Supplemental

Supplemental methods

Scaled Brier Score

As per Steyerberg et al 2010, the scaled Brier Score ($Brier_{scaled}$) can be calculated as follows:

$$Brier_{scaled} = 1 - \frac{Brier}{Brier_{max}}$$

where

$$Brier = \frac{1}{N}\sum_{t=1}^{N} (Y_t - p_t)^2$$
$$Brier_{max} = \bar{p}\,(1 - \bar{p})$$

given the following definitions:

$Brier_{scaled}$	scaled Brier score
Brier	Brier score
$Brier_{max}$	max Brier score
N	Number of observations = number of predictions
Y	Vector of binary outcomes
p	Vector of predictions
Y_t	Binary outcome for event t
p_t	Prediction for event t
$\bar{p}$	Mean of all predictions p

By its definition, a Brier score is a mean squared error between an observation Y and a prediction p. Consequently, a Brier score of 0 would imply a perfect prediction (e.g. p always predicts Y with 100% accuracy) and a Brier score of 1 would imply a perfect error (e.g. p predicts Y with 0% accuracy - or, rephrased - p always predicts the opposite of Y with 100% accuracy).

The scaled Brier score can be difficult to interpret. It is affected by both prediction accuracy (how well p predicts Y, including calibration) and event incidence, scaled by the performance relative to a trivial solution $Brier_{max}$. A perfect prediction, with $Brier = 0$, would result in $Brier_{scaled} = 1$. Classically, the range of $Brier_{scaled}$ is quoted as [0, 1], assuming that the lower limit of performance is to be at least as accurate as $Brier_{max}$.

$Brier_{max}$ is the equivalent performance of a trivial, noninformative prediction at the event rate. For example, for an event with 30% incidence, this would be the Brier score performance of always predicting that the event would occur with 30% probability.

We now proceed to examine different Brier scores and how their interpretation can vary by event incidence.

In the case of an event with 50% incidence:
A Brier score of 0.25 would be equal to $Brier_{max}$, which would be the noninformative prediction.
A Brier score of 0 (perfect prediction) would result in a $Brier_{scaled}$ of 1.
A Brier score of 0.25 would result in a $Brier_{scaled}$ of 0, which would indicate prediction equivalent to a trivial solution of always predicting p = 0.50.
A Brier score of 1 (perfect error) would result in a $Brier_{scaled}$ of -3.0.

In the case of an event with 10% incidence:
A Brier score of 0.09 would be equal to $Brier_{max}$, which would be the noninformative prediction.
A Brier score of 0 (perfect prediction) would result in a $Brier_{scaled}$ of 1.
A Brier score of 0.09 would result in a $Brier_{scaled}$ of 0, which would indicate prediction equivalent to a trivial solution of always predicting p = 0.10.
A Brier score of 1 (perfect error) would result in a $Brier_{scaled}$ of -10.111.

In the case of an event with 5% incidence:
A Brier score of 0.0025 would be equal to $Brier_{max}$, which would be the noninformative prediction.
A Brier score of 0 (perfect prediction) would result in a $Brier_{scaled}$ of 1.
A Brier score of 0.0025 would result in a $Brier_{scaled}$ of 0, which would indicate prediction equivalent to a trivial solution of always predicting p = 0.05.
A Brier score of 1 (perfect error) would result in a $Brier_{scaled}$ of -399.

As we can see, $Brier_{scaled}$ technically has a range of (-∞, 1]. In this case, we note that in situations where $Brier_{scaled}$ < 0, the model is less informative than a trivial solution.

Finally, Brier scores - and, consequently, scaled Brier scores - are also affected by calibration due to the difference between Y and p as a mean squared error. Even with perfect discrimination (that is, being 100% able to distinguish between $Y_t = 0$ and $Y_t = 1$), if predictions were poorly calibrated - for example, instead of ranging from [0, 1], ranged instead from [0.45, 0.55] with perfect prediction accuracy (e.g. p_t = 0.45 always predicted$Y_t = 0$, and p_t = 0.55 always predicted $Y_t = 1$):

The Brier score of this poorly calibrated function would be 0.2025.

In the case of an event with 50% incidence, with a resulting $Brier_{max} = 0.25$, would result in a scaled Brier score of 0.19, indicating a model that is more informative than the noninformative model $Brier_{max}$.

In the case of an event with 10% incidence, with a resulting $Brier_{max} = 0.09$, would result in a scaled Brier score of -1.25, indicating a model that is less informative than the noninformative model $Brier_{max}$.

Supplemental Table 1a. A list of predictor data features extracted, grouped by type.

Demographics (2 features): Gender, Patient Age at Visit
Vitals (8 features): Temperature, Respiration rate, Heart rate (HR), Oxygen saturation (O2Sat), End-tidal CO2 (EtCO2), Systolic blood pressure (SBP), Diastolic blood pressure (DBP), Mean arterial pressure (MAP)
Laboratory (34 features): Sodium, Potassium, Chloride, Bicarbonate, Blood urea nitrogen (BUN), Creatinine, Creatinine clearance, Glucose, Magnesium, Phosphate, Aspartate aminotransferase (AST), Albumin, Alkaline phosphatase, Ammonia, Amylase, Direct bilirubin, Indirect bilirubin, Total bilirubin, Calcium, Total protein, Ionized calcium, Creatinine kinase, Hemoglobin A1c, Iron, LDL cholesterol, Lactate, Lactate dehydrogenase, Lipase, Troponin, Partial thromboplastin time (PTT), D-dimer, Fibrinogen, Haptoglobin, International normalized ratio (INR)
Blood gases (12 features): Arterial blood gas (pH, pO2, pCO2, base excess, bicarbonate, O2 saturation), Venous blood gas (pH, pO2, pCO2, base excess, bicarbonate, O2 saturation)
Oxygen therapy (14 features): RoomAir, Result_FiO2, NasalCannulaOrSimpleMask, NasalCannulaOrSimpleMask_fio2, NasalCannulaOrSimpleMask_flow, ModerateFlowNasalCannulaOrMask, ModerateFlowNasalCannulaOrMask_fio2, ModerateFlowNasalCannulaOrMask_flow, NocturnalNippv, Nonrebreather, Nonrebreather_flow, VenturiMask, VenturiMask_fio2, VenturiMask_flow,

Supplemental Table 1b. A list of oxygen therapy data features that were used to verify presence of the outcome AdvRS but not provided to the model for training.

