THE LANCET
Digital Health

Supplementary appendix 1

Supplement to: Rasmy L, Nigo M, Kannadath BS, et al. Recurrent neural network models (CovRNN) for predicting outcomes of patients with COVID-19 on admission to hospital: model development and validation using electronic health record data. *Lancet Digit Health* 2022; published online April 21. https://doi.org/10.1016/S2589-7500(22)00049-8.

Supplemental Material A: Data Sources

Cerner Real World Data^{TM} (CRWD)

Cerner Real-World Data^{TM} is a national, de-identified, person-centric dataset solution provided by Cerner Corporation to enable researchers to leverage longitudinal electronic health records (EHR) data from contributing organizations. Cerner offers one-year free access to a COVID-19 data science workspace, which includes access to a CRWD COVID-19 de-identified data cohort hosted on the HealtheDataLab^{TM}—the Cerner data science ecosystem, built and deployed on Amazon Web Services (AWS).

Data in the CRWD are extracted from the EHR of hospitals and clinics across the United States that have consented to such use. Encounters may include pharmacy, clinical, and microbiology laboratory, admission, and billing information from affiliated patient care locations. All admissions, medication orders, dispensing, laboratory orders, and specimens are date- and time-stamped, providing a temporal relationship between treatment patterns and clinical information. Cerner de-identifies the CRWD in compliance with the Health Insurance Portability and Accountability Act (HIPAA).

In our study, we used the latest freely available version (Q3), which includes patient data up to the end of September 2020.

Optum® de-identified COVID-19 Electronic Health Record Dataset (OPTUM)

Given the urgent need to clinically understand the novel virus of COVID 19, Optum developed a low latency data pipeline that enables minimal data lag, while preserving as much clinical data as possible. The data are sourced from Optum's longitudinal EHR repository, which is derived from dozens of healthcare provider organizations in the United States, including more than 700 hospitals and 7,000 clinics. The data are certified as de-identified by an independent statistical expert, following HIPAA statistical de-identification rules, and managed according to Optum® customer data-use agreements. The COVID-19 data asset incorporates a wide swath of raw clinical data, including new, unmapped COVID-specific clinical data points from inpatient and ambulatory electronic medical records (EMRs), practice management systems, and numerous other internal systems. Information is processed from across the continuum of care, including acute inpatient stays and outpatient visits. The COVID-19 data capture point of care diagnostics specific to the COVID-19 patient during initial presentation, acute illness, and convalescence, with over 500 mapped labs and bedside observations, including COVID-19 specific testing.

The Optum COVID-19 data elements include patient-level information: demographics, mortality, and clinical interventions, such as medications prescribed and administered. The data are composed of multiple tables that can be linked by a common patient identifier (an anonymous, randomized string of characters). The COVID-19 patient base includes patients in the EHR database who have documented clinical care from January 2007 to the most current monthly data release (October 2020) with a documented diagnosis of COVID-19 or acute respiratory illness after February 1, 2020, and/or documented COVID-19 testing (positive or negative result).

