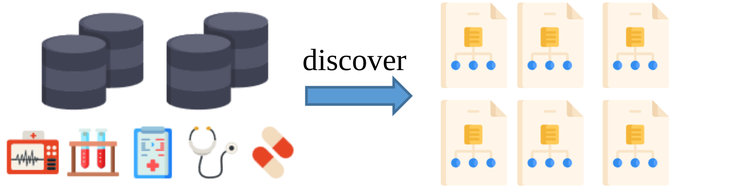
Learning Phenotypes and Dynamic Patient Representations via RNN Regularized Collective Non-negative Tensor Factorization

Kejing Yin[1], Dong Qian[1], William K. Cheung[1], Benjamin C. M. Fung[2], Jonathan Poon[3]

[1]Hong Kong Baptist University [2] McGill University [3]Hong Kong Hospital Authority

{cskjyin, dongqian, william}@comp.hkbu.edu.hk, ben.fung@mcgill.ca, jonathan@ha.org.hk

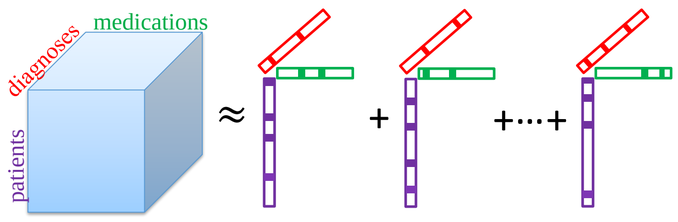
Introduction



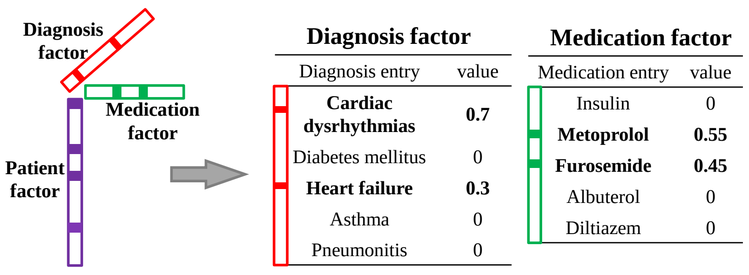
Raw EHR
Largely missing,
Heavily noisy,
Potentially biased.

Phenotypes
(Combinations of Rx, Dx,…)
Clinically meaningful,
Highly interpretable,
Reveal true disease status.

❖ Phenotyping using Tensor CP Factorization



- **CP factorization** approximates the interaction tensor with the sum of rank-one tensors.

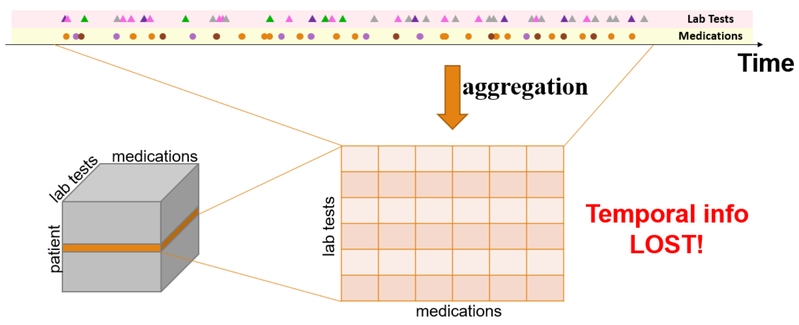


Diagnosis factor	
Diagnosis entry	value
Cardiac dysrhythmias	**0.7**
Diabetes mellitus	0
Heart failure	**0.3**
Asthma	0
Pneumonitis	0

Medication factor	
Medication entry	value
Insulin	0
Metoprolol	**0.55**
Furosemide	**0.45**
Albuterol	0
Diltiazem	0

- Each rank-one tensor is defined as a candidate of the resulting phenotypes.
- The patient factor can be used as patient representations.

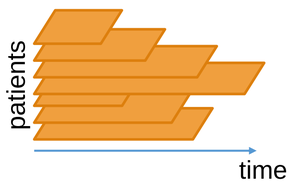
❖ Limitation: Temporal Dynamics Unexplored



- The temporal dynamics are simply ignored.
- Consequently, disease states appear at different times are **mixed together**.
- Especially in inpatients data and ICU cases.

❖ Adding Time as a Dimension?

- **Aligning Patients**: Patients have different Length-of-Stay (LoS), thus very difficult to represent all patients with one temporal tensor.



- **Phenotypes being dynamic**: With a temporal tensor for all patients, the global temporal patterns are embedded into the phenotypes, making it difficult to interpret.

