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]

*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

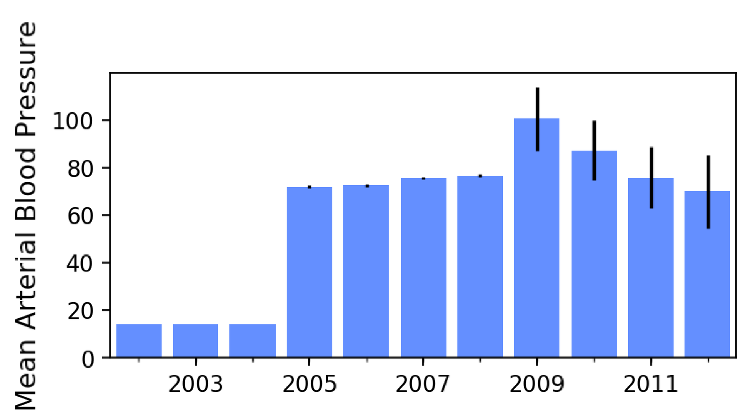
Is Machine Learning Resilient to Clinical Practice Change?

Models trained on de-identified, date-obscured data may not endure as care practice evolves

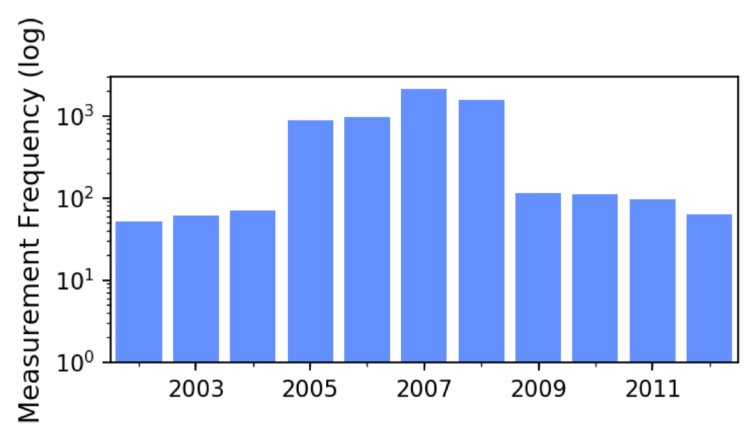
- De-identification neglects concept drift
- Adaptive computation with explicit control over tradeoff between speed and numerical precision.

Illustration of Concept Drift in Clinical Practice

Values of the collected data changes
(Underlying physiology of humans does not)