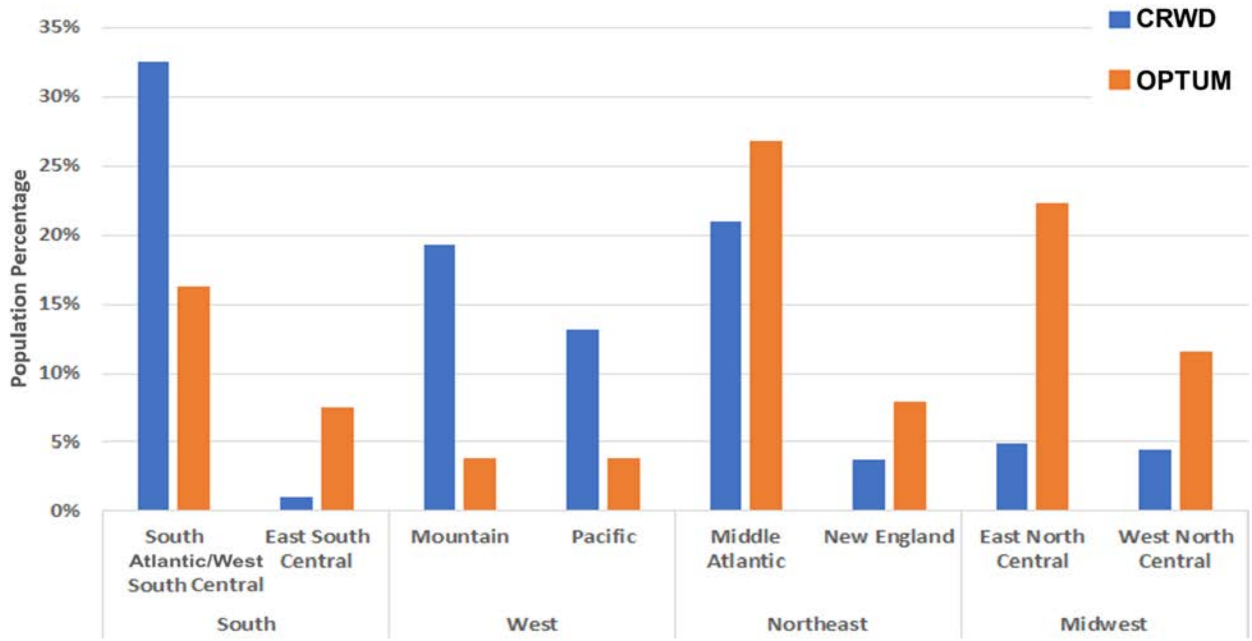
In our study, we used the 1015 version, which includes patient data until the October 15, 2020. We only included COVID-19 patients with a hospital stay longer than one day and a confirmed COVID-19 diagnosis, either through a positive COVID-19 testing results or a documented COVID-19 diagnosis code (U071).

Major differences and similarities between CRWD and OPTUM

A key difference between both datasets is that they applied different de-identification strategies. The CRWD applied date shifting for date de-identification, whereas OPTUM did not apply date shifting for events but, rather, masked the exact day for patient identifiable dates, such as date of birth. OPTUM did not provide the exact encounter disposition for expired patients, whereas this is available on CRWD.

CRWD and OPTUM used nearly the same standard codes for diagnosis, procedures, laboratory, and assessment results; thus, we did not need to apply any terminology normalization. Cerner used Multum codes for medications, whereas OPTUM used NDC codes. Thus, we used the Cerner Multum drug database mapping tool to map the OPTUM medication NDC codes to Multum codes.

The CRWD results table includes the clinical interpretations of laboratory results and does not include the normal range for numerical values. Although the OPTUM laboratory results include the normal range for numerical values, they do not include the interpretation. As in our study, we converted all numerical laboratory results into a “below normal low, normal, or above normal high” classification, by using the interpretation value provided in the CRWD, or by defining the result categories, based on the assigned normal result ranges for OPTUM. In regard to demographics, the OPTUM version that we used did not include a race group for native Alaskans, and the South Atlantic and West South regions were merged, which is not the case in the CRWD. Supplementary Figure 1 shows the geographical coverage for the CRWD and OPTUM cohorts.



Supplementary Figure 1. Geographical distribution of CRWD and OPTUM cohorts

Supplemental Material B: Data Preparation

We predefined our prediction point as the first COVID-19 admission date, and we refer to it as the index date. For training and internal validation, we excluded all patients who stayed in the hospital for less than 1 day or died within 1 day (24 hrs) to ensure that there was no information leakage and to train the model on more difficult cases. We extracted all patient data available on or before the index date.

For diagnoses, we included the diagnosis code types along with the diagnosis codes recorded before the index admission date. The CRWD and OPTUM datasets used mainly ICD-9, ICD-10, or SNOMED CT codes for recording diagnosis information. Therefore, we relied on these codes without any further normalization. We excluded diagnosis codes of other or unknown types. For medications, as the CRWD used Multum codes, whereas OPTUM used NDC codes, and given that we had access to the NDC to Multum mappings, we converted NDC codes to Multum drug identifiers that correspond to the drug generic name and major dosage form and used the multum drug identifiers and the multum therapeutic categories in our input variables. For procedures, we included all procedure codes, specifically ICD-9PCS, ICD10PCS, CPT, and HCPCS. For laboratory results, in addition to converting to categorical variables, we also mainly used LOINC codes. For an example patient record, we converted the information to look as follows: [ICD9_789.22, loinc_1244-1$High , Multumdnum_d03807, MCat_Antidiabetic, g_Female, r_White, a_87 . . .].