❖ Our Goal

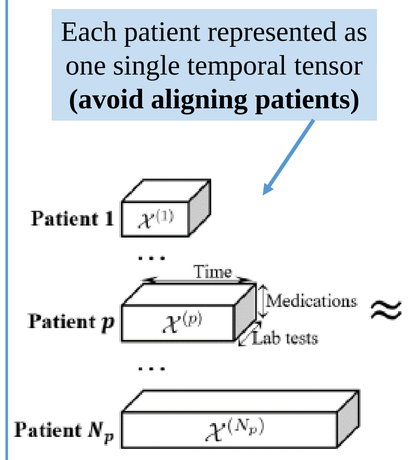
- **Avoid aligning patients.**
- **Keep the phenotypes static.**
- **Embed the individual disease progression into the dynamic patient representation.**
- **Explore the global temporal patterns as well.**

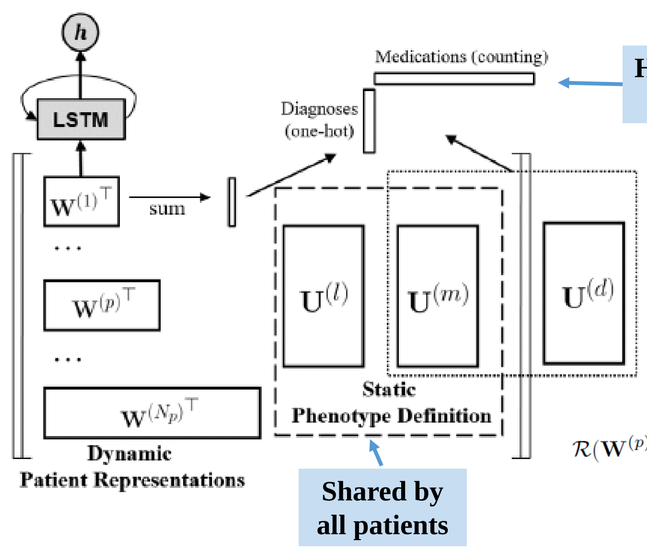
Related Works

- **Marble**[2]: Poisson non-negative tensor factorization model for computational phenotyping task, with a bias tensor to capture the global information.
- **Rubik**[3]: Incorporating guidance of existing medical knowledge, and adding pairwise constraint to make phenotypes distinct.
- **SPARTan**[6]: Represent the EHR data using an "irregular" tensor and solve for the factors using PARAFAC2 decomposition.
- However, this model does not explicitly consider the interaction between the diagnoses and medications.

Proposed Method

Each patient represented as one single temporal tensor **(avoid aligning patients)**





HITF for incorporating **non-temporal** Dx.

$$\ell = \alpha_1 f^{CNTF} + \alpha_2 \sum_{p=1}^{N_p} f_p^{HITF} + \beta \sum_{p=1}^{N_p} \mathcal{R}(\mathbf{W}^{(p)})$$

where $f^{CNTF} \equiv \sum_{p=1}^{N_p} \frac{1}{T_p} \left(\sum_{ijk} \hat{x}_{ijk}^{(p)} - x_{ijk}^{(p)} \log \hat{x}_{ijk}^{(p)} \right)$

$$f_p^{HITF} \equiv \sum_i \hat{d}_i^{(p)} - d_i^{(p)} \log(e^{\hat{d}_i^{(p)}} - 1) + \sum_j \hat{m}_j^{(p)} - m_j^{(p)} \log \hat{m}_j^{(p)}$$

$$\mathcal{R}(\mathbf{W}^{(p)}) = \frac{1}{T_p} \sum_{t=2}^{T_p} \left\| g(\mathbf{w}_{t-1}) - \mathbf{w}_t \right\|_2^2$$

$g(w_{t-1})$: the output given by LSTM.

Experiments and Results

- **Data Set**: MIMIC-III, an open-source, large-scale, de-identified ICU patients related EHR data set.
- Extract a subset containing 4,590 adult patients with LoS longer than 7 days.
- **Baselines**: **Rubik** (one of the state-of-the-art NTF computational phenotyping models),
  CP-APR (widely used Poisson non-negative tensor factorization model).

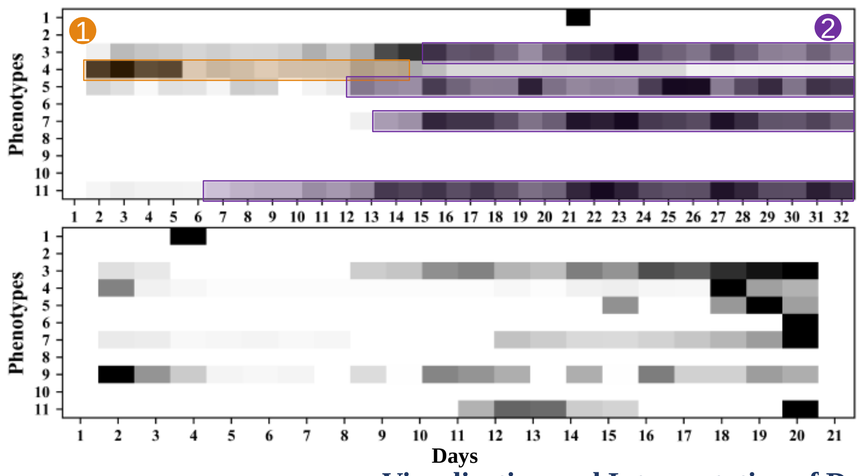
Phenotype Comparison: Phenotypes derived by our proposed model (top) and Rubik (bottom)

