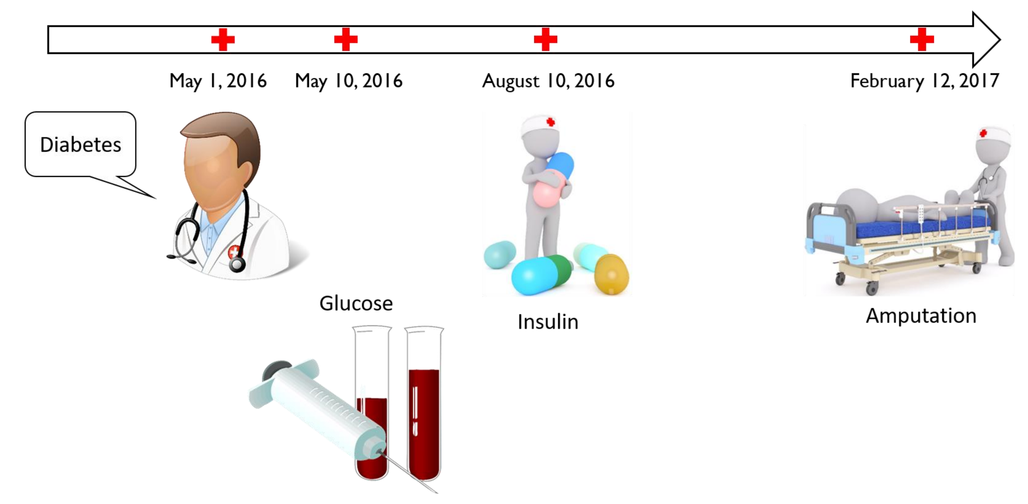
Resolving the Bias in Electronic Medical Records

Kaiping Zheng, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, Wei Luen James Yip

National University of Singapore, National University Health System

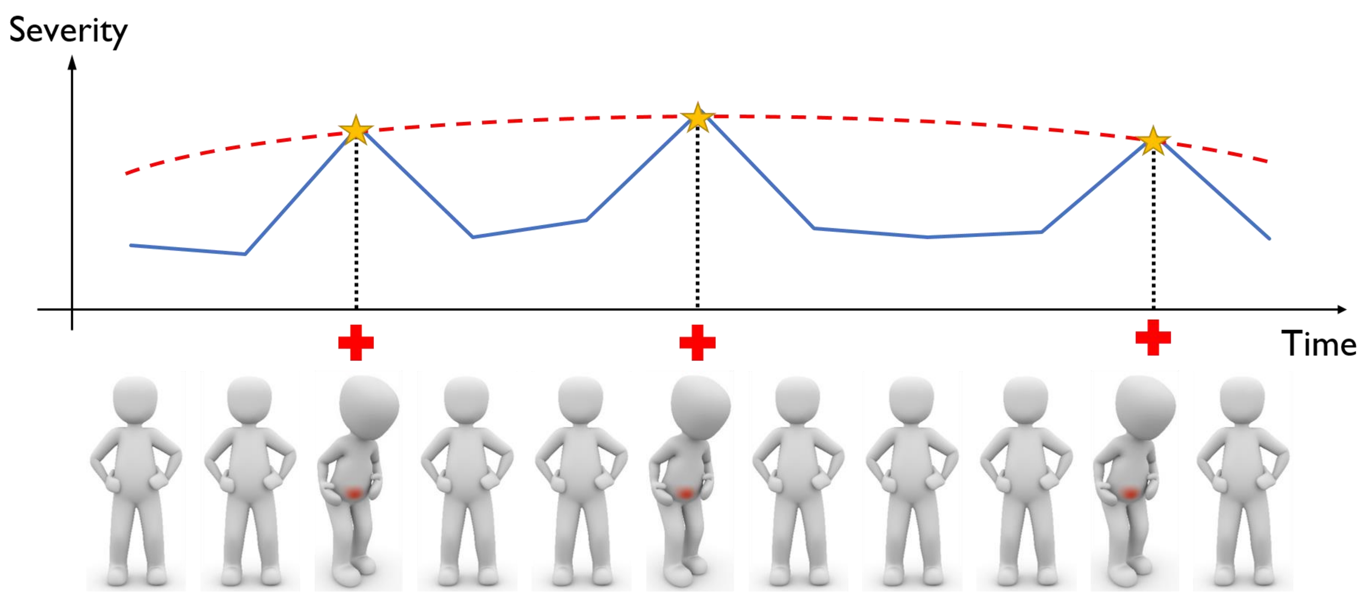
INTRODUCTION

Electronic Medical Records (EMR Data)