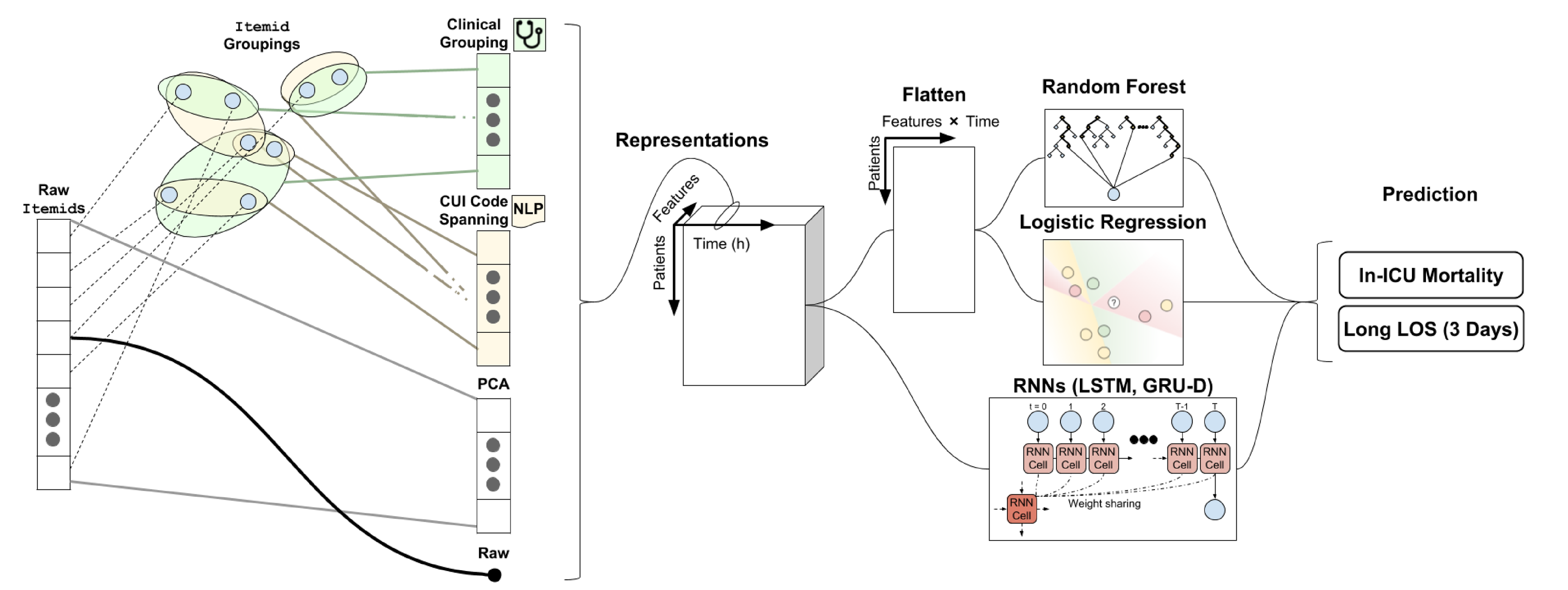


Frequency of data collection changes

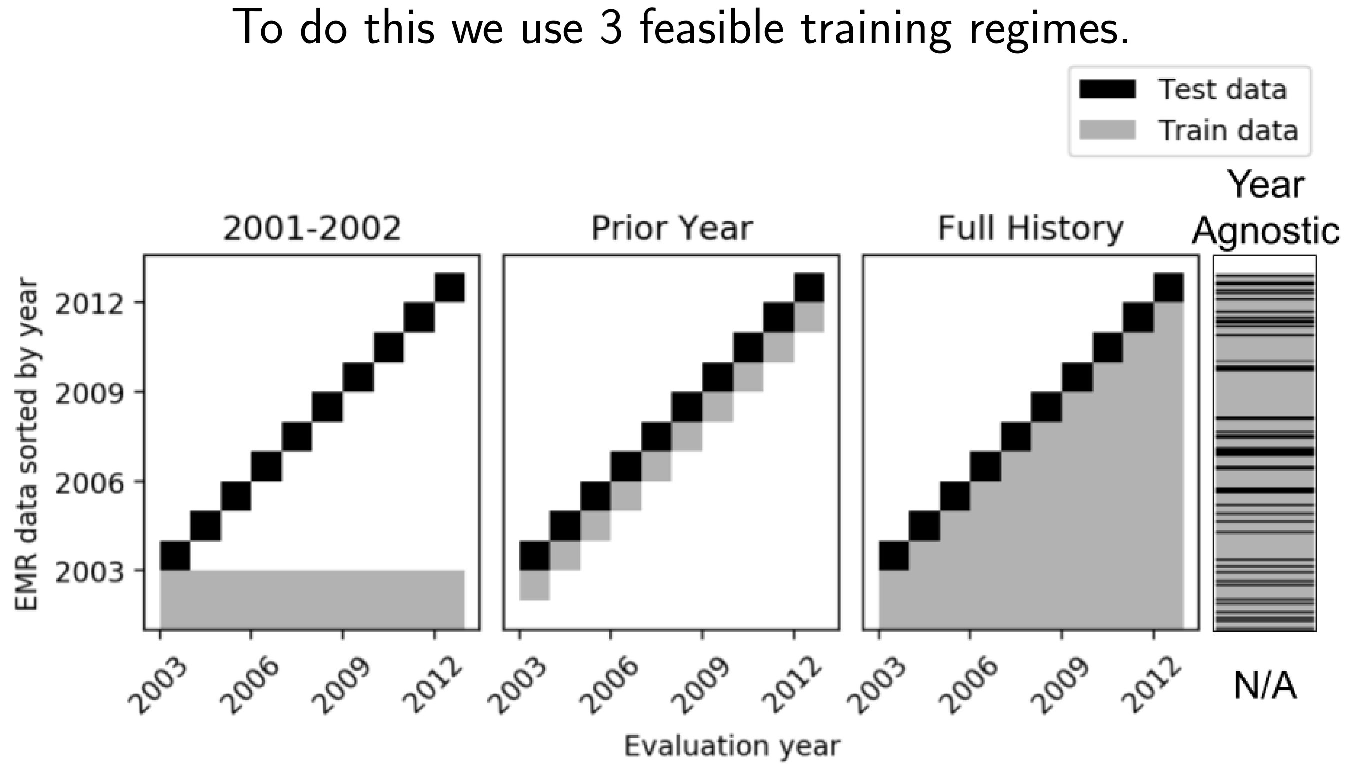


Experiments

We established a standard pipeline that selects a representation then trains any model on a classification task.