Supplementary Table 1: Clinical codes used to define mVent outcome on CRWD

Code Type	**Codes**
ICD-10-PCS	5A1955Z, 5A1945Z, 5A1935Z, 5A09357, 5A09457, 5A09557, 5A09358, 5A0935Z, 5A0945Z, 5A0955Z, 5A09458, 5A09558, 5A09559, 5A09459
ICD-9-PCS	93.9, 96.71, 96.72
CPT-4	94002, 94660, 94003, 78582
SNOMED CT	47545007, 243142003, 251901004, 26261000175105
LOINC	19834-1, 19835-8, 19839-0, 19840-8, 19932-3, 19976-0, 19994-3, 19996-8, 20054-3, 20055-0, 20056-8, 20058-4, 20063-4, 20068-3, 20077-4, 20079-0, 20112-9, 20116-0, 20124-4 ,33429-2, 33438-3, 33446-6, 60794-5, 76007-4, 76222-9

Supplemental Material C: Implementation Details

We used Scikit-learn package v.0.24 for LR, LGBM package v.3.1.1 for LGBM, and Pytorch v.1.7 for CovRNN. For survival evaluation and visualizations, we used lifelines package v.0.23.7. The 95-confidence intervals were calculated using 5000 stratified bootstrap replicates using Scikit Learn resampling with replacement function. For hyperparameter tuning, we used the Tree-structured Parzen Estimator (TPE) algorithm available through Optuna package v.2.5 to search for the best hyperparameters combination, using a sample cohort extracted from OPTUM data. We evaluated the same on CRWD, and it showed improved model performance compared to the default parameters used in Experiment 1 (Table 4). Therefore, for later results, we used the following hyperparameter: For LGBM, we used a learning rate (lr) of 0.05, feature fraction (proportion of features included on each iteration) as 0.55, min_child_samples (minimum number of data in one leaf) as 77, min_split_gain (minimal gain to perform split) as 0·1, n_estimators (number of boosting iterations) as 150, bagging_fraction (randomly selecting part of the data without resampling) as 0·68, bagging_freq (frequency for bagging) as 4, num_leaves (max number of leaves in one tree) as 139, reg_alpha (L1 regularization) as 0·52. For LR, we found that a weighted L2 regularized model trained with a liblinear solver (algorithm to use in optimization problem) and with C (inverse of regularization strength) as 0·004, was associated with the best performance. For RNN based models, we used the Adagrad optimizer with a starting learning rate (lr) of 0·05, weight decay (L2 penalty) of 0·0001, and the eps (term added to the denominator to improve numerical stability) as 1e-4. We used embedding and hidden dimensions of 64, as they were associated with one of the best performances, as well as the efficient model size and running time. We used a training batch size of 128 for binary classification tasks and 256 for survival models.

Supplementary Table 2: Descriptive analysis for different test sets

Characteristics	CRWD Training	CRWD Valid	CRWD Multi–Hospital Test	Hospital 1	Hospital 2	OPTUM Fine-tuning	OPTUM Test
	n = 170,626	*n* = 24,378	*n* = 48,781	*n* = 3,469	*n* = 706	*n* = 29,416	*n* = 6,724
Age on index Median (IQR)	57 (36–72)	57 (35–72)	57 (36–72)	58 (40–71)	43 (30–57)	60 (44–72)	59 (43–71)
Gender							
Female	89,844 (52%)	12,843 (52%)	25,693 (52%)	1,814 (52%)	346 (49%)	14,898 (50%)	3,339 (49%)
Male	80,269 (47%)	11,467 (47%)	22,915 (46%)	1,643 (47%)	359 (50%)	14,505 (49%)	3,380 (50%)
Race & Ethnicity							
Caucasian	116,342 (68%)	16,577 (68%)	33,278 (68%)	1,786 (51%)	623 (88%)	16,047 (54%)	3,657 (54%)
African American	24,748 (14%)	3,545 (14%)	7,034 (14%)	1,395 (40%)	40 (5%)	6,427 (21%)	1,409 (20%)
Asian	3,843 (2%)	539 (2%)	1,088 (2%)	15 (0%)	9 (1%)	742 (2%)	188 (2%)
American Indian /Alaska Native	2,919 (1%)	453 (1%)	908 (1%)	4 (0%)	1 (0%)	NA	NA
Hispanic	50,114 (29%)	7,247 (29%)	14,113 (28%)	101 (2%)	493 (69%)	4,708 (16%)	1,074 (15%)
Comorbidities							
Hypertension (HTN)	78,260 (46%)	11,274 (46%)	22,576 (46%)	2,105 (61%)	172 (24%)	18,039 (61%)	3,996 (59%)
Diabetes (DM)	43,918 (26%)	6,326 (26%)	12,471 (26%)	1,182 (34%)	126 (18%)	10,706 (36%)	2236 (33%)
Congestive Heart Failure (CHF)	24,598 (14%)	3,565 (15%)	7,189 (15%)	643 (19%)	45 (6%)	5,428 (18%)	1,140 (16%)
Chronic Kidney Disease (CKD)	23,827 (14%)	3,469 (14%)	6,794 (14%)	661 (19%)	38 (5%)	6,208 (21%)	1,309 (19%)
Cancer	13,074 (8%)	1,910 (8%)	3,826 (8%)	310 (9%)	25 (4%)	4,229 (14%)	865 (12%)