Oxygen therapy (27 features): *Nippv, NippvEpap, NippvIpap, NippvPctLeak, NippvRespiratoryRateSet, NippvRespiratoryRateSpontaneous, NippvRoute, SpontaneousBreathingTrial, TrachCollar, TrachCollar_flow, HeatedHumidifiedHighFlow, HeatedHumidifiedHighFlow_FiO2, HeatedHumidifiedHighFlow_flow, BPAP,* *Intubated, VentilatorAutoPEEP, VentilatorFiO2, VentilatorMeanAirwayPressure, VentilatorPEEP, Ventilator pressure control above PEEP, VentilatorPeakInspiratoryPressure, VentilatorPlateauPressure, VentilatorPressureSupport, VentilatorRateSet, VentilatorRateTotal, VentilatorTidalVolumeExhaled, VentilatorTidalVolumeSetPerKg*

Supplemental Table 2. Confusion matrix breakdowns by method and dataset. **TP** true positive, **FP** false positive, **TN** true negative, **FN** false negative.

method	TP	FP	TN	FN
train				
PARFAIT	5471 (5319 - 5624)	629 (564 - 694)	6550 (6487 - 6612)	1619 (1465 - 1773)
MEWS > 3	1165 (1132 - 1198)	575 (538 - 612)	6603 (6564 - 6643)	5926 (5891 - 5960)
MEWS > 4	522 (506 - 537)	132 (102 - 162)	7047 (7017 - 7077)	6569 (6553 - 6584)
MEWS > 5	244 (233 - 255)	37 (19 - 54)	7142 (7125 - 7158)	6846 (6837 - 6856)
test				
PARFAIT	1256 (1240 - 1272)	2845 (2662 - 3028)	14862 (14679 - 15045)	528 (511 - 545)
MEWS > 3	292 (258 - 327)	1409 (1361 - 1458)	16298 (16249 - 16346)	1491 (1457 - 1525)
MEWS > 4	131 (116 - 146)	327 (309 - 345)	17380 (17362 - 17398)	1653 (1638 - 1667)
MEWS > 5	61 (51 - 72)	95 (79 - 111)	17612 (17596 - 17628)	1722 (1711 - 1733)
external validation				
PARFAIT	1808 (1765 - 1851)	6652 (6322 - 6981)	30913 (30584 - 31243)	770 (727 - 813)
MEWS > 3	587	3301	34264	1991
MEWS > 4	273	837	36728	2305
MEWS > 5	123	277	37288	2455
temporal validation, COVID-tested, any result				
PARFAIT	312 (308 - 317)	559 (514 - 603)	1469 (1425 - 1514)	61 (56 - 65)
MEWS > 3	89	125	1903	284
MEWS > 4	48	36	1992	325
MEWS > 5	25	12	2016	348
temporal validation, COVID+				
PARFAIT	88 (87 - 90)	59 (51 - 66)	204 (197 - 212)	6 (4 - 7)
MEWS > 3	27	23	240	67
MEWS > 4	15	7	256	79
MEWS > 5	8	3	260	86

Supplemental Table 3. Prediction metrics for PARFAIT in comparison to MEWS > 3-5. ***SEN*** sensitivity, ***SPE*** specificity, ***ACC*** accuracy, ***PPV*** positive predictive value, ***NPV*** negative predictive value, ***NNE*** number needed to examine, ***PRE*** precision, ***REC*** recall, ***true prev*** true prevalence, ***AUROC*** area under the receiver operating curve, ***AUPRC*** area under the precision recall curve