Bias in EMR Data

- Patients tend to visit hospital more often when they feel sick
- Doctors tend to prescribe the lab examinations that show abnormality



METHODOLOGY

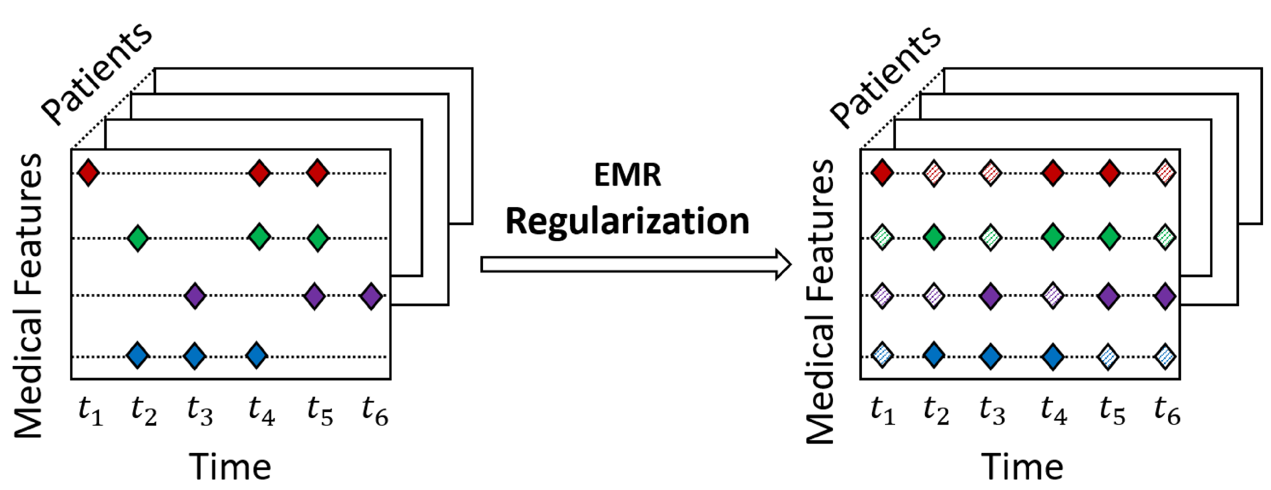
EMR Regularization

Inspiration

- EMR Series Ψ is not a randomly sampled subset of Patients' Hidden Conditions Φ
- Probability that one tuple $< p, t, d, v >$ (p: patient, t: time point, d: feature, v: value) is observed may depend on the medical feature and its value

Target of our work

- Estimate the unobserved hidden conditions $\Phi - \Psi$ using EMR series Ψ



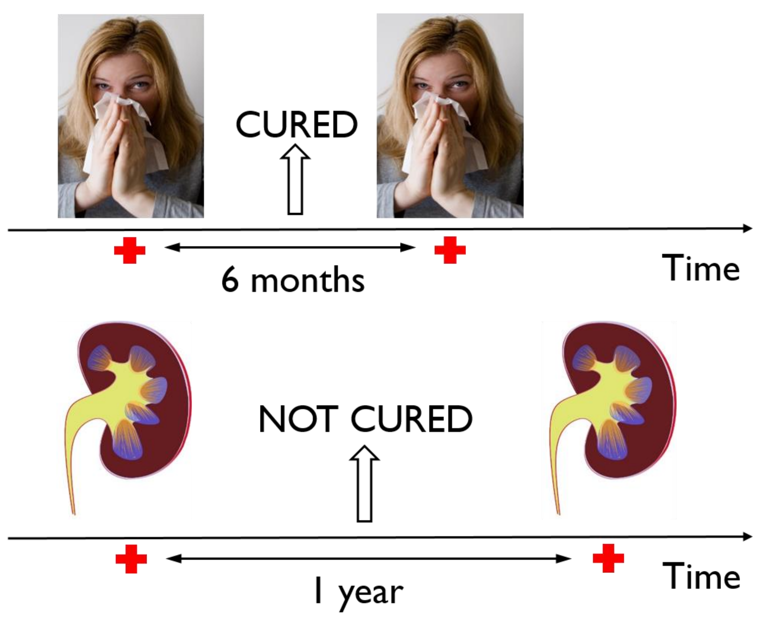
Characteristics of Medical Features

Condition Change Rate (CCR)

