


Hands on with MIMIC-III


ADOPTED FROM IRENE CHEN (MIT)



What is MIMIC-III?

- Largest open dataset for clinical healthcare (for authorized researchers)
 - Dataset of 26 tables (e.g. admissions, patients)
 - Maintained by Roger Mark's lab at MIT
- 

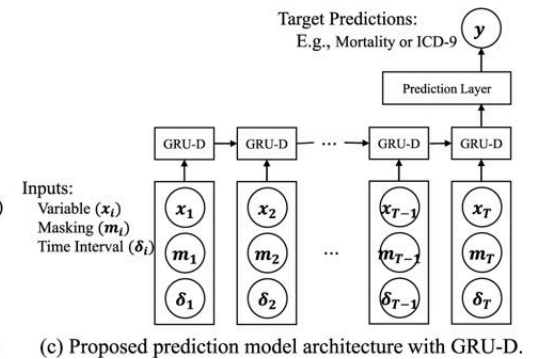
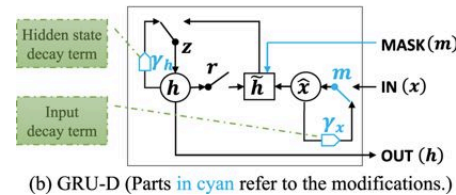
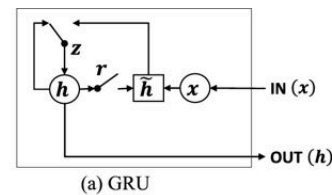
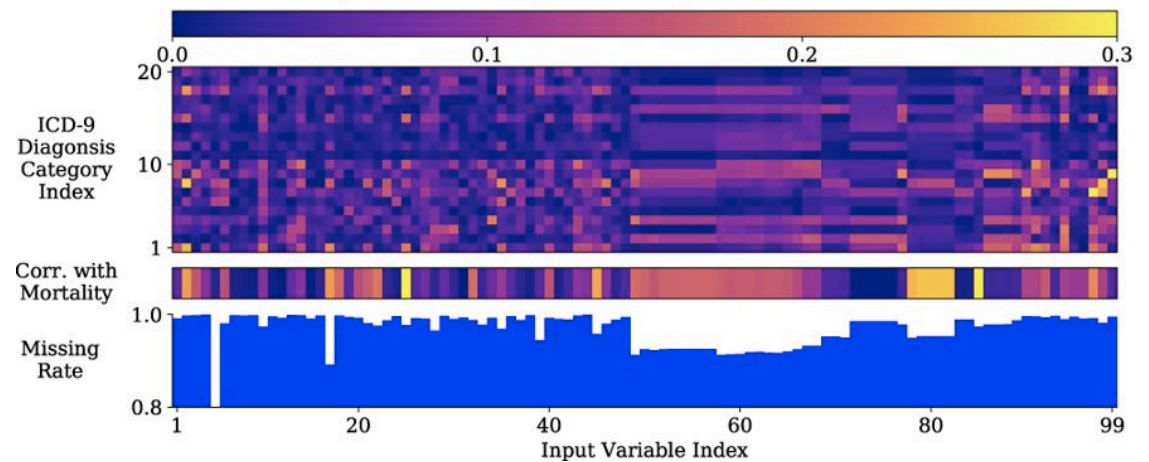
Today we'll get to know MIMIC better

1. Examples of recent academic papers (all from MIT!) using MIMIC-III data
 2. Live coding to compute statistics from MIMIC-III
 3. Walk through logistic regression
- 

Recurrent Neural Networks for Multivariate Time Series with Missing Values

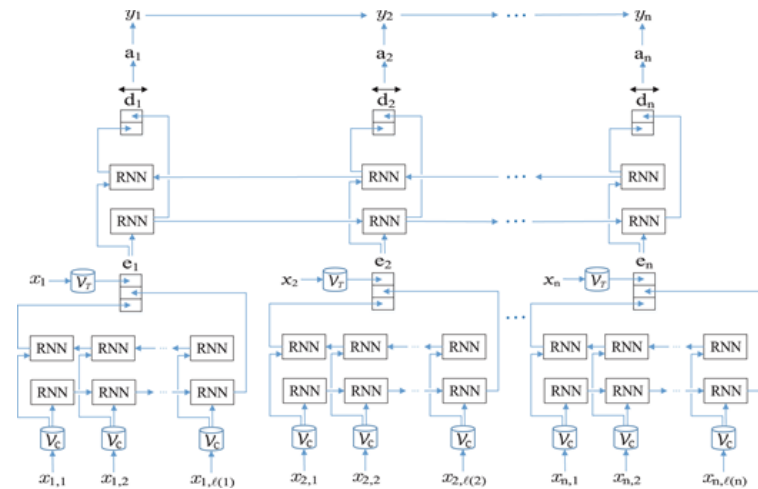
Che et al, 2018 (Nature Scientific Reports)