We train these models **only** using retrospective data and test on prospective data. To do this we use 3 feasible training regimes.



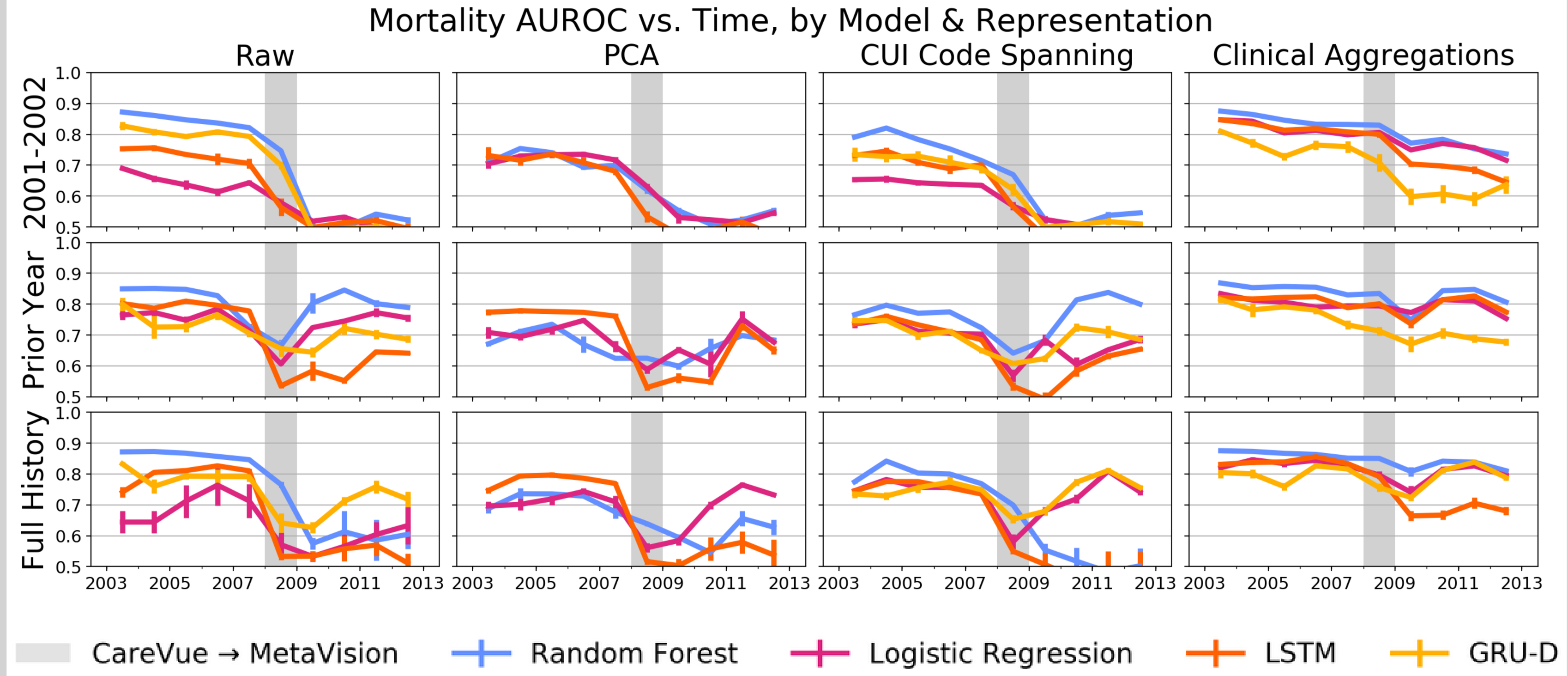
Model Performance Under Practical Training Regimes

Task 1: In ICU Mortality

First, we show the performance on models trained without knowledge of the years (randomised CV splits).