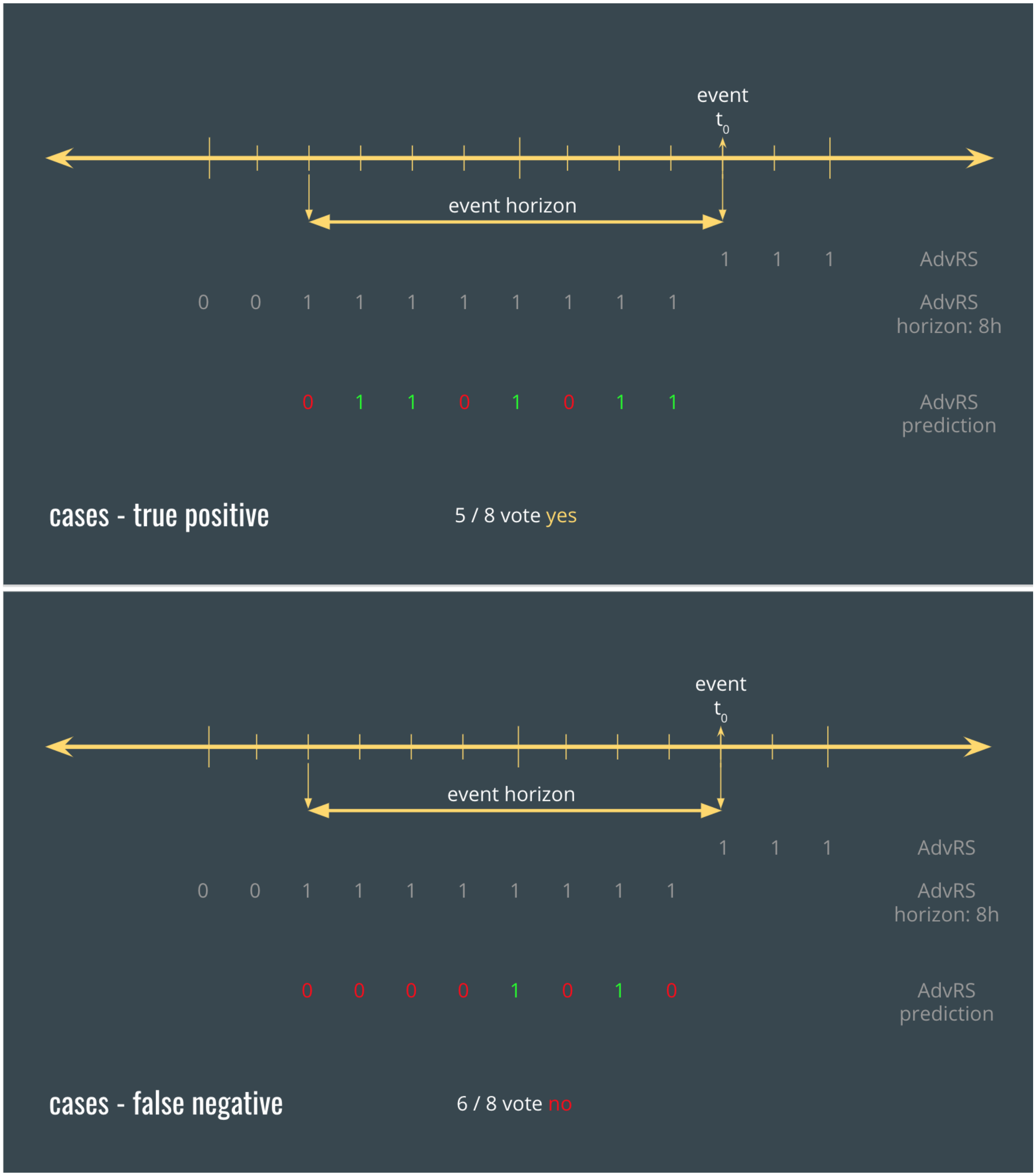
method	SEN	SPE	ACC	PPV	NPV	PRE	REC	true prev	calibration slope	calibration intercept	AUROC	AUPRC
train												
PARFAIT	0.77 (0.75 - 0.79)	0.91 (0.90 - 0.92)	0.84 (0.83 - 0.85)	0.90 (0.89 - 0.91)	0.80 (0.79 - 0.82)	0.90 (0.89 - 0.91)	0.77 (0.75 - 0.79)	0.50 (0.50 - 0.50)	7.36 (7.11, 7.62)	-3.43 (-3.48, -3.38)	0.93 (0.92 - 0.93)	0.93 (0.93 - 0.94)
MEWS > 3	0.16 (0.16 - 0.17)	0.92 (0.91 - 0.93)	0.54 (0.54 - 0.55)	0.67 (0.65 - 0.69)	0.53 (0.52 - 0.53)	0.67 (0.65 - 0.69)	0.16 (0.16 - 0.17)	0.50 (0.50 - 0.50)	3.59 (3.53, 3.66)	-2.81 (-2.82, -2.8)	0.57 (0.57 - 0.58)	0.60 (0.59 - 0.60)
MEWS > 4	0.07 (0.07 - 0.08)	0.98 (0.98 - 0.99)	0.53 (0.53 - 0.53)	0.80 (0.76 - 0.84)	0.52 (0.52 - 0.52)	0.80 (0.76 - 0.84)	0.07 (0.07 - 0.08)	0.50 (0.50 - 0.50)	3.63 (3.56, 3.69)	-2.83 (-2.83, -2.82)	0.58 (0.57 - 0.58)	0.60 (0.59 - 0.60)
MEWS > 5	0.03 (0.03 - 0.04)	0.99 (0.99 - 1.00)	0.52 (0.52 - 0.52)	0.87 (0.81 - 0.93)	0.51 (0.51 - 0.51)	0.87 (0.81 - 0.93)	0.03 (0.03 - 0.04)	0.50 (0.50 - 0.50)	3.63 (3.54, 3.72)	-2.83 (-2.84, -2.82)	0.58 (0.57 - 0.58)	0.60 (0.59 - 0.60)
test												
PARFAIT	0.70 (0.69 - 0.71)	0.84 (0.83 - 0.85)	0.83 (0.82 - 0.84)	0.31 (0.29 - 0.32)	0.97 (0.96 - 0.97)	0.31 (0.29 - 0.32)	0.70 (0.69 - 0.71)	0.09 (0.09 - 0.09)	7.27 (6.96, 7.59)	-3.41 (-3.46, -3.36)	0.85 (0.85 - 0.86)	0.44 (0.43 - 0.46)
MEWS > 3	0.16 (0.14 - 0.18)	0.92 (0.92 - 0.92)	0.85 (0.85 - 0.85)	0.17 (0.16 - 0.19)	0.92 (0.91 - 0.92)	0.17 (0.16 - 0.19)	0.16 (0.14 - 0.18)	0.09 (0.09 - 0.09)	3.4 (3.12, 3.68)	-2.78 (-2.82, -2.74)	0.57 (0.56 - 0.59)	0.15 (0.14 - 0.16)
MEWS > 4	0.07 (0.06 - 0.08)	0.98 (0.98 - 0.98)	0.90 (0.90 - 0.90)	0.29 (0.26 - 0.32)	0.91 (0.91 - 0.91)	0.29 (0.26 - 0.32)	0.07 (0.06 - 0.08)	0.09 (0.09 - 0.09)	3.38 (3.09, 3.67)	-2.79 (-2.83, -2.75)	0.57 (0.56 - 0.59)	0.15 (0.14 - 0.15)
MEWS > 5	0.03 (0.03 - 0.04)	0.99 (0.99 - 1.00)	0.91 (0.91 - 0.91)	0.39 (0.33 - 0.46)	0.91 (0.91 - 0.91)	0.39 (0.33 - 0.46)	0.03 (0.03 - 0.04)	0.09 (0.09 - 0.09)	3.42 (3.17, 3.68)	-2.79 (-2.84, -2.75)	0.58 (0.56 - 0.59)	0.15 (0.14 - 0.15)
external validation												
PARFAIT	0.70 (0.68 - 0.72)	0.82 (0.81 - 0.83)	0.82 (0.81 - 0.82)	0.21 (0.21 - 0.22)	0.98 (0.97 - 0.98)	0.21 (0.21 - 0.22)	0.70 (0.68 - 0.72)	0.06	6.42 (6.2, 6.65)	-3.65 (-3.69, -3.62)	0.84 (0.84 - 0.85)	0.37 (0.36 - 0.38)
MEWS > 3	0.23	0.91	0.87	0.15	0.95	0.15	0.23	0.06	4.87	-3.4	0.61	0.13
MEWS > 4	0.11	0.98	0.92	0.25	0.94	0.25	0.11	0.06	4.66	-3.38	0.61	0.13
MEWS > 5	0.05	0.99	0.93	0.31	0.94	0.31	0.05	0.06	4.82	-3.41	0.61	0.13
temporal validation, COVID-tested, any result												
PARFAIT	0.84 (0.82 - 0.85)	0.72 (0.70 - 0.75)	0.74 (0.73 - 0.76)	0.36 (0.34 - 0.37)	0.96 (0.96 - 0.96)	0.36 (0.34 - 0.37)	0.84 (0.82 - 0.85)	0.16	6.86 (6.27, 7.44)	-3.15 (-3.29, -3.01)	0.86 (0.86 - 0.87)	0.60 (0.59 - 0.62)
MEWS > 3	0.24	0.94	0.83	0.42	0.87	0.42	0.24	0.16	3.73	-2.18	0.66	0.32
MEWS > 4	0.13	0.98	0.85	0.57	0.86	0.57	0.13	0.16	3.58	-2.17	0.65	0.32
MEWS > 5	0.07	0.99	0.85	0.67	0.85	0.67	0.07	0.16	3.67	-2.19	0.66	0.33
temporal validation, COVID+												
PARFAIT	0.94 (0.92 - 0.96)	0.78 (0.75 - 0.81)	0.82 (0.80 - 0.84)	0.60 (0.57 - 0.63)	0.97 (0.97 - 0.98)	0.60 (0.57 - 0.63)	0.94 (0.92 - 0.96)	0.26	5.37 (5.0, 5.74)	-2.34 (-2.48, -2.2)	0.93 (0.92 - 0.95)	0.80 (0.76 - 0.84)
MEWS > 3	0.29	0.91	0.75	0.54	0.78	0.54	0.29	0.26	2.74	-1.49	0.73	0.5
MEWS > 4	0.16	0.97	0.76	0.68	0.76	0.68	0.16	0.26	2.85	-1.51	0.71	0.49
MEWS > 5	0.09	0.99	0.75	0.76	0.75	0.76	0.09	0.26	2.79	-1.51	0.71	0.48

