Data Analytics with Electronic Health Records

Fei Wang
Associate Professor
Department of Computer Science and Engineering
University of Connecticut
fei_wang@uconn.edu

Slides available through my website
http://www.engr.uconn.edu/~fwang/tutorials/AAAI15_tutorial.pdf

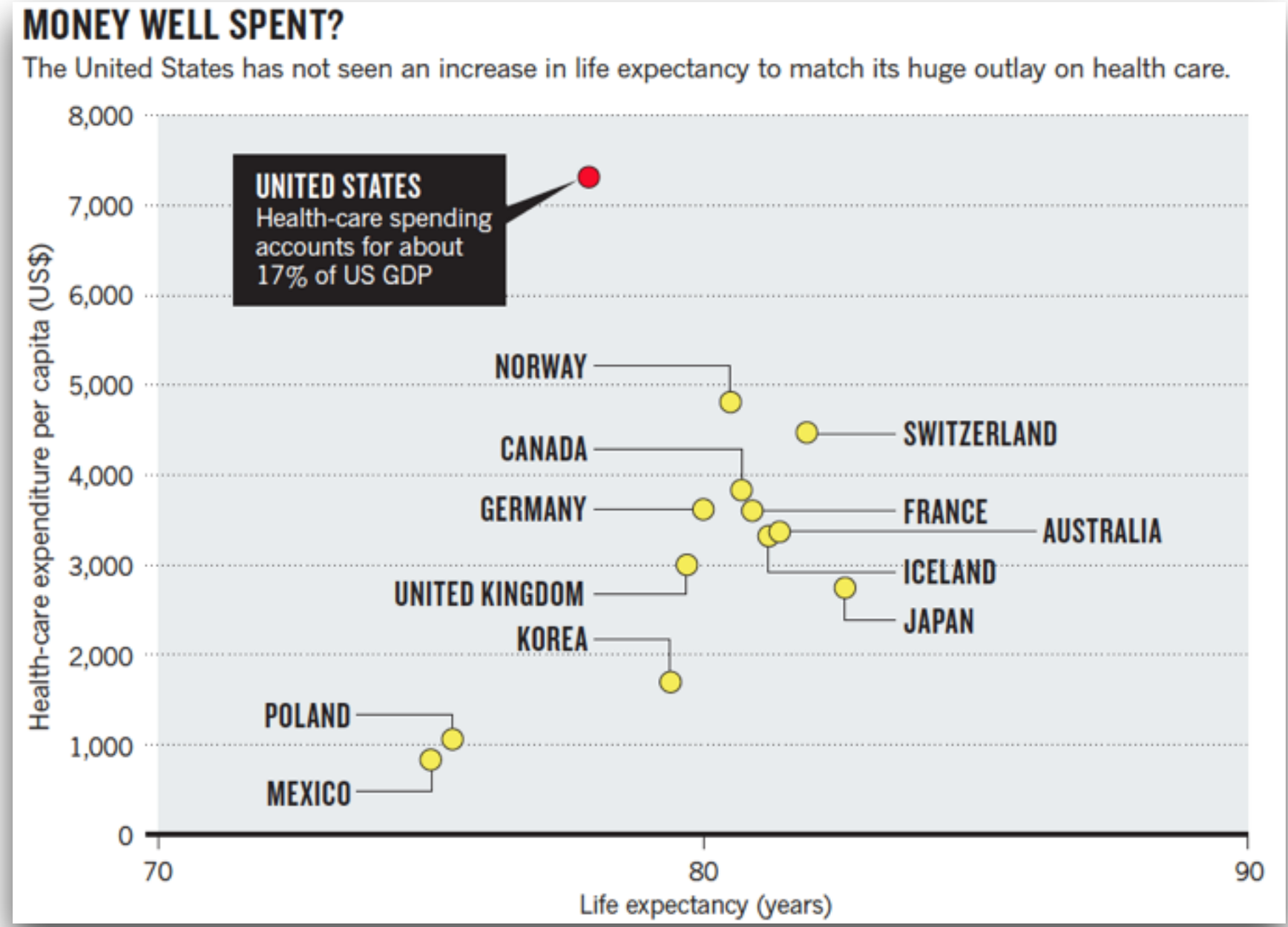
Roadmap

- Background
- Electronic Health Records
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Roadmap