Outcomes							
Mortality (iMort)	9,324 (5%)	1,321 (5%)	2,666 (5%)	263 (7%)	33 (4%)	3,946 (13%)	885 (13%)
Median TTE	3 (1–6)	3 (1–6)	3 (1–6)	2 (1–6)	1 (1–5)	5 (3–10)	5 (3–10)
Mechanical ventilation (mVent)	23,127 (13%)	3,225 (13%)	6,556 (13%)	496 (14%)	101 (14%)	7,845 (26%)	1,737 (25%)
Intubated on first day (% of mVent)	12,270 (53%)	1,703 (52%)	3,557 (54%)	215 (43%)	66 (65%)	3,676 (46%)	790 (45%)
Median TTE	2 (1–4)	2 (1–4)	2 (1–4)	2 (1–4)	1 (1–3)	3 (2–7)	3 (2–7)
Length of stay Median (IQR)	3 (1–6)	3 (1–6)	3 (1–6)	2 (1–6)	1 (1–5)	5 (3–10)	5 (3–10)

Supplementary Table 3: Additional prediction accuracy metrics for CovRNN binary classification models

Cohort	task	AUROC	AUPRC	Specificity @ 95% Sensitivity	Sensitivity*	F1-score*	Specificity*
CRWD Training Set	iMort	95·33%	63·74%	79·62%	90·35%	42·02%	86·15%
	mVent	95·91%	86·36%	76·88%	89·95%	65·41%	86·66%
	pLOS	90·75%	71·04%	64·98%	86·70%	61·38%	77·90%
CRWD Valid Set	iMort	92·30%	49·75%	67·52%	83·80%	38·38%	85·51%
	mVent	92·56%	77·93%	60·88%	83·01%	59·33%	85·24%
	pLOS	86·58%	59·07%	56·43%	80·32%	55·83%	75·70%
CRWD Multi-hospital Test Set	iMort	93·03%	52·84%	70·93%	84·92%	39·22%	85·65%
	mVent	92·90%	79·51%	63·48%	83·39%	59·73%	85·12%
	pLOS	86·50%	60·00%	55·56%	79·74%	56·51%	76·28%
Hospital 1	iMort	91·77%	51·24%	70·87%	82·51%	44·24%	84·37%
	mVent	91·54%	73·71%	62·33%	76·01%	61·20%	87·92%
	pLOS	87·15%	57·86%	59·68%	76·96%	58·53%	79·64%
Hospital 2	iMort	97·00%	59·57%	86·18%	90·91%	51·72%	92·12%
	mVent	96·02%	85·25%	85·12%	83·17%	69·42%	90·58%
	pLOS	88·33%	61·41%	63·95%	75·42%	55·80%	80·95%
OPTUM Test Set	iMort	91·27%	70·63%	59·34%	82·94%	56·44%	83·18%
	mVent	91·46%	83·19%	55·02%	89·06%	67·54%	73·99%
	pLOS	80·97%	70·24%	35·52%	85·97%	64·42%	57·25%

*****_Sensitivity, Specificity, and F1-Score are at the best identified threshold of 7·5% for iMort, 10% for mVent, and 20% for pLOS_

Supplemental Material D: Additional Experiments

Experiment 1: Ablation study investigating the impact of different data categories

The first experiment was an ablation experiment to evaluate the added value for each clinical data category, starting with diagnosis information, followed by medication, medication categories, laboratory tests, assessments results, procedures, and, lastly, demographics.

