Understanding Behavior of Clinical Models under Domain Shifts

Jayaraman J. Thiagarajan[1], Deepta Rajan[2] and Prasanna Sattigeri[2]
[1]Lawrence Livermore National Laboratory, [2]IBM Research AI

Deep Learning for Clinical Diagnosis

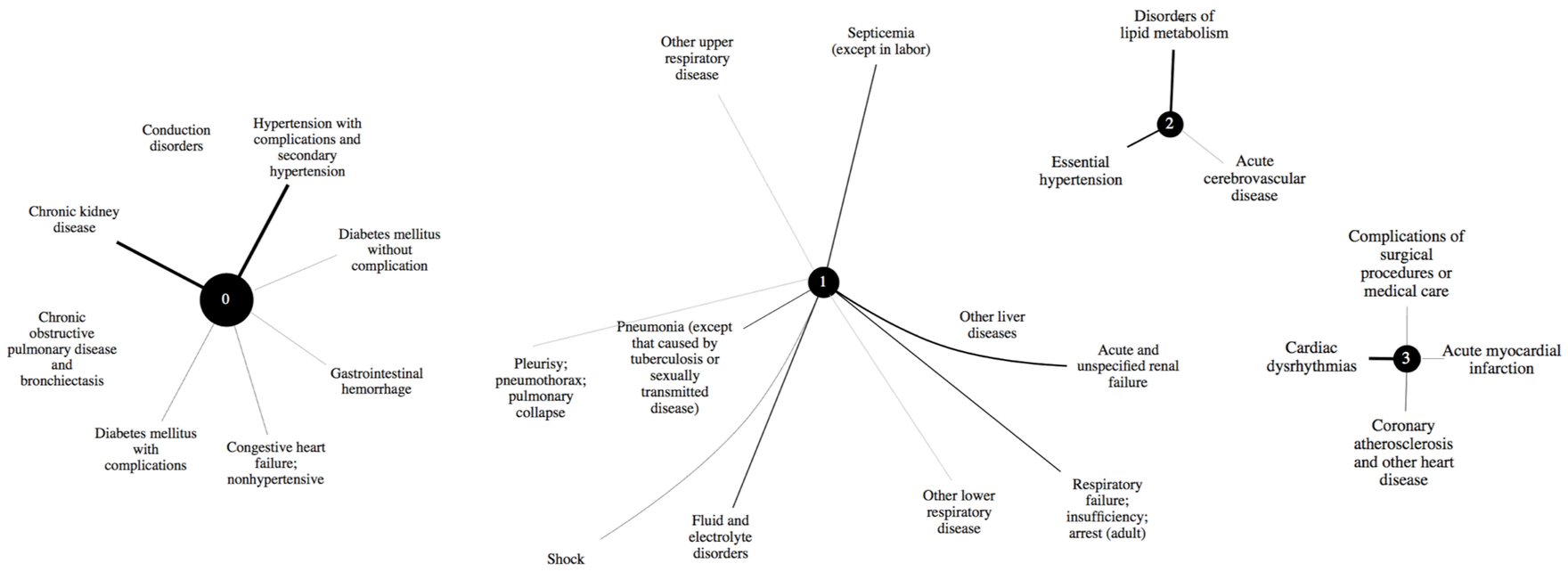
- **Why use AI in Healthcare?**: To leverage knowledge from complex, large-scale information systems, and to build efficient models that are reliable enough to be adopted for diagnosis.
- **Current Practice**: Deep learning (CNNs, RNNs, GANs) has been a game changer in building predictive models for clinical diagnosis.

Evaluating Model Behavior

- **Challenge**: No amount of curated data can be fully representative of what the model might encounter when deployed.
- **Proposed Work**: Characterize classes of domain shifts that can occur in clinical settings, and evaluate the behavior of a predictive model in light of those shifts in order to quantify its reliability.