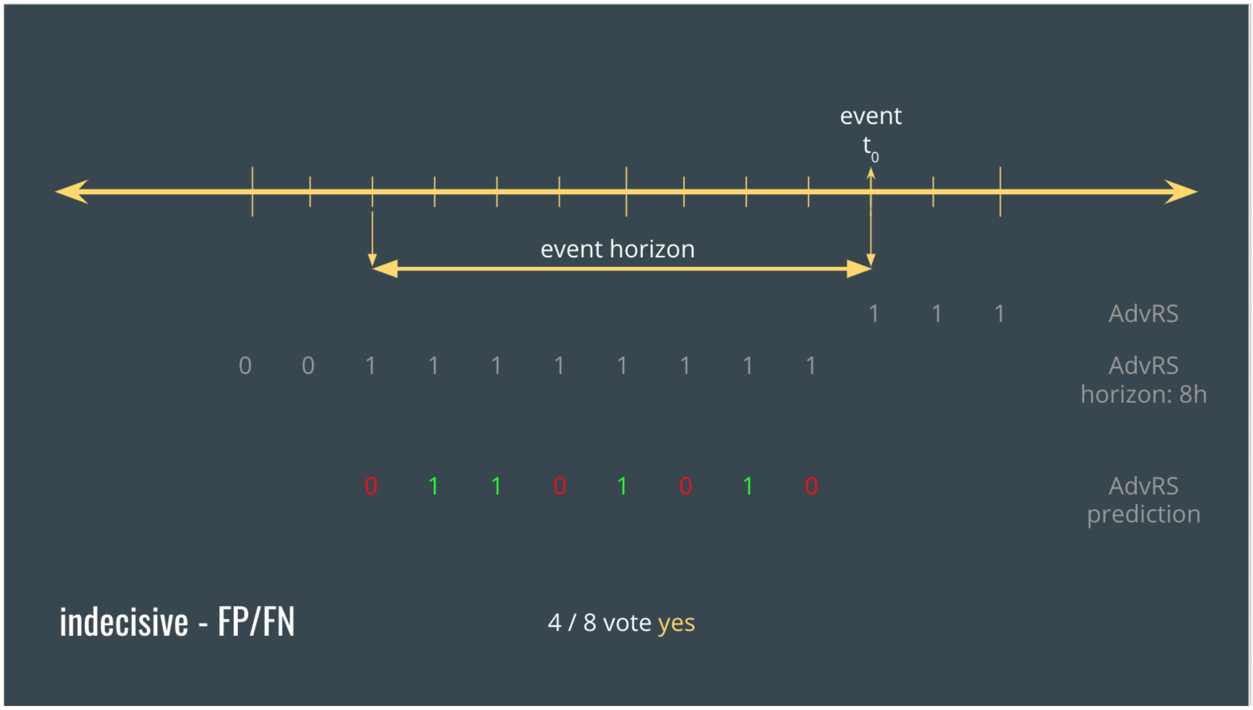
Supplemental Table 4. Summary statistics of Scaled Brier Scores from each of the 5 folds in each dataset. The scaled Brier score incorporates aspects of both discrimination and calibration, where 1.0 is perfectly informative and 0.0 is uninformative relative to the event rate. Note that the Brier score (used to generate the scaled Brier score) can range from 0.0 to 1.0 as a mean squared error. However, although the scaled Brier score is classically quoted as ranging from 1.0 (perfectly informative) to 0.0 (uninformative), the lower bound is -∞, where scores < 0 imply scores less informative than the trivial solution of predicting the event rate.

method	dataset	scaled Brier score
MEWS > 3	train (Hospitals 1-3)	0.01 (0.01, 0.01)
	test (Hospitals 1-3)	0.01 (0.01, 0.02)
	external hospital validation (Hospital 4)	0.01
	temporal validation (COVID-tested, any result, Hospitals 1-4)	0.00
	temporal validation (COVID+, Hospitals 1-4)	-0.08
MEWS > 4	train (Hospitals 1-3)	0.01 (0.01, 0.01)
	test (Hospitals 1-3)	0.01 (0.01, 0.02)
	external hospital validation (Hospital 4)	0.01
	temporal validation (COVID-tested, any result, Hospitals 1-4)	-0.01
	temporal validation (COVID+, Hospitals 1-4)	-0.08
MEWS > 5	train (Hospitals 1-3)	0.01 (0.01, 0.01)
	test (Hospitals 1-3)	0.01 (0.01, 0.02)
	external hospital validation (Hospital 4)	0.01
	temporal validation (COVID-tested, any result, Hospitals 1-4)	-0.01

	temporal validation (COVID+, Hospitals 1-4)	-0.08
PARFAIT	**train (Hospitals 1-3)**	0.27 (0.26, 0.29)
	test (Hospitals 1-3)	0.28 (0.26, 0.3)
	external hospital validation (Hospital 4)	0.17 (0.16, 0.17)
	temporal validation (COVID-tested, any result, Hospitals 1-4)	0.35 (0.33, 0.38)
	temporal validation (COVID+, Hospitals 1-4)	0.42 (0.39, 0.45)

Supplemental Figure 1. Voting example during an event horizon. The label is AdvRS. The 8 hour horizon is labeled by AdvRS horizon and implies the time period during which the event horizon is applicable. The prediction here refers to the thresholded prediction, where predictions over a binary threshold are labeled as ‘1’ (green) and predictions under a threshold are labeled as ‘0’ (red).