Our experiment showed that each clinical data category contributes to an increase in the model prediction accuracy. For example, the addition of medication or laboratory results contributed to a 4% increase in the prediction accuracy for iMort or mVent tasks, not only for deep learning-based models but also for LR and LGBM (Supplementary Table 4).

Supplementary Table 4. Experiment 1- Ablation study investigating the impact of different data categories

Characteristics	**Number of covariates**	**Mortality (iMort)**			**Ventilator Use (mVent)**		
		LR	LGBM	CovRNN	LR	LGBM	CovRNN
Diagnosis only (ICD-9 / ICD-10 /SNOMED CT)	49,074	77·6	83·8	85·9	75·5	80·9	83·7
Diagnosis + Medication *(Multum dNUM & Multilevel categories)*	52,177	81·4	86·6	88·3	80·8	85·2	87·8
Diagnosis + Medication + Lab results *(LOINC codes with categorical results/ interpretations)*	80,203	85·3	90·1	92·1	84·6	89·2	91·4
Diagnosis + Medication + Lab and other assessments results + Procedures *(CPT-4, HCPCS, SNOMED CT, ICD-9/10Pcs)*	125,821	85·4	90·7	92·5	86·2	90·1	92·1
Diagnosis + Medication + Lab and other assessments results + Procedures + Demographics *(Race, gender, age, location)*	125,917	86·4	90·9	92·7	86·2	90·1	92·3

LR: logistic regression, LGBM: light gradient boost machine

Experiment 2: Model performance using the most recent visit data versus using full patient history

The second experiment was a subgroup analysis to evaluate the validity of CovRNN for new patients who were admitted to the hospital for the first time and had no past medical history available in their records. Therefore, we compared the performance of CovRNN models on a modified version of the multi-hospital test set that includes only the last (index) visit information against the original full-history multi-hospital test set.

The experiment results showed that the use of the full patient history continuously had a better performance than using only the last (index) visit information only (Supplementary Table 5). Notably, the models' performance remains acceptable without the use of previous medical records, especially for the iMort and mVent tasks, which show an AUROC of 92%. (Supplementary Table 5). For the pLOS task, there is a higher decrease in the prediction accuracy, by 3.5%. Interestingly, this decrease in pLOS prediction accuracy also aligns with the higher prediction accuracy improvement of 6% for CovRNN models compared to LR-based model for the pLOS task versus an improvement of only 3% for the iMort and mVent tasks.

Supplementary Table 5. Experiment 2 - Model performance using only last visit data versus using full patient history

Outcome	Full History	Last Visit only
In-hospital Mortality	93·0	92·2
Mechanical Ventilation	92·9	91·6
Hospital Stay > 7 days	86·5	83·0
In-hospital Mortality – Survival*	86·0	85·9
Mechanical Ventilation - Survival*	92·6	91·3

Experiment 3: Effect of possible label leakage on the need for mechanical ventilation prediction

