# A Comprehensive Benchmark for COVID-19 Predictive Modeling Using Electronic Health Records in Intensive Care

**Junyi Gao***, Yinghao Zhu*, Wenqing Wang*, Guiying Dong, Wen Tang, Hao Wang, Yasha Wang, Ewen M. Harrison, Liantao Ma

THE UNIVERSITY of EDINBURGH | usher institute

## Motivation

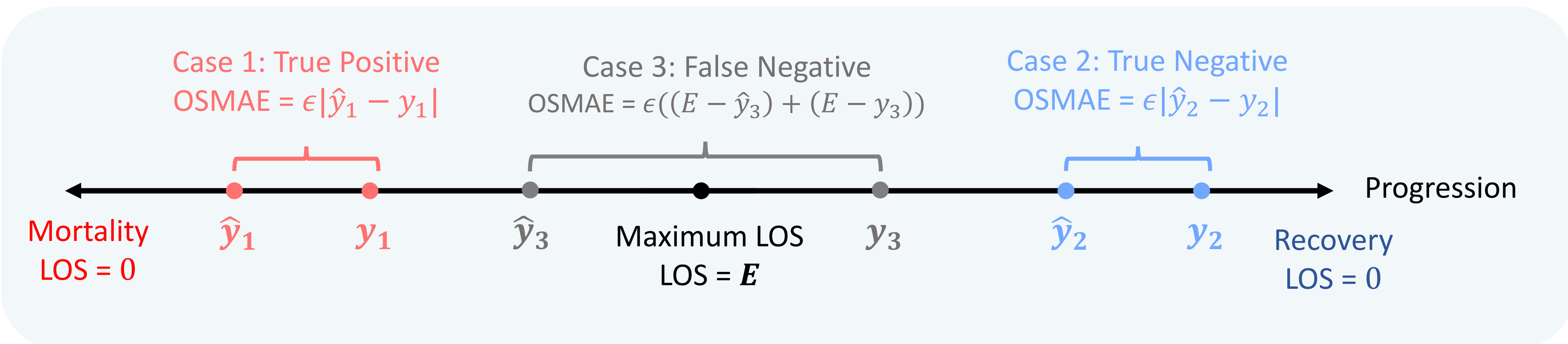
- How to adapt traditional prediction tasks for COVID-19 patients in intensive care units?
- How to choose the best model among various options of predictive models?

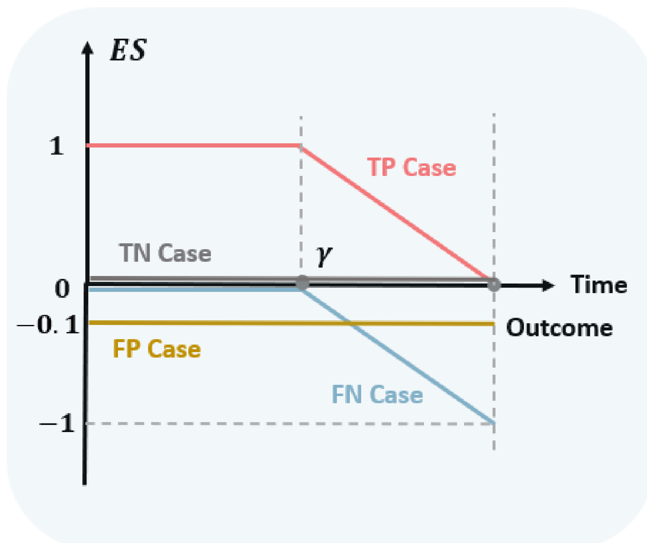
**Objective:** Build a **fair**, **comprehensive** and **open-source** benchmark for COVID-19 prediction models in ICU.

## Prediction Tasks

- **Task 1 (Early mortality prediction):** Predict the mortality risk at the early stage.
- **Task 2 (Outcome-specific length-of-stay prediction):** Predict the outcome and length-of-stay simultaneously.