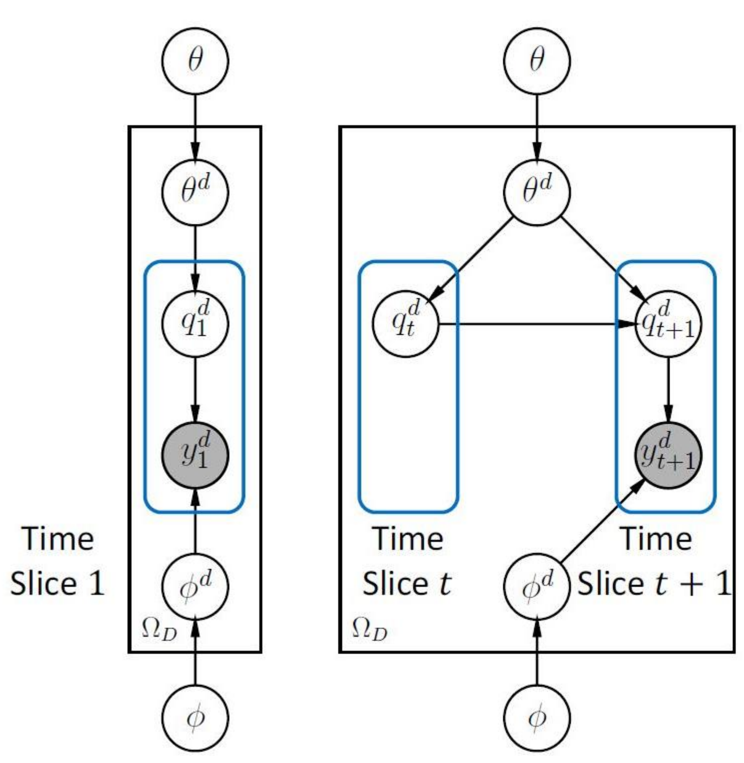
- How a medical feature is likely to change from its condition in the previous observation

Observation Rate (OR)

- Probability that a medical feature is exposed at a time point based on its actual condition at that time point



A Hidden Markov Model (HMM) Variant for Learning and Inference



Algorithm 1: EMR regularization with smoothing

Input: medical features Ω_D, observation sequences $\Omega_S^d = \{Y^{d,s} | Y^{d,s} = y_1^{d,s}, \cdots, y_T^{d,s}\}$ for each feature d and for each sequence s. A's prior for feature d is $Beta(a_A^d, b_A^d)$, B's prior for feature d is $Beta(a_B^d, b_B^d)$.

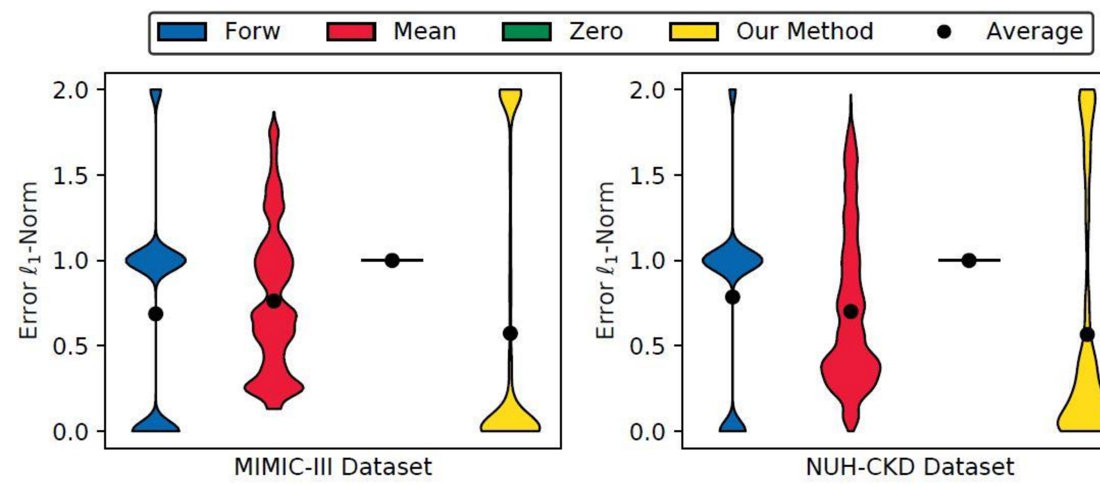
Output: parameters $\lambda^d = (\Pi^d, A^d, B^d)$ for each $d \in \Omega_D$, hidden state probability sequence $P(q_t^{d,s} = z_i | Y^{d,s}, \lambda^d)$.

```
For each medical feature d ∈ Ω_D
    Initialize λ^d = (Π^d, A^d, B^d)
    Iterate EM process until convergence
    E-Step:
        For each observation sequence s ∈ Ω_S^d
            Compute ξ_t(q_t^{d,s} = z_i, q_{t+1}^{d,s} = z_j) (Equation 3)
            Compute γ_t(q_t^{d,s} = z_j) (Equation 4)
    M-Step:
        Update Π̂_i^d (Equation 5)
        Update transition matrix Â_{i,j}^d (Equation 6)
        Update emission matrix B̂_{j,v_k}^d (Equation 7)
    Compute P(q_t^{d,s} = z_i | Y^{d,s}, λ^d) (Equation 8)
    return λ^d = (Π^d, A^d, B^d), P(q_t^{d,s} = z_i | Y^{d,s}, λ^d)
```

