Feature Robustness in Non-stationary Health Records:

Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks

Bret Nestor[*,0,1], Matthew B. A. McDermott[*,2], Willie Boag[2], Gabriela Berner[3], Tristan Naumann[4], Michael C. Hughes[5], Anna Goldenberg[0,1,6], Marzyeh Ghassemi[0,1]

August 9, 2019

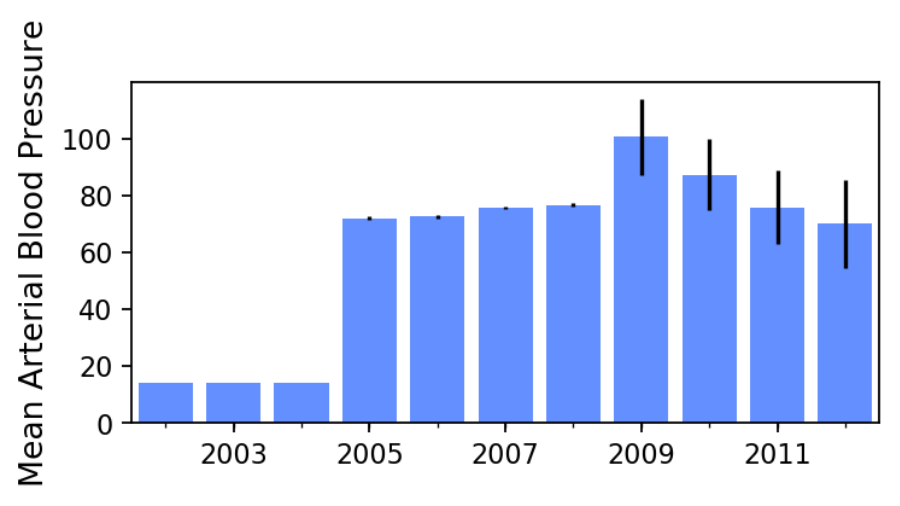
[*]Equal Contribution [0] University of Toronto, [1] Vector Institute, [2] Massachusetts Institute of Technology, [3]Harvard University, [4]Microsoft Research, [5]Tufts University, [6]Hospital for Sick Children

Illustration of Concept Drift in Clinical Practice

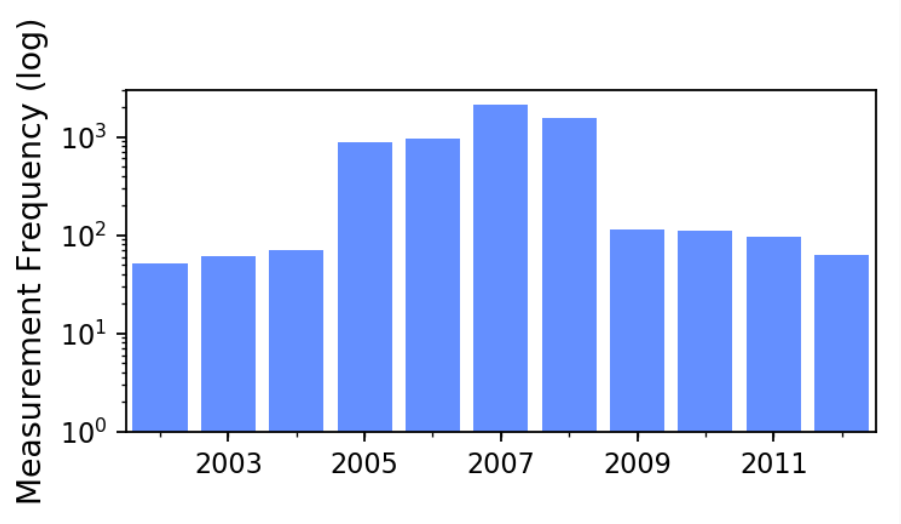
There are two worrisome effects of concept drift.

