours	100	100	100	-

(c) Length of Stay

Table S4. Statistical Significance Matrix with Bootstrap Resampling. All the scores are percentages of samples where the base model performs better than the secondary model. Each sample is built by resampling the original test set and then scoring the base/secondary model on it. For example, the last row in (a) shows the base model (LSTM with S + Ours) is better than the secondary models (LSTM with S or S + NCR or S + CB) on 100% samples (i.e. with statistical significance). Here, S refers to Structured, NCR to Neural Concept Recognizer[16], CB to ClinicalBERT, and Ours to our phenotyping model.

A.6 4-Fold cross validation results

Classification Model	Features Design	4-Fold Cross Validation Aggregate			
		AUC-ROC		AUC-PR	
		Mean	SD	Mean	SD
SAPS-II	-	0.754	0.006	0.322	0.031
APACHE-III	-	0.732	0.008	0.326	0.018
Random Forest	S	0.810	0.008	0.418	0.018
	S + NCR	0.819	0.014	0.472	0.013
	S + CB	0.804	0.012	0.423	0.005
	S + Ours	**0.834**	0.008	**0.477**	0.016
LSTM	S	-	-	-	-
	S	0.829	0.007	0.441	0.016
	S + NCR	0.836	0.011	0.478	0.008
	S + CB	0.829	0.007	0.459	0.007
	S + Ours	**0.845**	0.004	**0.496**	0.014

(a) In-hospital mortality

Classification Model	Features Design	4-Fold Cross Validation Aggregate			
		AUC-ROC		AUC-PR	
		Mean	SD	Mean	SD
Random Forest	S	0.815	0.003	0.127	0.009
	S + NCR	0.820	0.003	0.125	0.007
	S + CB	0.818	0.004	0.123	0.008
	S + Ours	**0.844**	0.004	**0.165**	0.013
LSTM	S	-	-	-	-
	S	0.819	0.003	0.136	0.016
	S + NCR	0.820	0.003	0.134	0.013
	S + CB	0.821	0.006	0.128	0.022
	S + Ours	**0.833**	0.008	**0.144**	0.023

(b) Physiological decompensation

Classification Model	Features Design	4-Fold Cross Validation Aggregate			
		Kappa		MAD	
		Mean	SD	Mean	SD
Random Forest	S	0.381	0.005	142.010	4.665
	S + NCR	0.382	0.008	148.003	4.180
	S + CB	0.369	0.005	149.221	3.789
	S + Ours	**0.405**	0.006	**116.940**	5.674
LSTM	S	-	-	-	-
	S	0.375	0.003	134.373	17.293
	S + NCR	0.393	0.013	127.165	17.484
	S + CB	0.374	0.015	127.678	8.608
	S + Ours	**0.416**	0.012	**116.198**	6.904

(c) Length of Stay

Table S5. Results for (a) In-Hospital Mortality, (b) Physiological Decompensation, and (c) Length of Stay on the training set. The best score for each classifier is highlighted in bold. Here, S refers to Structured, NCR to Neural Concept Recognizer[16], CB to ClinicalBERT, and Ours to our phenotyping model.

A.7 Ablation study on phenotype persistency

Model	Phenotypic propagation	4-fold Cross Validation Aggregate				Test Set	
		AUC-ROC		AUC-PR		AUC-ROC	AUC-PR
		Mean	SD	Mean	SD		
RF	without	0.807	0.008	0.413	0.021	0.799 (0.772, 0.824)	0.351 (0.297, 0.407)
	with	**0.834**	0.008	**0.477**	0.016	**0.845 (0.826, 0.873)**	**0.462 (0.404, 0.524)**
LSTM	without	0.833	0.014	0.457	0.024	0.831 (0.807, 0.853)	0.421 (0.361, 0.483)
	with	**0.844**	0.004	**0.495**	0.013	**0.845 (0.823, 0.868)**	**0.464 (0.405, 0.523)**