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

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

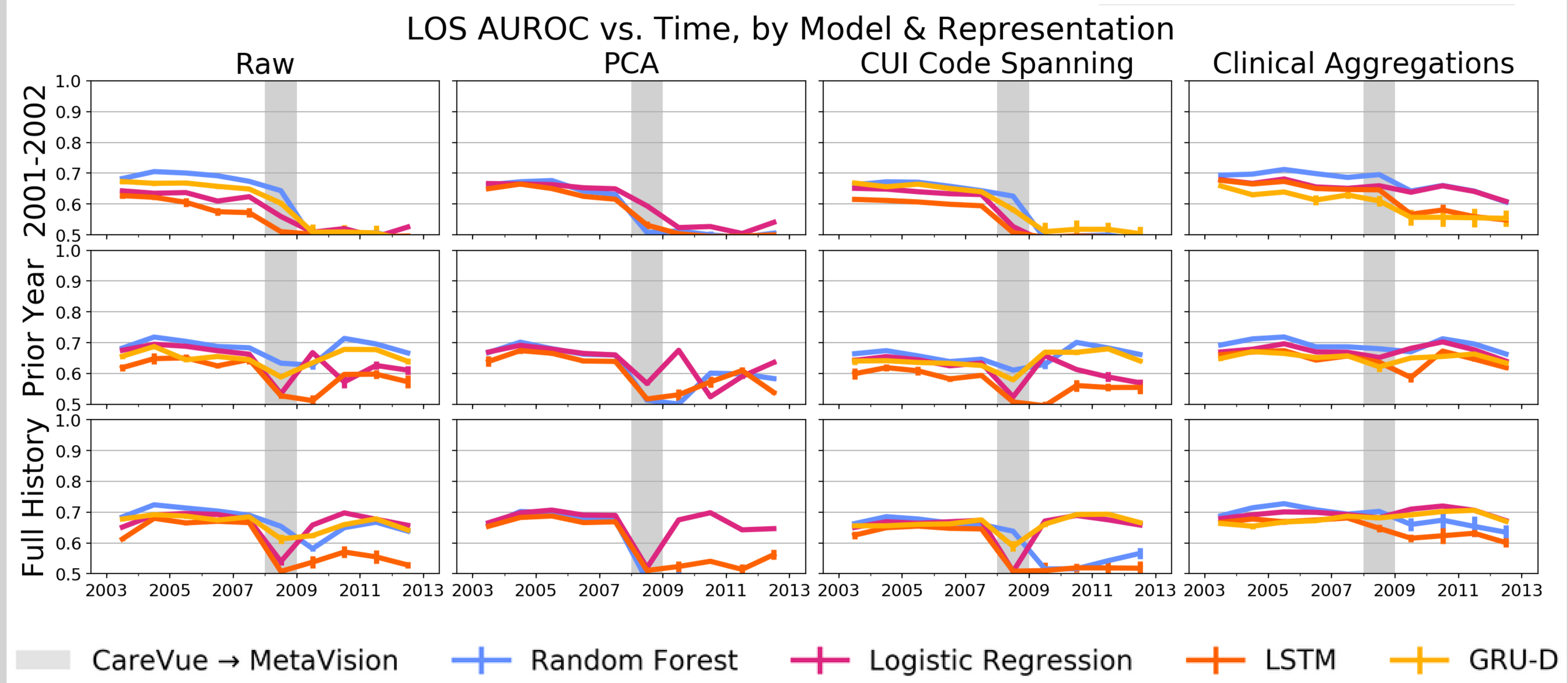


Task 2: Length of Stay Greater Than 3 Days (Classification)

First, we show the performance on models trained without knowledge of the years (5-2 randomised CV splits).