Absolute Values of Pearson Correlations between Variable Missing Rates and Labels (Mortality and ICD-9 Diagnosis Categories on MIMIC-III Dataset)



De-identification of patient notes with recurrent neural network

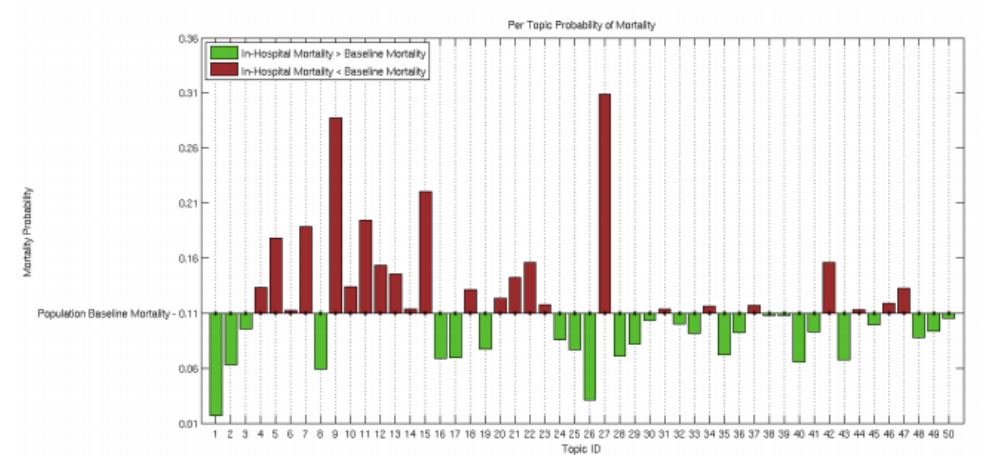
Dernoncourt et al, 2017 (JAMIA)



Model	i2b2			MIMIC		
	Precision	Recall	F1	Precision	Recall	F1
Nottingham	99.000	96.400	97.680	-	-	-
MIST	91.445	92.745	92.090	95.867	98.346	97.091
CRF	98.560	96.528	97.533	99.060	98.987	99.023
ANN	98.320	97.380	97.848	99.208	99.251	99.229
CRF + ANN	97.920	97.835	97.877	98.820	99.398	99.108

Unfolding Physiological State: Mortality Modeling in Intensive Care Units

Ghassemi et al, 2014 (KDD)

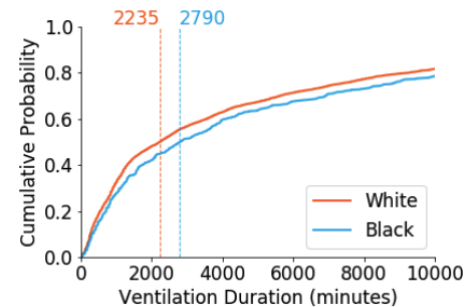


	Topic	Top Ten Words	Possible Topic
In-hospital Mortality	27	name, family, neuro, care, noted, status, plan, stitle, dr, remains	Discussion of end-of-life care
	15	intubated, vent, ett, secretions, propofol, abg, respiratory, resp, care, sedated	Respiratory failure
	7	thick, secretions, vent, trach, resp, tf, tube, coarse, cont, suctioned	Respiratory infection
	5	liver, renal, hepatic, ascites, dialysis, failure, flow, transplant, portal, ultrasound	Renal Failure

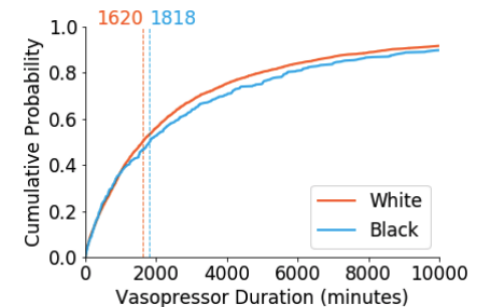
Racial Disparities and Mistrust in End-of-Life Care

Boag et al, 2018 (MLHC)

Social: Pt **refused to sign ICU consent** and expressed wishes to be DNR/DNI, seemingly **very frustrated** and **mistrusting of healthcare system** in relation to [REDACTED]. Also, w/ hx of **poor medication compliance and follow-up**



(a) CDF of ventilation duration by race ($p = .005$).