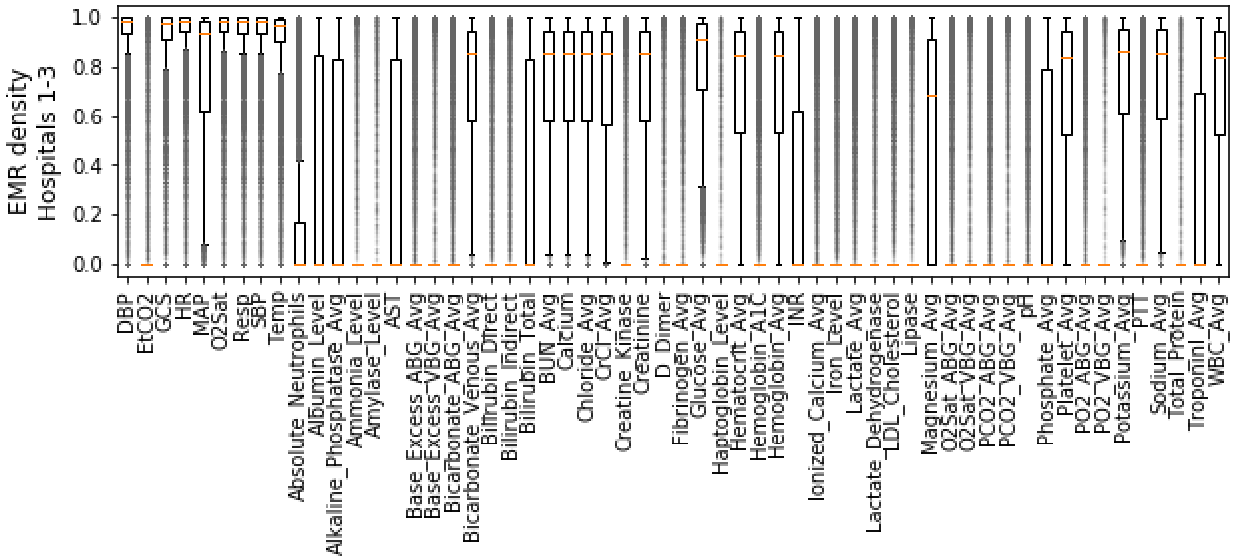
										1	1	1	AdvRS
0	0	1	1	1	1	1	1	1	1				AdvRS horizon: 8h
		0	1	1	0	1	0	1	0				AdvRS prediction