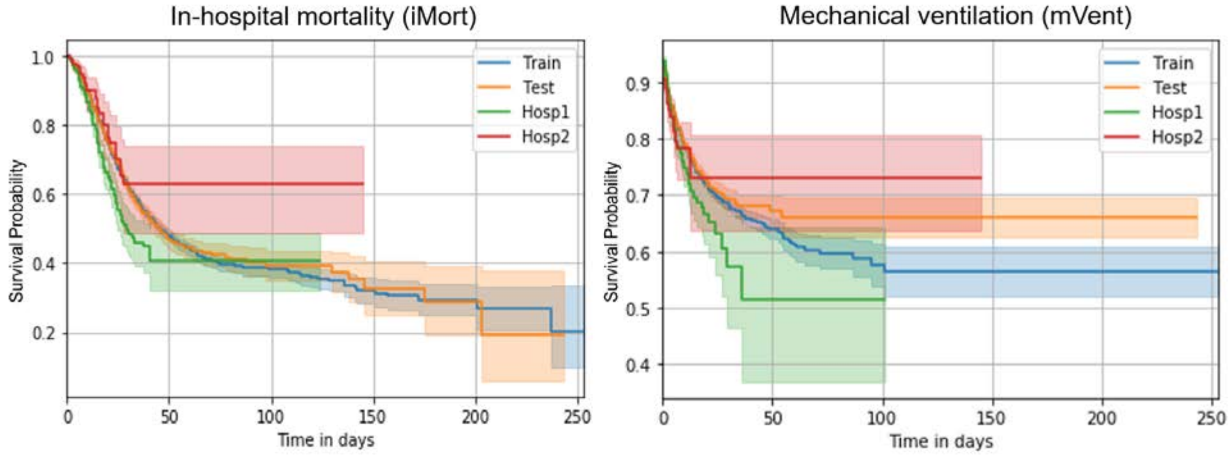
For a better understanding of the impact of any possible label leakage during model training, we conducted our third experiment using the binary classification CovRNN for the mVent task. As our cohort definition excluded any patients with a stay of less than one day, our cohort did not include any patients who died within one day of admission. Nevertheless, nearly half of the intubated patients were intubated on their first day. Therefore, we evaluated the effect of excluding such patients, which we refer to as a “restricted” dataset, and compared the performance against our original “full” cohort.

We found that training a version of the mVent binary prediction model, using the “restricted” training set, reduced the prediction accuracy by 3% on the full test set for CovRNN and 5% for LR and LGBM (SupplementaryTable 6). CovRNN performance remains constant on the “restricted” test set, regardless of which cohort it was originally trained on.

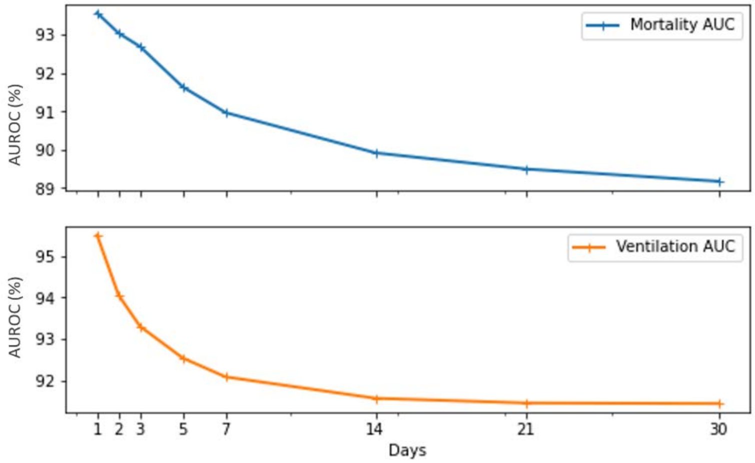
Supplementary Table 6. Experiment 3 - Effect of label leakage on the need for mechanical ventilation prediction

	Full Test Data			Restricted Test Data		
Trained	LR	LGBM	CovRNN	LR	LGBM	CovRNN
Full Data	89·5	91·2	92·9	81·5	82·8	85·9
Restricted Data	83·9	86·6	90·0	81·8	83·8	86·0