Values of the collected data change

(Underlying physiology of humans does not)

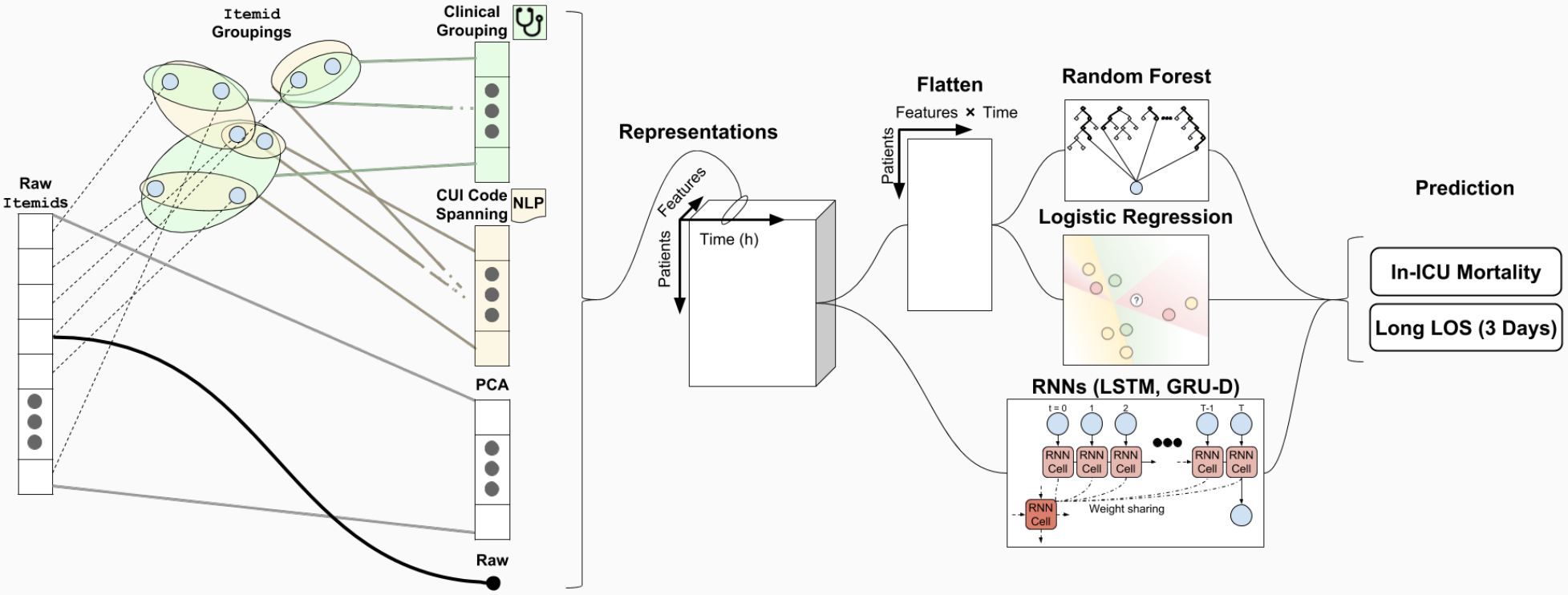


Frequency of data collection changes



Representation and Generalisability

We test the effect of representation on the longevity of model performance across multiple models.