Scenario Design – Characterizing Domain Shifts

- **Data**: De-identified EHR of ICU patients from the MIMIC-III database, with each record containing 76 vital measurements.
- **Task**: *Acute care phenotyping*, i.e., predicting the likely disease conditions using EHR (25 different conditions - critical, chronic and mixed).
- **Model**: Resnet-1D neural network architecture comprised of stacks of 1D convolution layers with skip connections.
- Each scenario represents the task of detecting the presence of "related" disease conditions – Co-occurrence of diseases varies between populations.



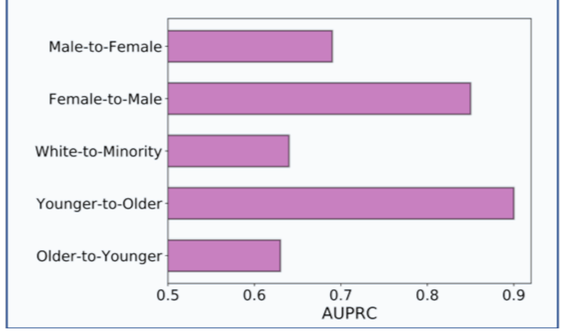
Analysis

Population Biases

Age: a. Older (60+) to younger (<=60); b. Younger to older

Gender: a. 90%M-10%F to 10%M-90%F; b. 10%M-90%F to 90%M-10%F

Race: Whites to Minority

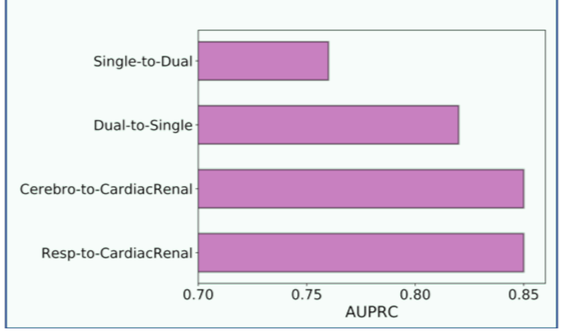


Label Distribution Shifts

Novel Diseases: a. [Resp] to [Resp + Renal + Cardiac]; b. [Cerebro] to [Cerebro + Renal + Cardiac]

Dual to Single: [Cardiac + Renal]

Single to Dual: [Cardiac], [Renal]

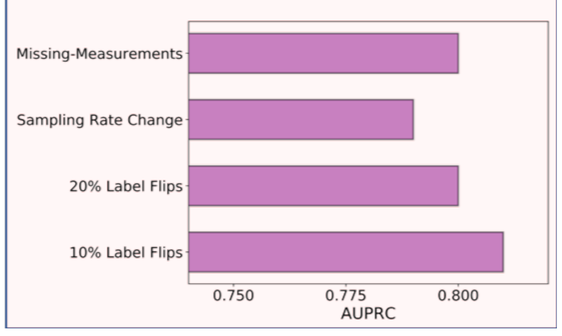


Measurement Discrepancies

Noisy Labels: [Resp] to [Resp + Renal + Cardiac], with 10% or 20% label flips

Sampling Rate Change: 96h to 48h window

Missing Meas.: pH, Temperature, Height, Weight, and all Verbal Response GCS



Shift	Source Diseases	Target Diseases
Older-to-Younger	Coronary atherosclerosis, Disorders of lipid metabolism, Essential hypertension	Coronary atherosclerosis, Disorders of lipid metabolism, Essential hypertension, Chronic kidney disease, Secondary hypertension
Male-to-Female	Dysrhythmia, Congestive heart failure	Dysrhythmia, Congestive Heart Failure, Coronary atherosclerosis
White-to-Minority	Dysrhythmia, Conduction disorder, Congestive heart failure	Dysrhythmia, Conduction disorder, Congestive heart failure, Chronic kidney disease, Secondary hypertension, Diabetes
Single-to-Dual	Cardiac-only or Renal-only	Both Cardiac and Renal Diseases

Lawrence Livermore National Laboratory

IBM