EXPERIMENTS

Imputation Accuracy Evaluation

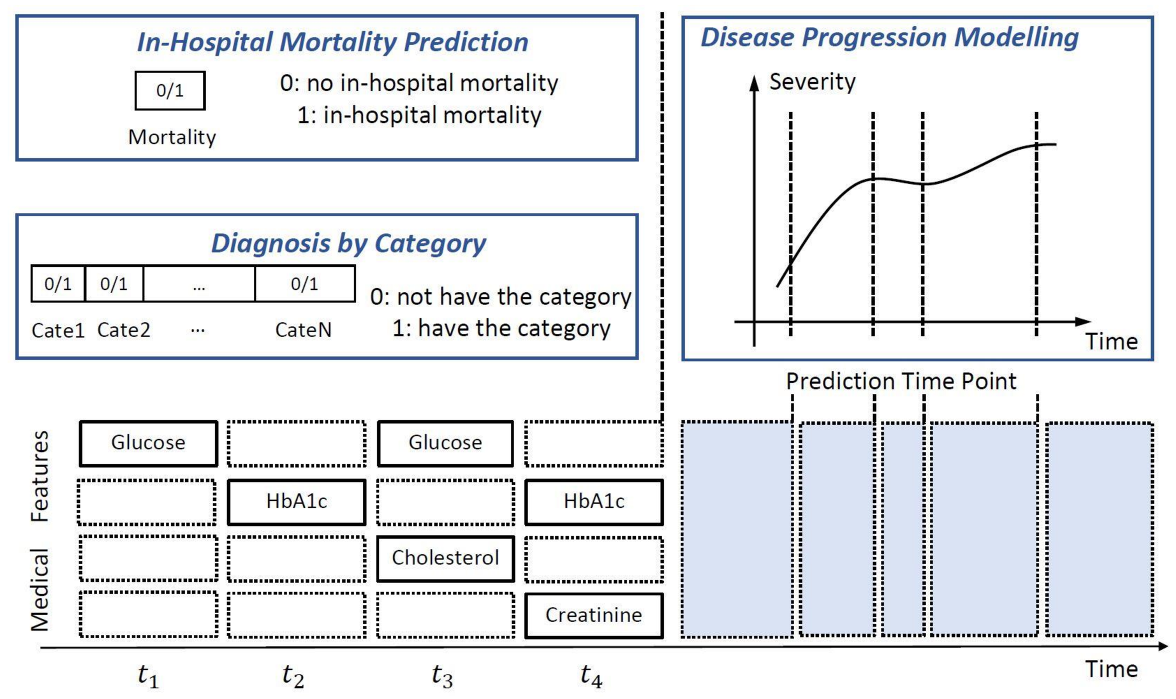
Imputation accuracy evaluation results in two datasets