indecisive - FP/FN

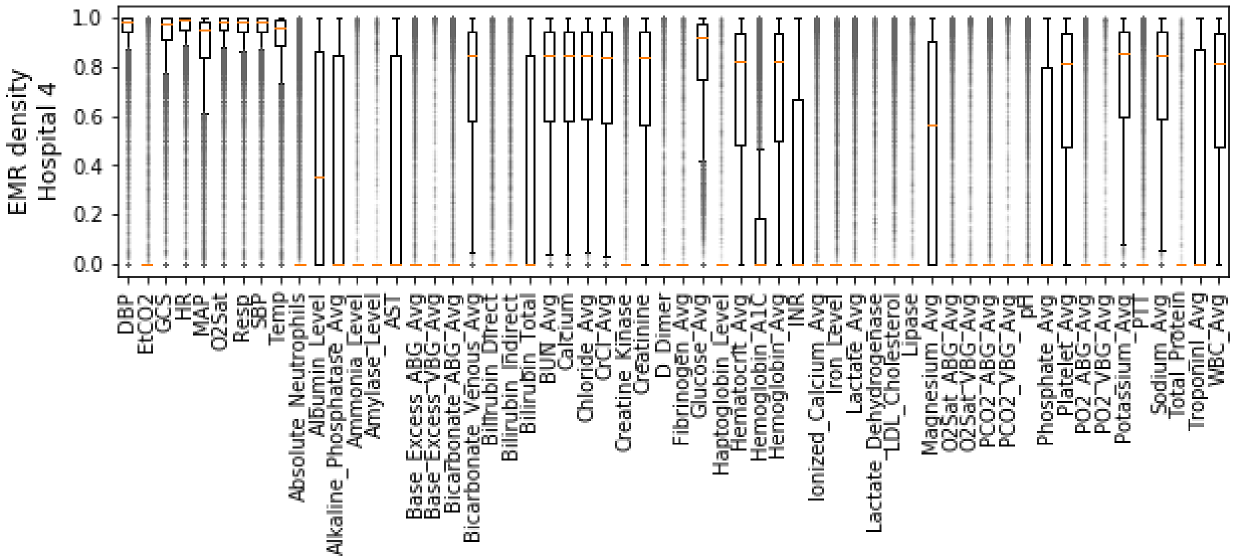
4 / 8 vote yes

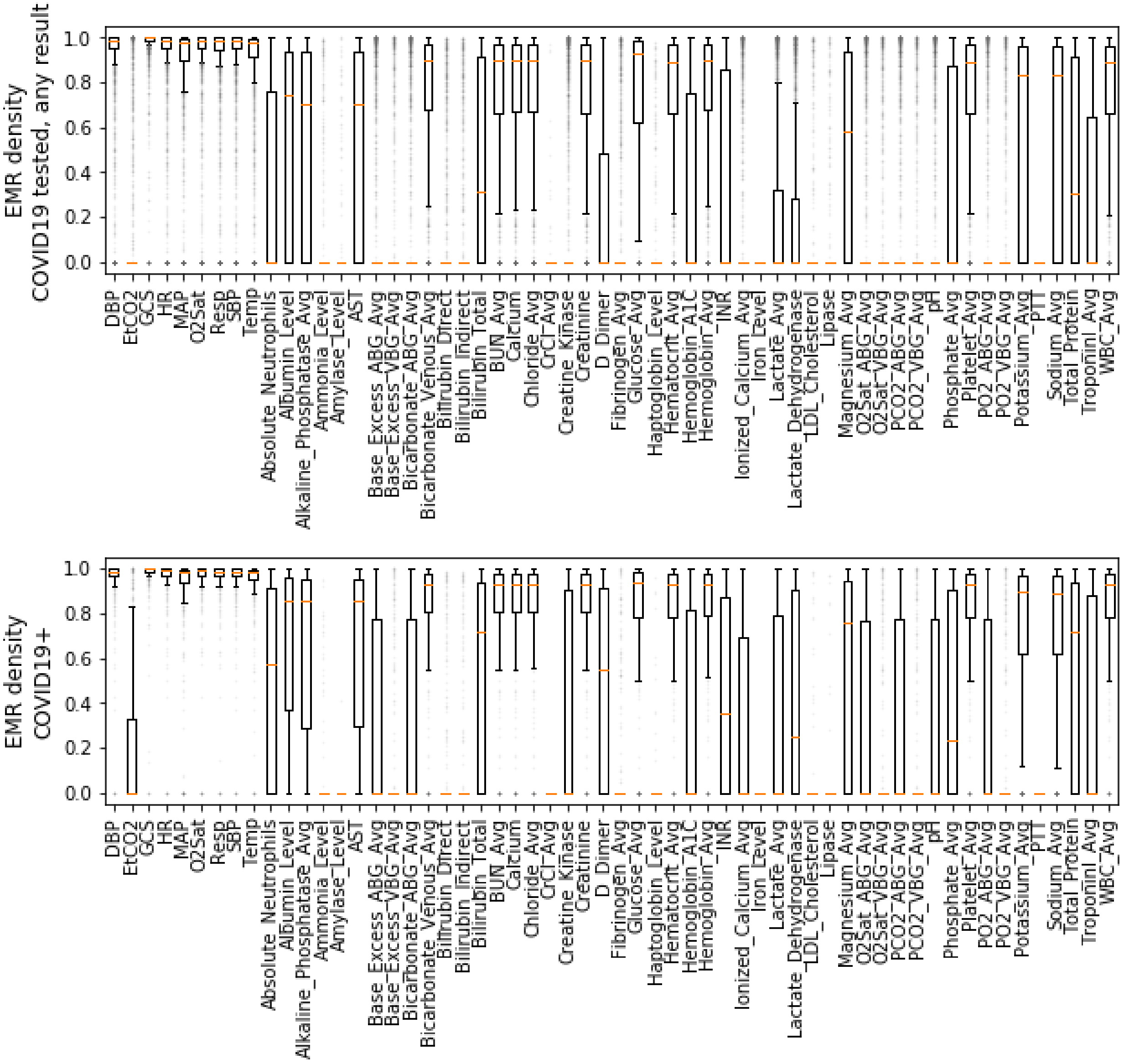
Supplemental Figure 2: Boxplot of EMR data density by variable for (2a) Hospitals 1-3, training set for crossvalidation, (2b) Hospital 4, external validation, (2c) temporal validation, COVID19 test, any result, (2d) temporal validation, COVID19+. No meaningful difference between AdvRS+ and AdvRS-.

(2a)



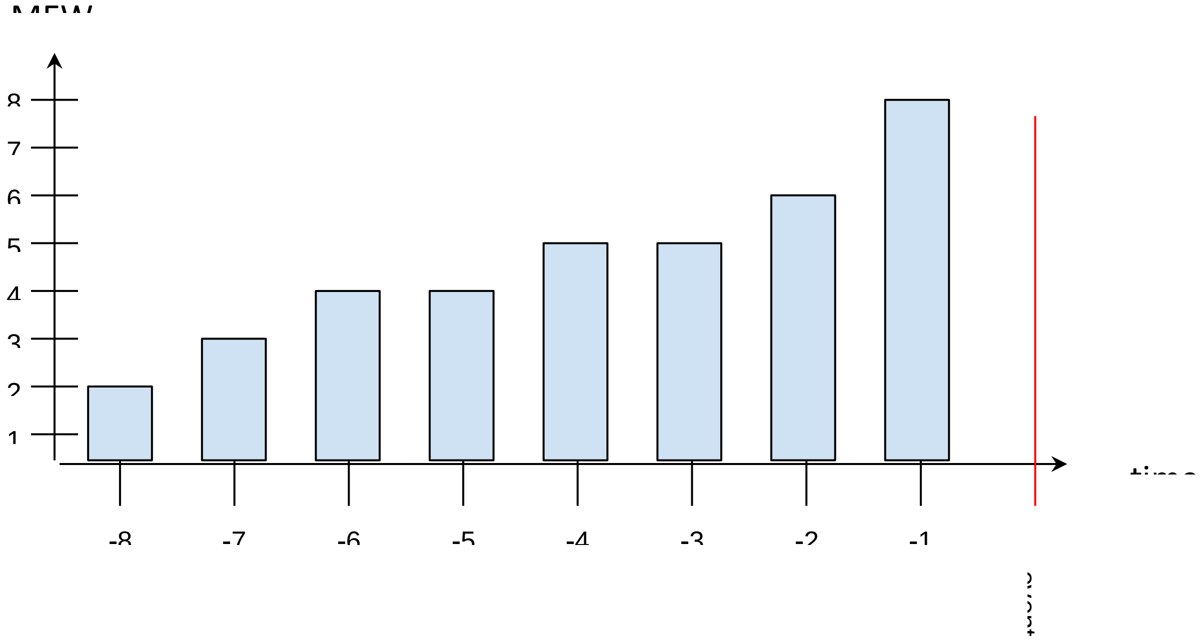
(2b)