**Evaluation Metrics:**

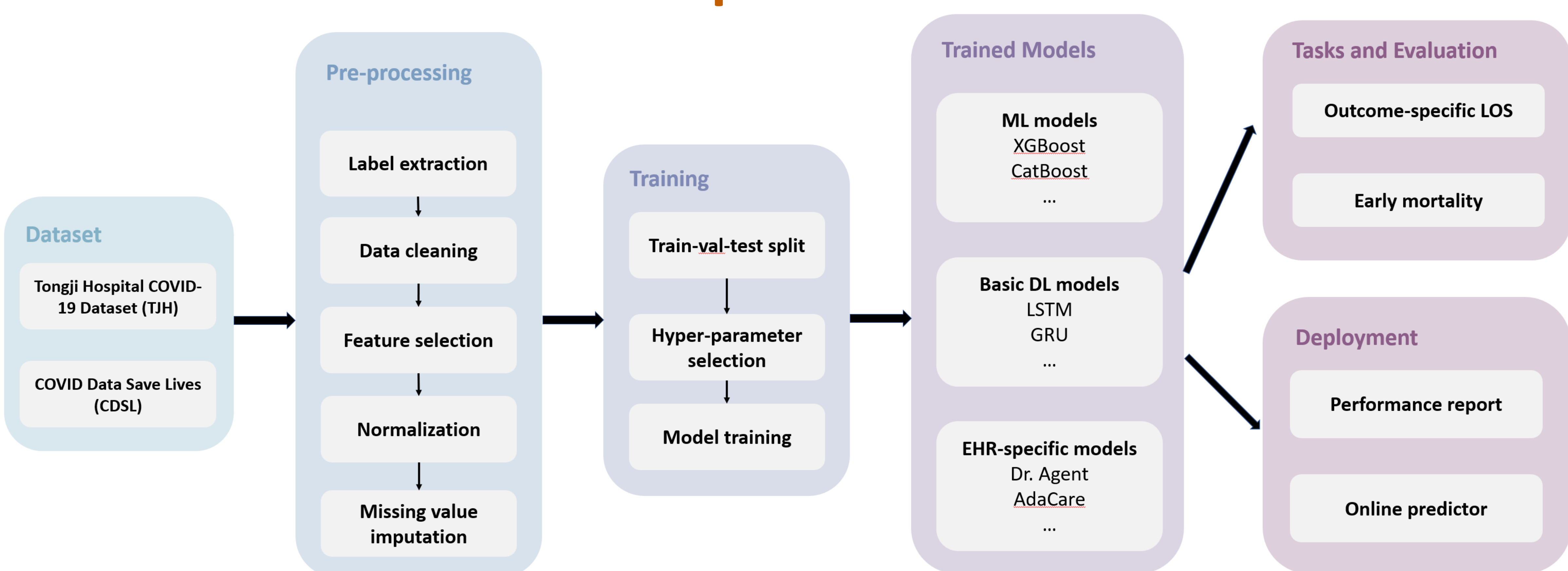



**1 Outcome-specific mean absolute error (OSMAE)**

**2 Early prediction score (ES)**

## Pipeline




## Datasets and Experiment Settings

- **Dataset 1 (Tongji Hospital COVID-19 Dataset (TJH)):**
  - 361 patients, 1,338 records, 45.98% (166) mortality rate
- **Dataset 2 (Covid Data Save Lives (CDSL)):**
  - 4,255 patients, 42,204 records, 45.98% (540) mortality rate

- 18 baseline models
- **1 Clinical scoring model:** 4C mortality score [1]
- **5 Machine learning models:** Decision tree (DT), Random forest (RF), ...
- **6 Basic deep learning models:** MLP, RNN, Transformer [2], …
- **6 EHR-specific predictive models:** RETAIN [3], AdaCare [4], …