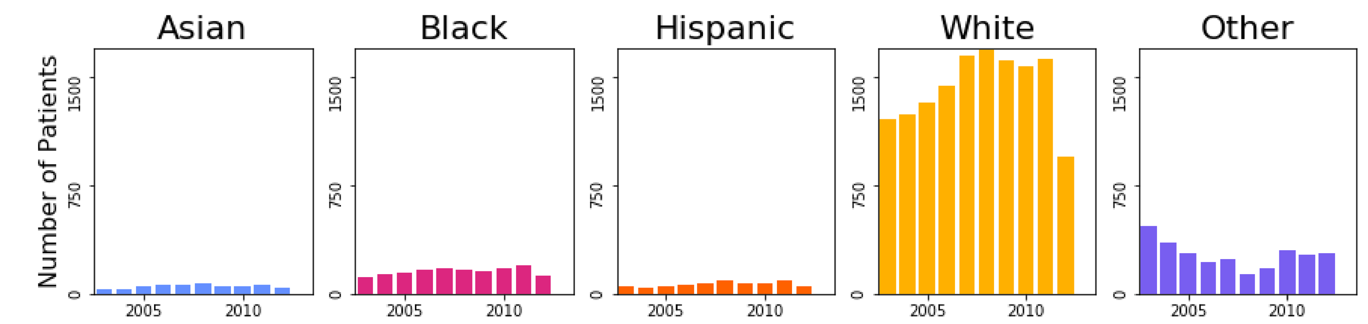
Model	Average AUROC			
	Raw	PCA	CUI Code Spanning	Clinical
LR	67.36 ± 1.91	68.37 ± 0.93	67.99 ± 0.61	70.47 ± 0.94
RF	69.89 ± 0.44	67.52 ± 0.60	66.83 ± 1.13	71.03 ± 0.72
LSTM	64.87 ± 1.09	61.86 ± 2.25	62.67 ± 1.90	68.75 ± 1.41
GRUD	68.95 ± 1.48	-	67.48 ± 0.87	69.89 ± 0.40

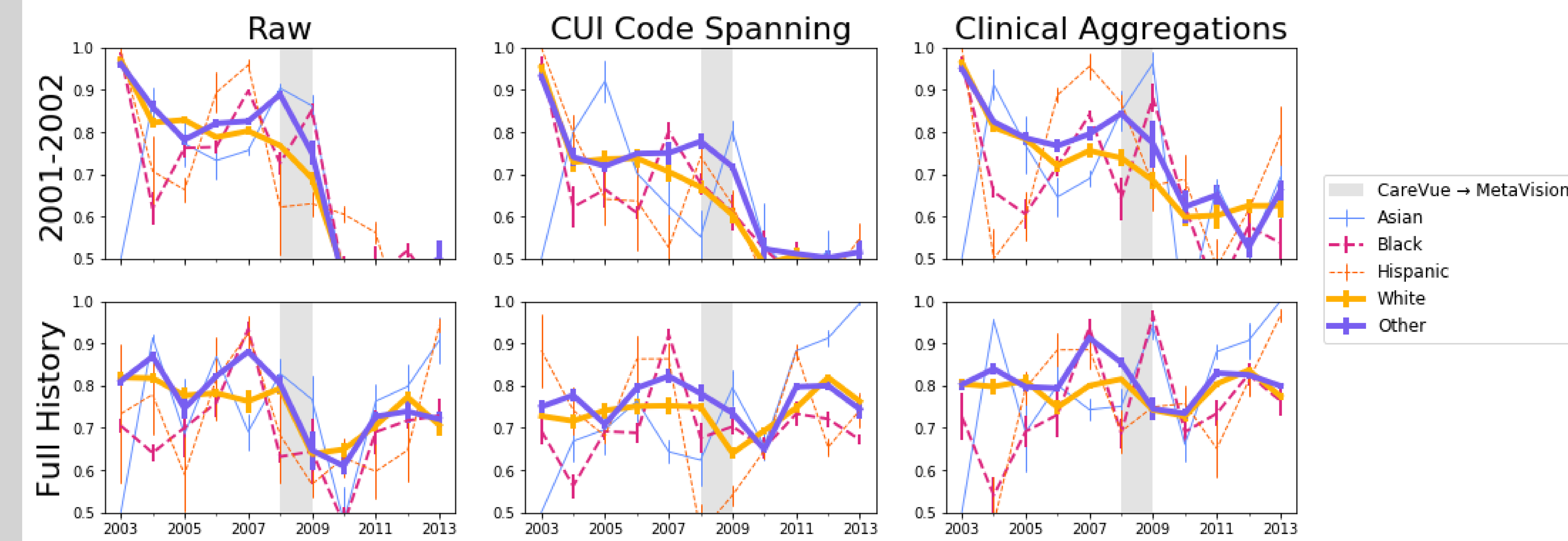
Below are the model performances when trained with feasible training regimes.