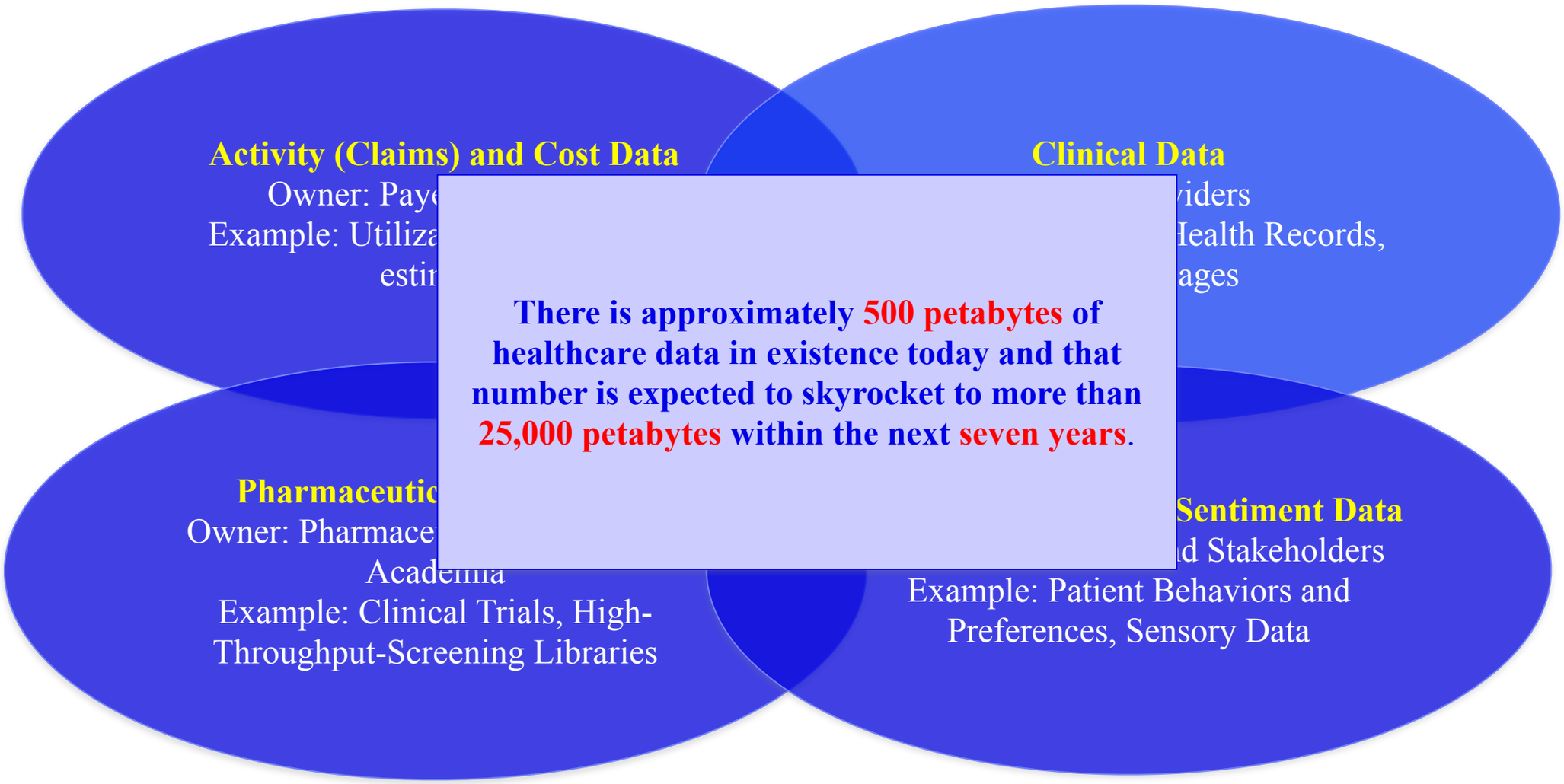
- Background
- Electronic Health Records
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Healthcare Is in Crisis



Hersh, W., Jacko, J. A., Greenes, R., Tan, J., Janies, D., Embi, P. J., & Payne, P. R. (2011). Health-care hit or miss? Nature, 470(7334), 327.