(a) In-hospital Mortality

Model	Phenotypic propagation	4-fold Cross Validation Aggregate				Test Set	
		AUC-ROC		AUC-PR		AUC-ROC	AUC-PR
		Mean	SD	Mean	SD		
RF	without	0.812	0.002	0.125	0.007	0.820 (0.815, 0.825)	0.127 (0.120, 0.135)
	with	**0.844**	0.004	**0.165**	0.013	**0.845 (0.840, 0.850)**	**0.180 (0.171, 0.190)**
LSTM	without	0.827	0.007	**0.146**	0.017	**0.841 (0.842, 0.851)**	**0.149 (0.141, 0.156)**
	with	**0.833**	0.007	0.144	0.022	0.839 (0.834, 0.844)	0.145 (0.138, 0.153)

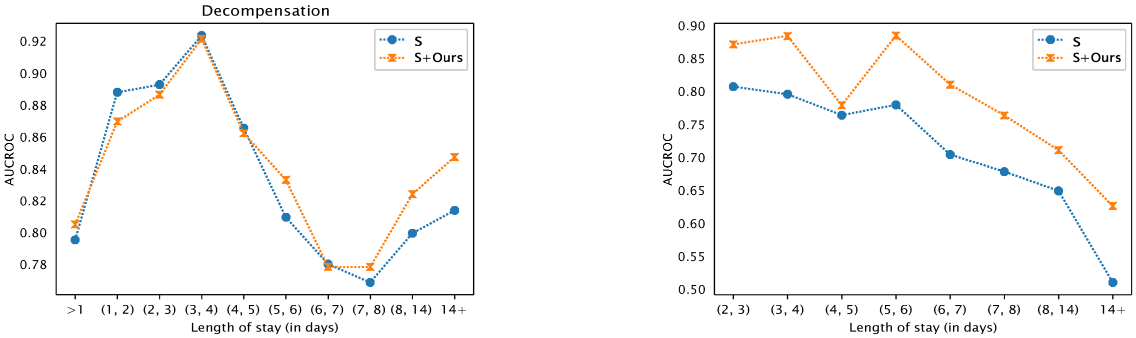
(b) Physiological Decompensation

Model	Phenotypic propagation	4-fold Cross Validation Aggregate				Test Set	
		Kappa		MAD		Kappa	MAD
		Mean	SD	Mean	SD		
RF	without	0.376	0.005	139.8	5.5	0.386 (0.380, 0.384)	135.0 (134.5, 135.6)
	with	**0.405**	0.006	**116.9**	5.6	**0.420 (0.418, 0.422)**	**110.3 (109.3, 111.3)**
LSTM	without	**0.427**	0.007	118.3	4.2	**0.441 (0.439, 0.440)**	**111.4 (110.9, 111.9)**
	with	0.416	0.012	**116.2**	6.9	0.430 (0.427, 0.432)	116.7 (116.2, 117.3)

(c) Length of Stay

Table S6. Results of ablation study on our phenotyping model to assess the importance of phenotypic modelling. Models without phenotypic propagation encounter high sparsity of phenotypes as data is only available at the timestep the clinical note is written. Models with phenotypic propagation observe phenotypes throughout all timesteps. The best score for each classifier is highlighted in bold.

A.8 Forecasts per total length of stay



(a) Physiological Decompensation

(b) In-hospital Mortality

Figure S1. AUC-ROC for (a) physiological decompensation and (b) in-hospital mortality for LSTM for patients with different LOS values. While the in-hospital mortality task benefits consistently for any duration of the ICU stay, decompensation sees the best improvements when patients stay the longest. This behaviour is a natural consequence of the fact that while near future forecasts can rely strongly on bedside measurements, forecasting without a fixed endpoint in time is significantly more difficult. Nevertheless, patients who stayed for less than two weeks still saw a benefit when introducing phenotypic features, as they calibrate better the algorithm's prediction. Here, S represents structured features and Ours refers to phenotypes from our phenotyping model.