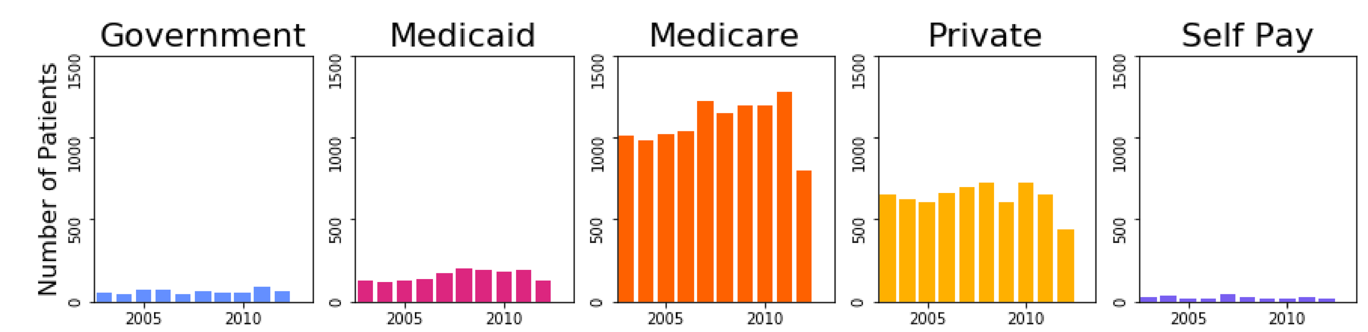
Do Models Deteriorate Faster for Underrepresented Groups?

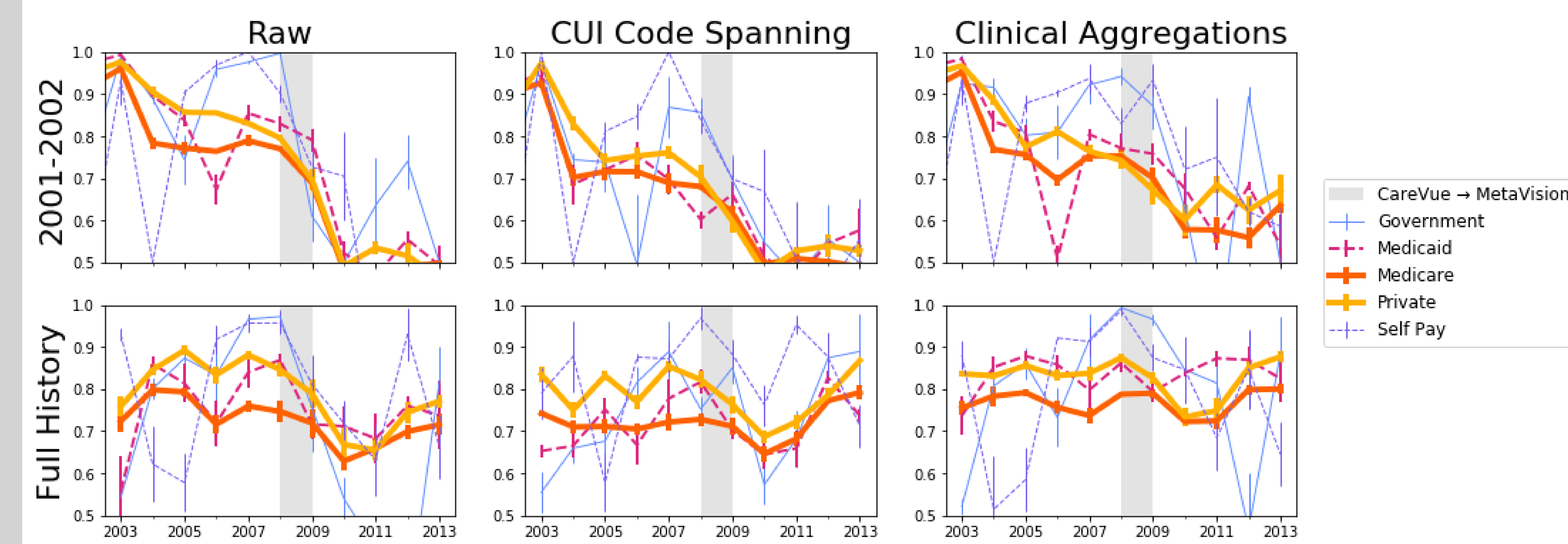
Distributions of ethnicity in MIMICIII by year.





Distribution of insurance types in MIMICIII by year.





Background

References

Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." Scientific data 3 (2016): 160035.

Che, Zhengping, et al. "Recurrent neural networks for multivariate time series with missing values." Scientific reports 8.1 (2018): 6085.

Resources

https://github.com/MLforHealth/MIMIC_Generalisation

https://arxiv.org/pdf/1908.00690.pdf