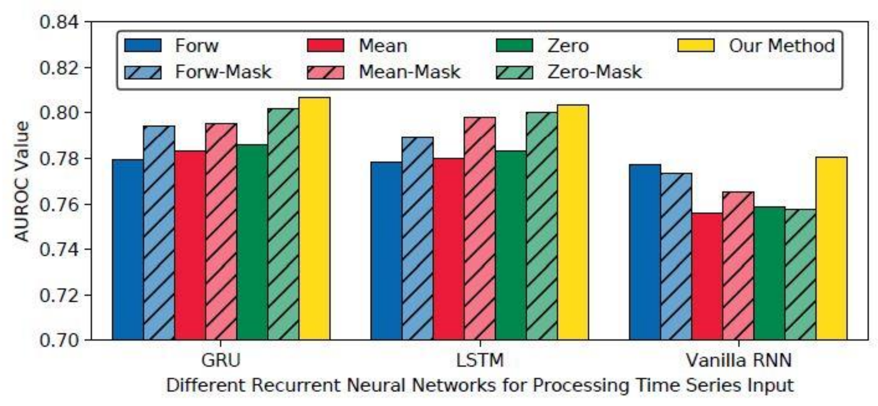


Benefits for Analytical Tasks

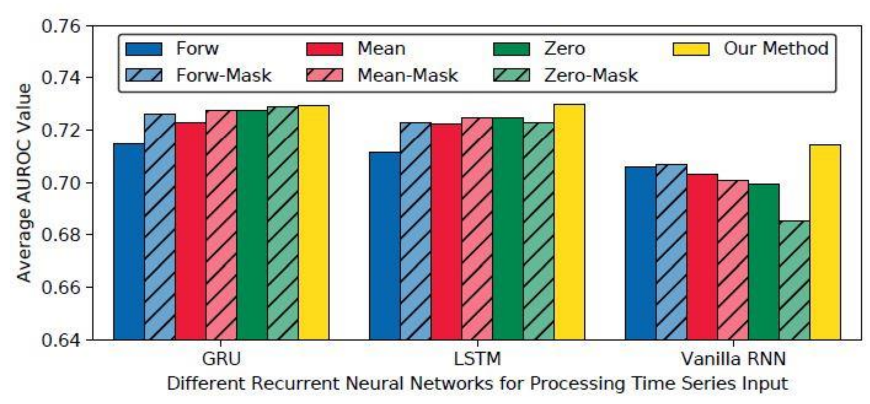
Analytical applications:

- In-hospital mortality
- Diagnosis by category
- Disease progression modelling

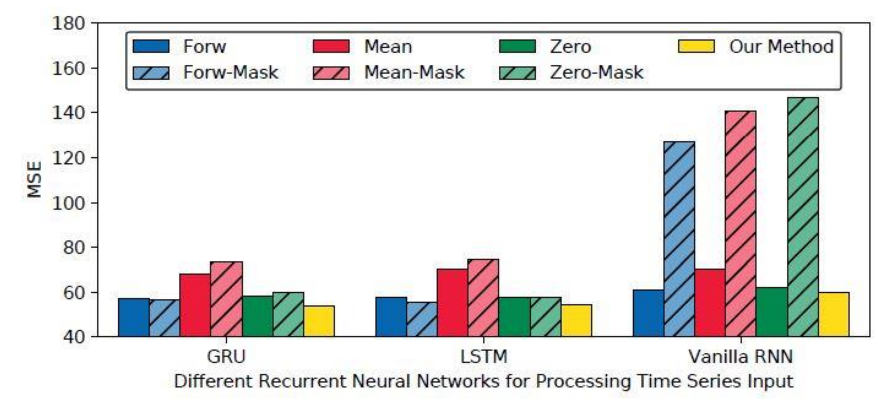




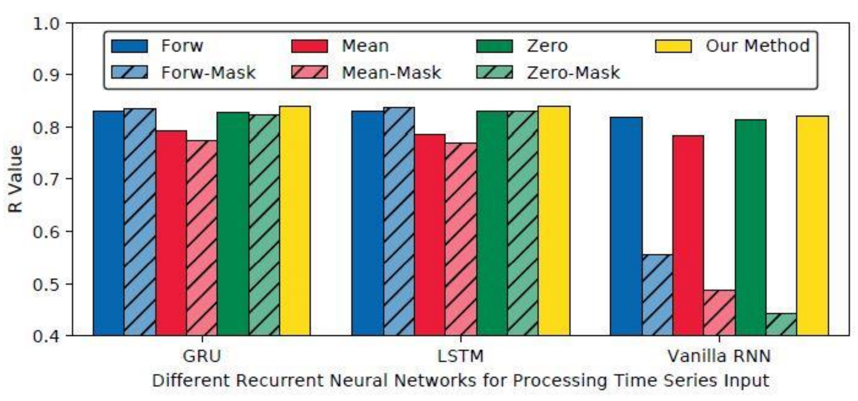
MIMIC-III in-hospital mortality prediction results



MIMIC-III diagnosis by category results



MSE for NUH-CKD disease progression modelling



R value for NUH-CKD disease progression modelling

CONCLUSION AND FUTURE WORK

EMR Regularization to Resolve Bias

- Consider CCR and OR as characteristics of medical features
- Employ an HMM variant for learning and inference
- Impute missing values in EMR data more accurately
- Improve the analytical performance after resolving the bias

Future Directions

- Model different diseases jointly in the probabilistic graphical model for capturing the relationships in between
- Model the patient personalization as different patients might behave differently in terms of CCR and OR

ACKNOWLEDGMENTS

NUS National University of Singapore

NUHS National University Health System