A.9 Case study for physiological decompensation

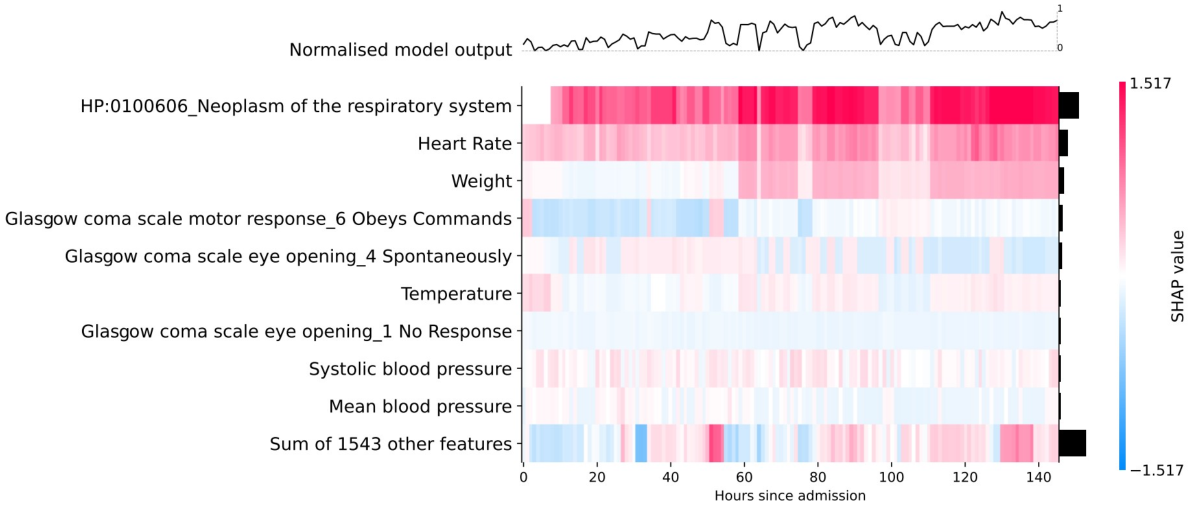


Figure S2. Time course of the physiological decompensation prediction for an illustrative patient in the test set. The top plot represents the time series of the prediction in probability (0 for no risk of decompensation, 1 for decompensation). The heatmap illustrates how the contribution of each feature (i.e., each row) varies across time for this subject. Features are sorted in decreasing order according to their importance for this patient, represented by the black horizontal bar at the right of each row. The colour of a row indicates how that feature contributes to the prediction at a moment in time, with red representing a positive contribution (i.e., that the patient will decompensate), and blue for a negative contribution. For this patient, although fluctuations in the prediction come from changes in structured data, taking into account the neoplasm of the respiratory system allows to better estimate the baseline risk of decompensation.