## Results

### Early mortality prediction (part):

| Dataset | TJH | | | CDSL | | |
|---|---|---|---|---|---|---|
| Metric | AUPRC(↑) | AUROC(↑) | ES(↑) | AUPRC(↑) | AUROC(↑) | ES(↑) |
| 4C | 89.84 ± 4.46 | 94.16 ± 2.57 | - | 23.93 ± 2.99 | 76.15 ± 4.06 | - |
| XGBoost | 95.70 ± 2.98 | 96.84 ± 2.09 | 76.41 ± 9.90 | 49.70 ± 5.06 | 84.59 ± 3.09 | -3.12 ± 4.44 |
| RNN | 96.03 ± 3.42 | 97.41 ± 2.19 | 78.33 ± 11.35 | 57.57 ± 6.55 | 87.81 ± 2.55 | 19.66 ± 8.86 |
| RNN-TA | 95.97 ± 3.65 | 97.38 ± 2.29 | 79.79 ± 11.06 | **58.21 ± 6.34** | 88.01 ± 2.36 | 20.15 ± 11.25 |
| Transformer | 93.47 ± 6.73 | 96.86 ± 2.13 | 79.21 ± 13.05 | 38.34 ± 5.07 | 81.32 ± 3.46 | -3.96 ± 13.23 |
| Transformer-TA | 93.86 ± 6.42 | 97.01 ± 2.01 | 80.73 ± 12.47 | 40.18 ± 5.00 | 82.36 ± 3.61 | -0.78 ± 11.42 |
| StageNet | 95.71 ± 3.77 | 97.27 ± 2.38 | 77.44 ± 12.52 | 56.57 ± 6.82 | 87.09 ± 3.54 | 23.93 ± 9.26 |
| StageNet-TA | 95.79 ± 3.86 | 97.32 ± 2.41 | 78.88 ± 12.29 | 58.19 ± 6.48 | **88.18 ± 2.82** | **24.55 ± 9.32** |
| AdaCare | 97.86 ± 1.09 | 98.53 ± 0.77 | 78.39 ± 7.51 | 55.32 ± 4.45 | 86.59 ± 1.99 | 17.46 ± 8.66 |
| AdaCare-TA | **98.11 ± 1.13** | **98.64 ± 0.89** | **84.51 ± 8.17** | 55.62 ± 5.44 | 87.00 ± 2.18 | 20.09 ± 9.86 |

* -TA denotes the model trained with time-aware loss

### Outcome-specific LOS prediction (part):

| Dataset | TJH | | | CDSL | | |
|---|---|---|---|---|---|---|
| Metric | MAE (↓) | MSE (↓) | OSMAE (↓) | MAE (↓) | MSE (↓) | OSMAE (↓) |
| $RF_t$ | 4.83 ± 0.53 | 40.94 ± 8.77 | 6.14 ± 0.99 | 4.05 ± 0.13 | 31.30 ± 4.42 | 4.90 ± 0.16 |
| $XGBoost_t$ | 4.78 ± 0.55 | 41.76 ± 9.47 | 5.77 ± 0.92 | 4.02 ± 0.13 | 31.10 ± 4.42 | 4.81 ± 0.15 |
| $GRU_t$ | **4.33 ± 0.49** | **35.10 ± 9.61** | 5.16 ± 1.31 | 3.75 ± 0.18 | 28.90 ± 4.68 | 4.34 ± 0.32 |
| $GRU_m$ | 4.80 ± 0.48 | 41.78 ± 9.49 | 5.38 ± 1.09 | 3.98 ± 0.18 | 32.14 ± 4.31 | 4.55 ± 0.22 |
| $Transformer_t$ | 5.04 ± 0.43 | 43.88 ± 7.22 | 5.96 ± 1.62 | 3.98 ± 0.20 | 32.35 ± 5.61 | 5.19 ± 0.19 |
| $Transformer_m$ | 5.06 ± 0.46 | 46.02 ± 11.63 | 6.61 ± 1.57 | 4.00 ± 0.20 | 32.02 ± 4.68 | 5.13 ± 0.19 |
| $StageNet_t$ | 4.60 ± 0.76 | 41.70 ± 14.21 | 5.59 ± 1.33 | **3.72** ± 0.15 | **28.81 ± 4.57** | 4.33 ± 0.22 |
| $StageNet_m$ | 4.49 ± 0.42 | 39.55 ± 8.52 | 7.10 ± 1.74 | 3.78 ± 0.18 | 29.93 ± 4.50 | **4.28 ± 0.19** |
| $Dr.Agent_t$ | 4.61 ± 0.58 | 40.85 ± 12.12 | 4.93 ± 1.14 | 3.80 ± 0.18 | 29.70 ± 4.96 | 4.46 ± 0.27 |
| $Dr.Agent_m$ | 4.41 ± 0.58 | 37.55 ± 11.38 | **4.75 ± 1.11** | 3.82 ± 0.19 | 29.45 ± 4.70 | 4.49 ± 0.28 |

* *m* denotes the model trained with multi-task setting, *t* denotes two-stage

Paper

Code

Online Platform

[1] Knight, S. R., Ho, A., Pius, R., Buchan, I., Carson, G., Drake, T. M., ... & Harrison, E. M. (2020). Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. *bmj*, *370*.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[3] Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. (2016). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, *29*.

[4] Ma, L., Gao, J., Wang, Y., Zhang, C., ... & Ma, X. (2020, April). Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 01, pp. 825-832).