(b) CDF of vasopressor duration by race ($p = 0.12$).

Reproducibility in critical care: a mortality prediction case study

Johnson et al, 2017 (MLHC)

We reproduced datasets for 38 experiments corresponding to 28 published studies using MIMIC. In half of the experiments, the sample size we acquired was 25% greater or smaller than the sample size reported. The highest discrepancy was 11,767 patients. While accurate reproduction of each study cannot be guaranteed, we believe that these results highlight the need for more consistent reporting of model design and methodology to allow performance improvements to be compared. We discuss the challenges in reproducing the cohorts used in the studies, highlighting the importance of clearly reported methods (e.g. data cleansing, variable selection, cohort selection) and the need for open code and publicly available benchmarks.

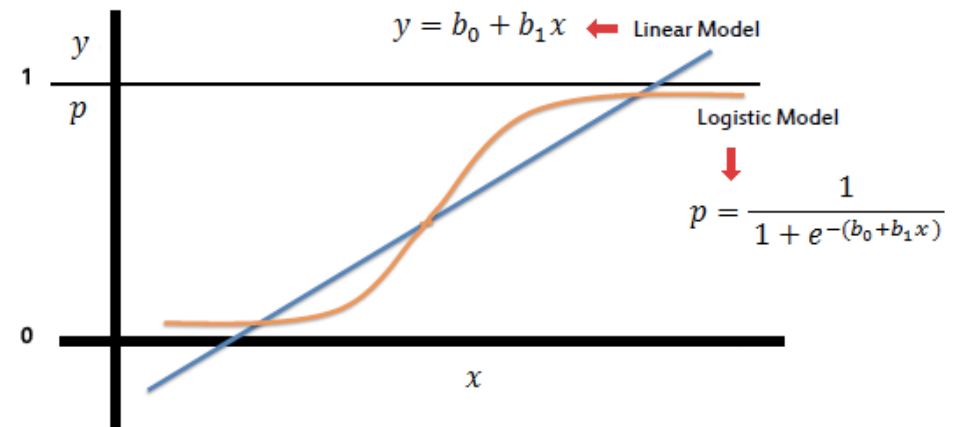
Study	Window, W (hours)	Inclusion criteria
Caballero Barajas and Akella (2015)	24	Age>18, Random fixed size subsample
Caballero Barajas and Akella (2015)	48	Age>18, Random fixed size subsample
Caballero Barajas and Akella (2015)	72	Age>18, Random fixed size subsample
Calvert et al. (2016b)	5*	Age>18, In MICU, >1 obs. for all features, LOS \geq 17hr, ICD-9 codes indicating alcohol withdrawal
Calvert et al. (2016a)	5*	Age>18, In MICU, >1 obs. for all features, 500hr \geq LOS \geq 17hr
Celi et al. (2012)	72	ICD-9 code 584.9
Celi et al. (2012)	24	ICD-9 code 430 or 852
Che et al. (2016) (b)	48	PhysioNet 2012 Challenge dataset
Ding et al. (2016)	48	PhysioNet 2012 Challenge dataset
Ghassemi et al. (2014)	12	Age>18, >100 words across all notes
Ghassemi et al. (2014)	24	Age>18, >100 words across all notes
Ghassemi et al. (2015)	24	Age>18, >100 words across all notes, >6 notes

Live coding time!



Logistic regression refresher

- Powerful (and simple!) predictive model for binary outcomes
- To avoid overfitting, L1 or L2 regularization commonly used
- Maximum likelihood estimation -> sklearn has built in solver



Live Coding Rules

- David controls keyboard and screen
- Each student gets to decide what to do next:
 - Dictate line of code
 - Google something (or Irene can directly tell them)
 - Look up MIMIC documentation, e.g.
<https://mimic.physionet.org/mimictables/admissions/>
 - Say “pass”
- If three “pass”-es in a row, David gives the answer.