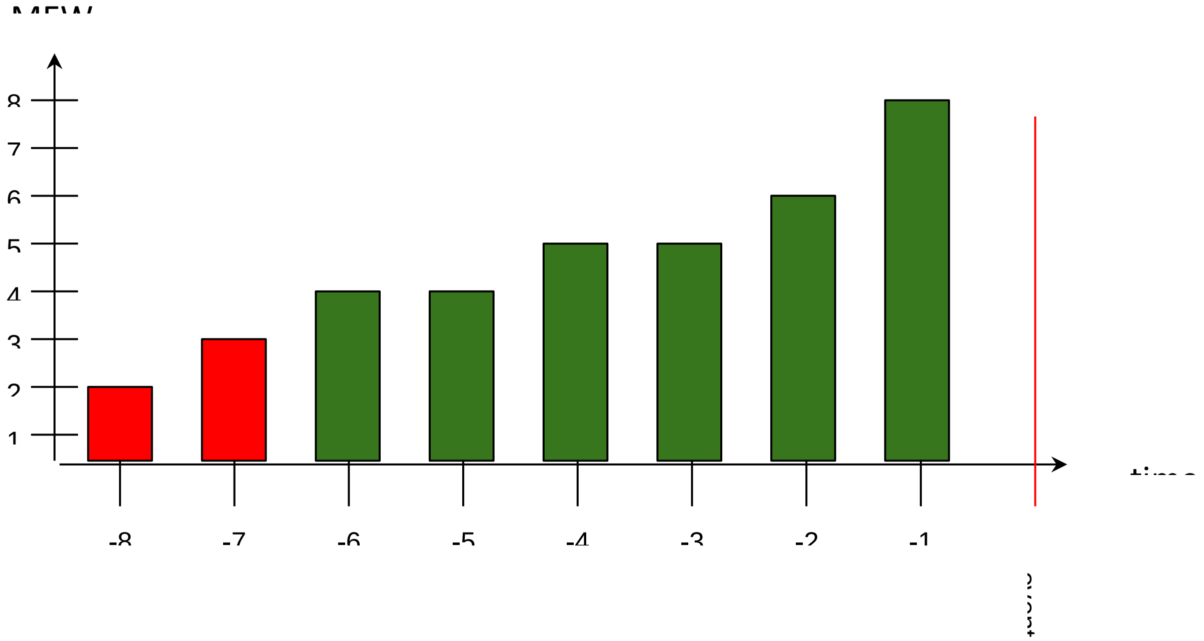

EMR density
COVID19 tested, any result
1.0
0.8
0.6
0.4
0.2
0.0
DBP
EtCO2
GCS
HR
MAP
O2Sat
Resp
SBP
Temp
Absolute_Neutrophils
Albumin_Level
Alkaline_Phosphatase_Avg
Ammonia_Level
Amylase_Level
AST
Base_Excess_ABG_Avg
Base_Excess_VBG_Avg
Bicarbonate_ABG_Avg
Bicarbonate_Venous_Avg
Bilirubin_Direct
Bilirubin_Indirect
Bilirubin_Total
BUN_Avg
Calcium
Chloride_Avg
CrCl_Avg
Creatine_Kinase
Creatinine
D_Dimer
Fibrinogen_Avg
Glucose_Avg
Haptoglobin_Level
Hematocrit_Avg
Hemoglobin_A1C
Hemoglobin_Avg
INR
Ionized_Calcium_Avg
Iron_Level
Lactate_Avg
Lactate_Dehydrogenase
LDL_Cholesterol
Lipase
Magnesium_Avg
O2Sat_ABG_Avg
O2Sat_VBG_Avg
PCO2_ABG_Avg
PCO2_VBG_Avg
pH
Phosphate_Avg
Platelet_Avg
PO2_ABG_Avg
PO2_VBG_Avg
Potassium_Avg
PTT
Sodium_Avg
Total_Protein
TroponinI_Avg
WBC_Avg
EMR density
COVID19+
1.0
0.8
0.6
0.4
0.2
0.0
DBP
EtCO2
GCS
HR
MAP
O2Sat
Resp
SBP
Temp
Absolute_Neutrophils
Albumin_Level
Alkaline_Phosphatase_Avg
Ammonia_Level
Amylase_Level
AST
Base_Excess_ABG_Avg
Base_Excess_VBG_Avg
Bicarbonate_ABG_Avg
Bicarbonate_Venous_Avg
Bilirubin_Direct
Bilirubin_Indirect
Bilirubin_Total
BUN_Avg
Calcium
Chloride_Avg
CrCl_Avg
Creatine_Kinase
Creatinine
D_Dimer
Fibrinogen_Avg
Glucose_Avg
Haptoglobin_Level
Hematocrit_Avg
Hemoglobin_A1C
Hemoglobin_Avg
INR
Ionized_Calcium_Avg
Iron_Level
Lactate_Avg
Lactate_Dehydrogenase
LDL_Cholesterol
Lipase
Magnesium_Avg
O2Sat_ABG_Avg
O2Sat_VBG_Avg
PCO2_ABG_Avg
PCO2_VBG_Avg
pH
Phosphate_Avg
Platelet_Avg
PO2_ABG_Avg
PO2_VBG_Avg
Potassium_Avg
PTT
Sodium_Avg
Total_Protein
TroponinI_Avg
WBC_Avg

Supplemental Figure 3. Example scoring method for MEWS. Bars in green are considered positive by the threshold. Bars in red are considered negative by the threshold.

3a. MEWS scores for 8 hours prior to an event. Median MEWS value is 4.5/14.

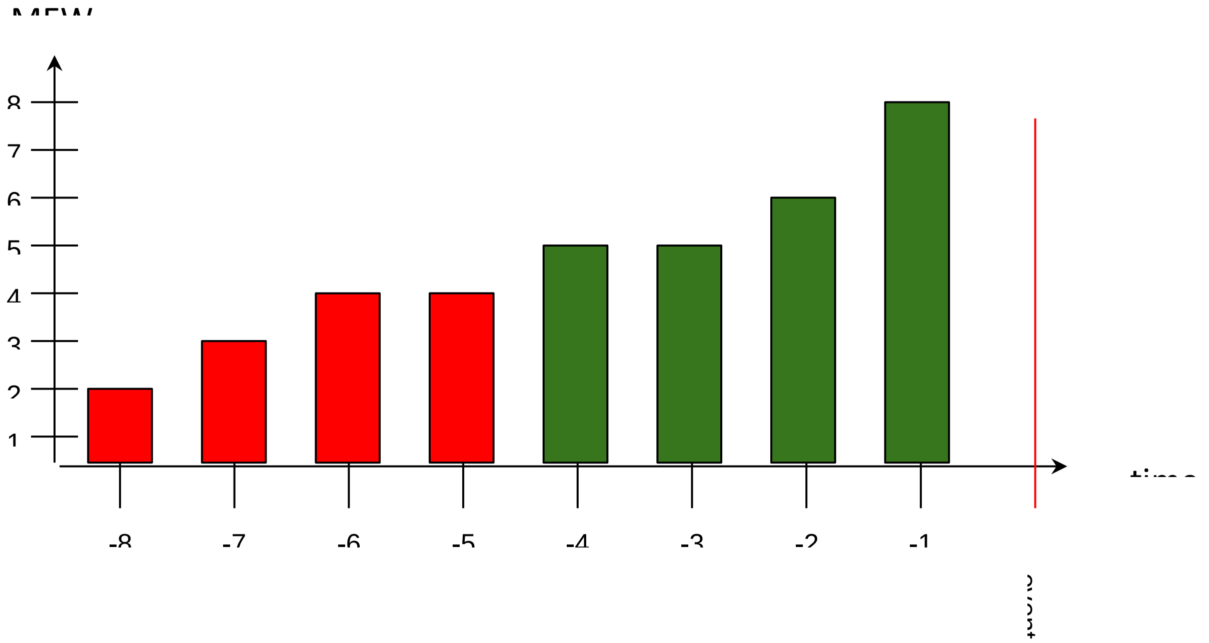


3b. MEWS > 3. All values ≤ 3 are considered negative. Median of the predicted positive MEWS scores is 5/14. The median of the predicted negative MEWS scores is 2.5/14. As the majority (6/8) predictions are positive, this is classed as a ‘true positive’ with median score 5/14 = 0.357.