A.10 Feature importance for Length-of-Stay

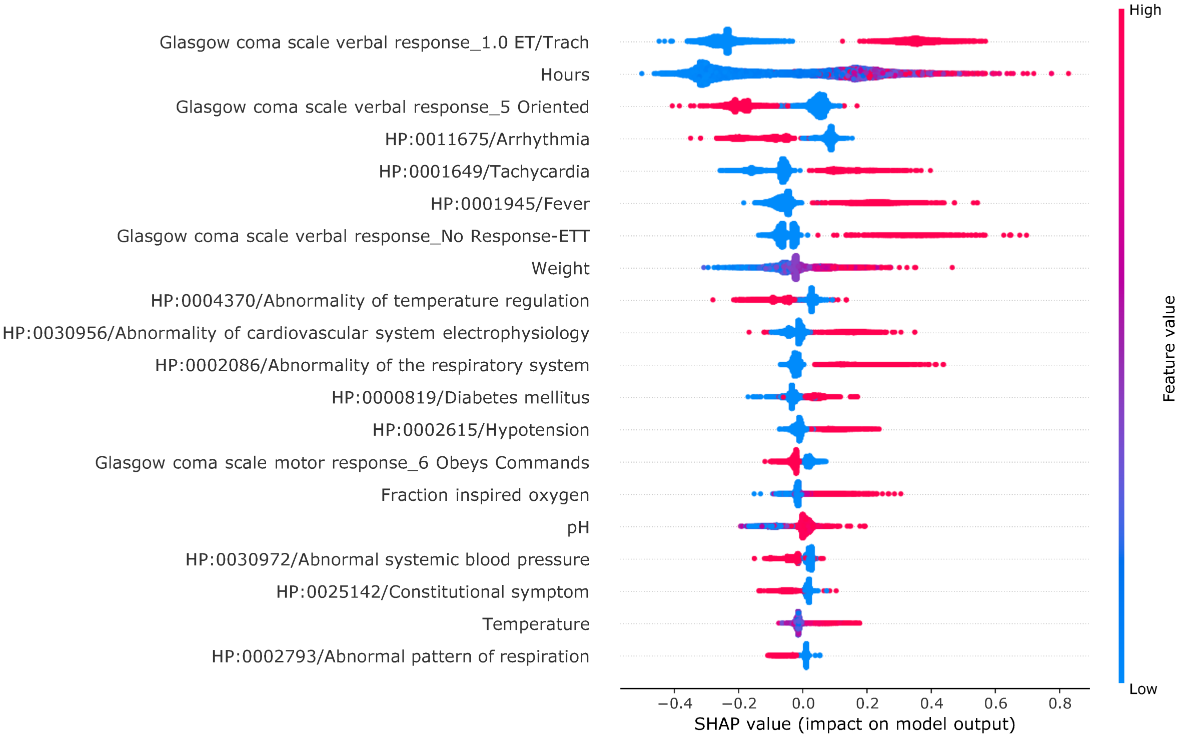
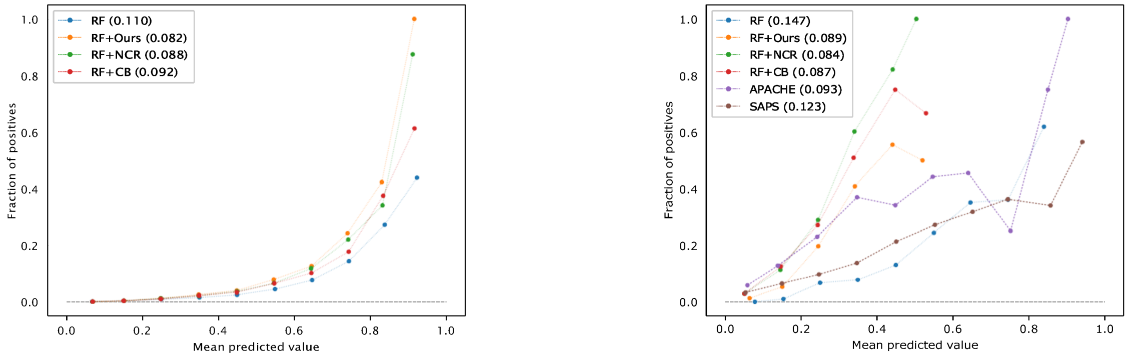


Figure S3. Top features for length-of-stay predicting stays of more than 1 week.

A.11 Calibration curves



(a) Physiological Decompensation

(b) In-hospital Mortality

Figure S4. Calibration curves with Random Forest for (a) physiological decompensation and (b) in-hospital mortality. RF in legend refers to using structured features only. Ours, NCR, CB: phenotypic features from our phenotyping model, NCR and ClinicalBERT, respectively.

A.12 Cohort study

Cohort	No. of Patients	No. of ICU Episodes	AUC-ROC
All	2119	2453	0.845
Cardiovascular Diseases	681	789	0.780
Diabetes	563	682	0.826
Cancer	277	304	0.822
Depression	119	122	0.783

(a) In-hospital Mortality.

Cohort	No. of Patients	No. of ICU Episodes	AUC-ROC
All	3683	4463	0.839
Cardiovascular Diseases	975	1197	0.792
Diabetes	927	1191	0.808
Cancer	489	565	0.806
Depression	216	240	0.820

(b) Physiological Decompensation.

Cohort	No. of Patients	No. of ICU Episodes	Kappa
All	3698	4483	0.430
Cardiovascular Diseases	980	1202	0.413
Diabetes	930	1195	0.424
Cancer	493	572	0.321
Depression	216	241	0.330

(c) Length of Stay

Table S7. Analysing the generalisability and robustness of our approach on cohorts with different diseases. The accuracies of the best LSTM models which use features from both structured and unstructured data are reported individually on each cohort for each ICU task.