Supplemental Material E: Additional Figures

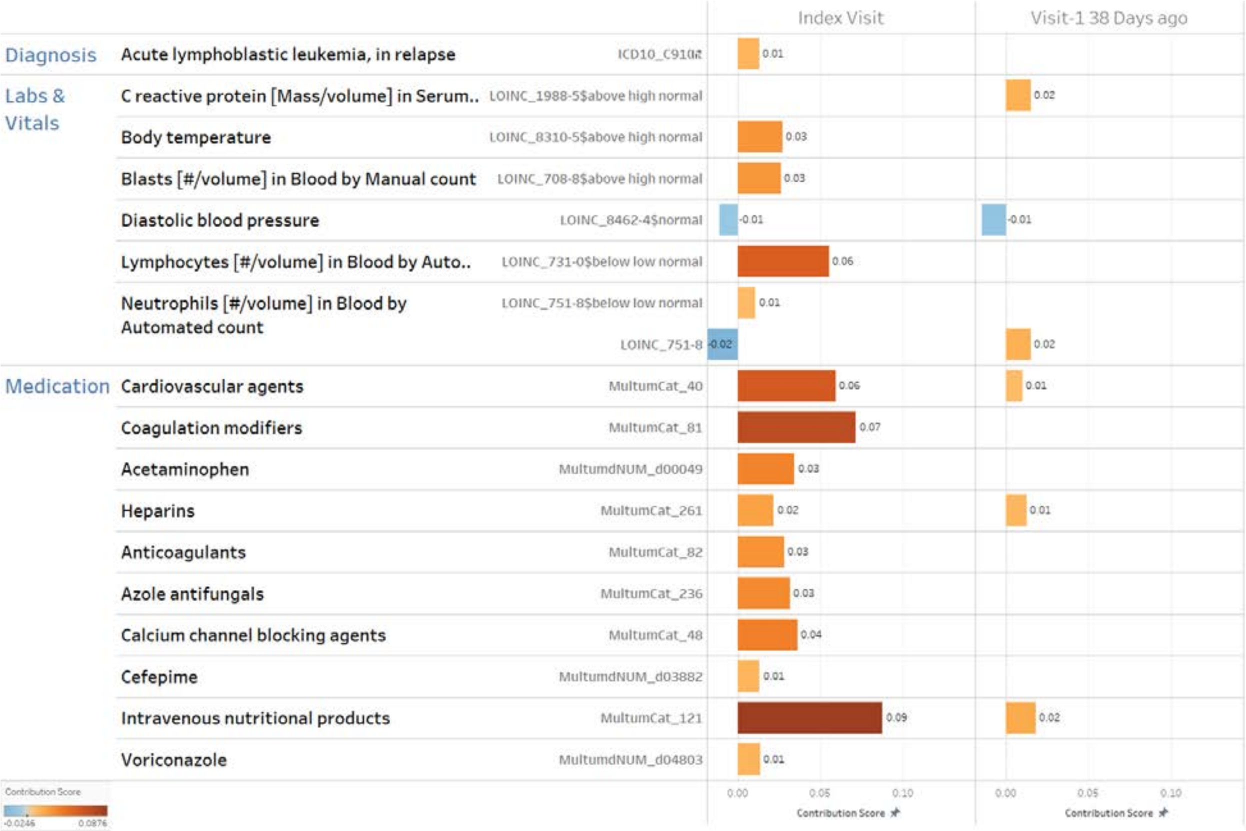


Supplementary Figure 2. K-M curve of in-hospital mortality and mechanical ventilation CRWD cohorts.



Days	In-hospital mortality	Mechanical ventilation
1	93.6%	95.5%
2	93.0%	94.1%
3	92.7%	93.3%
5	91.6%	92.5%
7	91.0%	92.1%
14	89.9%	91.6%
21	89.5%	91.5%
30	89.2%	91.5%
60	88.8%	91.4%
90	88.8%	91.4%
120	88.8%	91.4%

Supplementary Figure 3. AUROC across different time windows, using iMORT-Surv and mVent-Surv on the CRWD multi-hospital test set.



Supplementary Figure 4. Sample visit level explanation for a true positive pLOS case.

This is an example patient for whom the CovRNN model correctly predicted, with over 63% probability, would stay more than seven days in the hospital. The bar length and direction represent the contribution score calculated by the integrated gradient that predicts the prolonged hospital stay; i.e., a positive number means a stronger contribution to predicting the prolonged stay. For example, acute lymphoblastic leukemia, in relapse (ICD10_C9102), Blasts in the blood (LOINC_8867-4$above high normal), low lymphocyte counts (LOINC_731-0$below low normal), intravenous nutritional products (MultumCat_121), and coagulation modifiers (MulttumCat_81) are positively correlated to the positive prediction of the prolonged hospital stay, specifically for this patient. Notably, the contribution score can vary at the patient visit level; for example, the ordering of neutrophils count (LOINC_751-8) contribution score at an earlier visit was 0.02, whereas at the index visit, it is -0.02, as the patient had results during the index visits showing below normal neutrophils (LOINC_751-8$below_low_normal) and lymphocytes counts (LOINC_731-0$below_low_normal), which had a greater contribution to the patient predicted score.