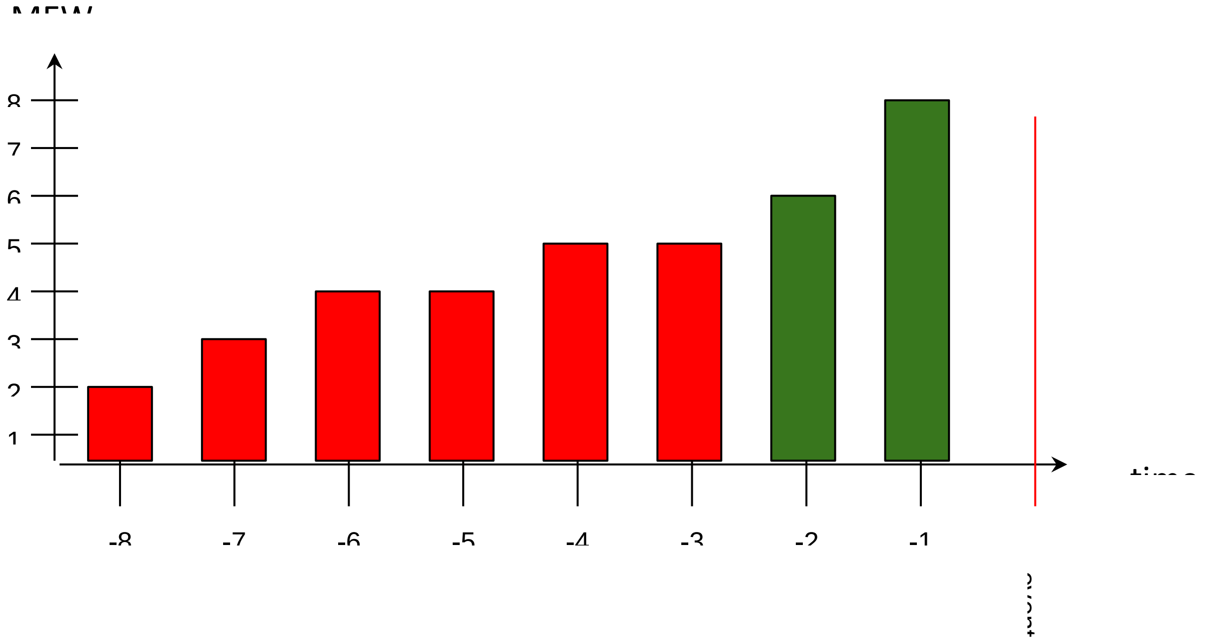


3c. MEWS > 4. All values ≤ 4 are considered negative. Median of the predicted positive MEWS scores is

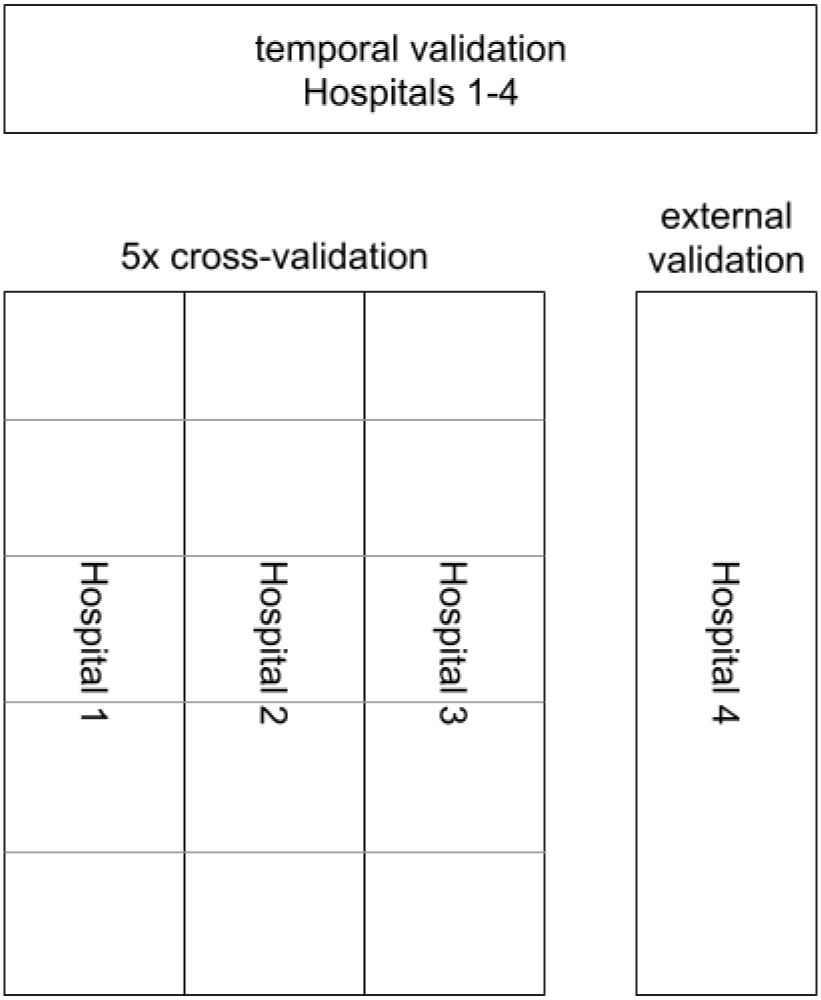
5.5/14. The median of the predicted negative MEWS scores is 3.5/14. As per criteria, a situation with an even number of ‘predicted positive’ and ‘predicted negative’ is classed as a false negative, with median score 3.5/14 = 0.250.



3d. MEWS > 5. All values ≤ 5 are considered negative. Median of the predicted positive MEWS scores is 7/14. The median of the predicted negative MEWS scores is 4/14. As the number of predicted negative MEWS scores (6/8) is greater than the number of positive scores, this is classed as a false negative case with a median score of 4/14 = 0.286.



Supplemental Figure 4. Data organizational schema for ARF modeling, including training, external validation, and temporal validation.


temporal validation
Hospitals 1-4
5x cross-validation
external
validation
Hospital 1
Hospital 2
Hospital 3
Hospital 4