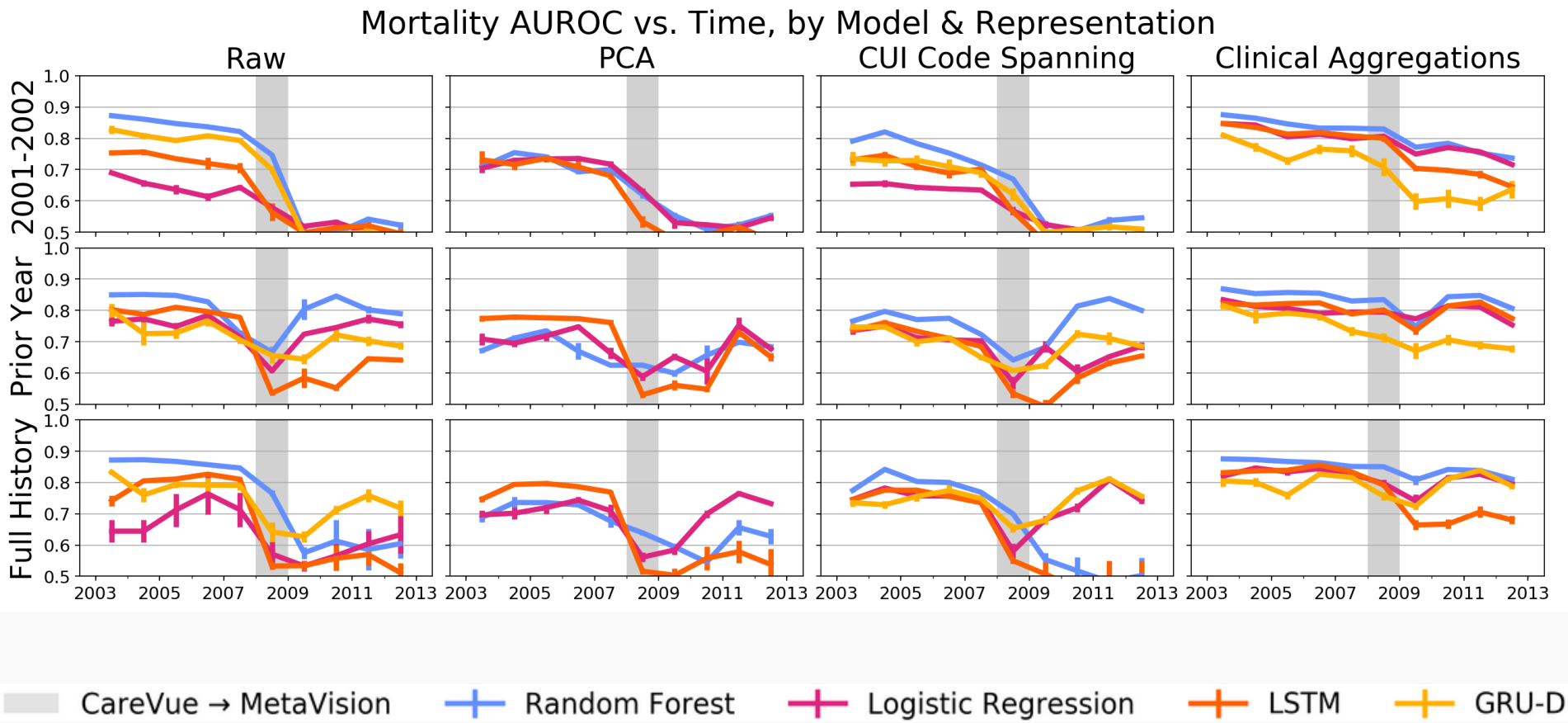
Task 1: In-ICU Mortality classification with Year-Agnostic Training

Table 1: Year agnostic model trained to predict in-ICU mortality with randomised CV splits based on the first 24 hours of the ICU stay.

Model	Average AUROC for Random Splits			
	Raw	PCA	CUI Code Spanning	Clinical
LR	71.30 ± 1.70	78.65 ± 1.49	68.37 ± 0.98	84.96 ± 1.26
RF	81.87 ± 2.21	77.01 ± 2.81	79.42 ± 1.90	85.87 ± 2.07
LSTM	70.15 ± 2.53	75.03 ± 0.81	68.45 ± 2.52	83.69 ± 0.90
GRUD	81.43 ± 3.59	-	79.84 ± 1.38	82.67 ± 2.40

Task 1: In ICU Mortality classification with Feasible Training

Below are the model performances when trained with feasible training regimes.



Feature Robustness in Non-stationary Health Records

Code is available at:

`https://github.com/MLforHealth/MIMIC_Generalisation`