	Phenotype 1	Phenotype 4	Phenotype 9
Clinically Highly Relevant	Chronic kidney disease (CKD) (0.536)	Other forms of chronic ischemic heart disease (0.507) Cardiac dysrhythmias (0.372) Essential hypertension (0.024)	Other diseases of lung (0.876)
Chronic Disease	RBC (Urine) (0.200) Osmolality, Measured (Blood) (0.117) Protein/Creatinine Ratio (Urine) (0.069)	Hematocrit (Blood) (0.072) Red Blood Cells (Blood) (0.071) Hemoglobin (Blood) (0.070)	pO2 (Blood Gas) (0.253) pCO2 (Blood Gas) (0.237) pH (Blood Gas) (0.215)
	Hydromorphone (0.336) Phenylephrine (0.038) Aspirin (0.033)	Acetaminophen (0.188) Metoclopramide (0.102) Insulin Human Regular (0.070)	Acetaminophen (0.113) Insulin (0.099) Bisacodyl (0.089)

Acute Respiratory Failure (classified "other disease of lung")

Phenotype 1	Phenotype 2	Phenotype 3
Other diseases of lung (0.045) Septicemia (0.040)	Other diseases of lung (0.040) Acute kidney failure (0.036)	Acute kidney failure (0.039) Other diseases of lung (0.037)
Certain adverse effects not elsewhere classified (0.039)	Certain adverse effects not elsewhere classified (0.032)	Cardiac dysrhythmias (0.033)
Glucose(Blood) (0.019) Red Blood Cells(Blood) (0.019) Hematocrit(Blood) (0.019)	Hematocrit(Blood) (0.017) Red Blood Cells(Blood) (0.017) Glucose(Blood) (0.017)	Glucose(Blood) (0.018) Hematocrit(Blood) (0.018) Red Blood Cells(Blood) (0.018)
Vancomycin (0.017) Insulin (0.015) Potassium Chloride (0.015)	Vancomycin (0.013) Potassium Chloride (0.013) Pantoprazole Sodium (0.012)	Vancomycin (0.015) Potassium Chloride (0.014) Heparin (0.014)

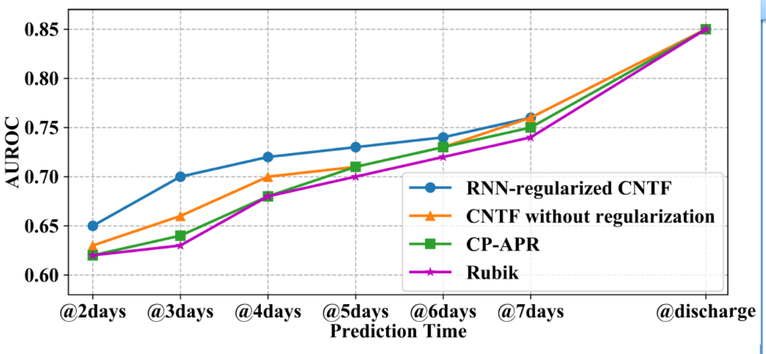
Dominated by acute diseases.



1. High value for phenotype 4 (Chronic Heart Disease) in the first several days.
2. High value for phenotype 3 (Other Disease of the Lung), phenotype 5 (Cardiac Dysrhythmias), phenotype 7 (Acute Kidney Failure), phenotype 11 (Cardiac Dysthymias with Heart Failure)

Patient **admitted with existing condition**, chronic heart disease, which is **treated unsuccessfully**, and **eventually developed multiple organ failure**. *(Supported by reviewing the clinical notes.)*

Visualization and Interpretation of Dynamic Patient Representations



Mortality Prediction

Contributions

- We proposed CNTF to jointly learn the dynamic patient representations and the static globally shared phenotypes.
- RNN-based regularization and HITF model are integrated to model the time dependency and incorporate the non-temporal modalities.
- The proposed model can derive clinically meaningful and interpretable phenotypes, and better **separate the disease states appearing at different time**.

References

[1] Chi, Eric C., and Tamara G. Kolda. "On tensors, sparsity, and nonnegative factorizations." *SIAM Journal on Matrix Analysis and Applications* 33.4 (2012): 1272-1299.

[2] Ho, Joyce C., Joydeep Ghosh, and Jimeng Sun. "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization." *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2014.

[3] Wang, Yichen, et al. "Rubik: Knowledge guided tensor factorization and completion for health data analytics." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[4] Kim, Yejin, et al. "Discriminative and distinct phenotyping by constrained tensor factorization." *Scientific Reports* 7.1 (2017): 1114.

[5] Gunasekar, Suriya, et al. "Phenotyping using Structured Collective Matrix Factorization of Multi-source EHR Data." *arXiv preprint* arXiv:1609.04466 (2016).

[6] Yin, Kejing, et al. "Joint Learning of Phenotypes and Diagnosis-Medication Correspondence via Hidden Interaction Tensor Factorization." *IJCAI*. 2018.

[7] Perros, Ioakeim, et al. "SPARTan: Scalable PARAFAC2 for large & sparse data." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.