Healthcare Data



Groves, P., Kayyali, B., Knott, D. & Kuilen, S. V. (2013). The 'Big-Data' Revolution in Healthcare. MicKinsey & Company Report.

Drug Network

Drug Repurposing

Genomic Medicine

Patient Network

Disease Network

Phenotype/ Genotype Association

TACSTD1
IL2
PTPN6
LAIR1
PTPN11
CD8A

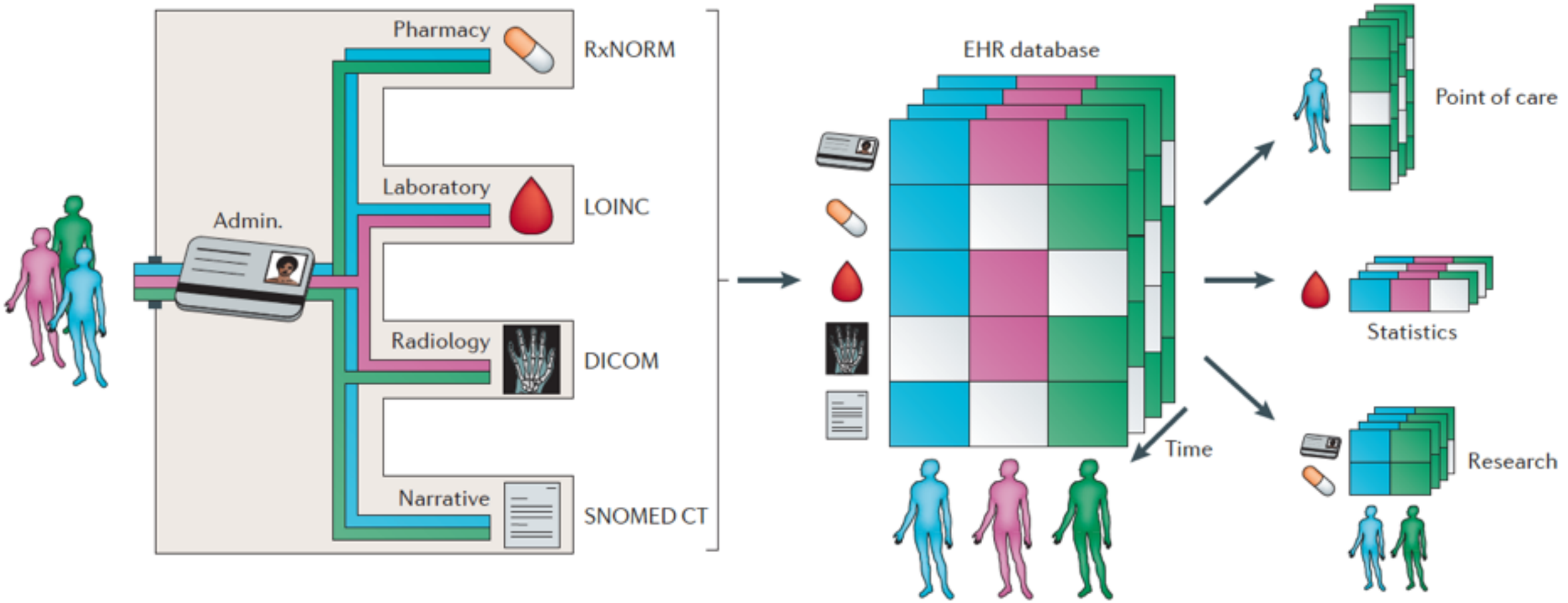
Gene Network

Roadmap

- Background
- Electronic Health Records
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Electronic Health Records

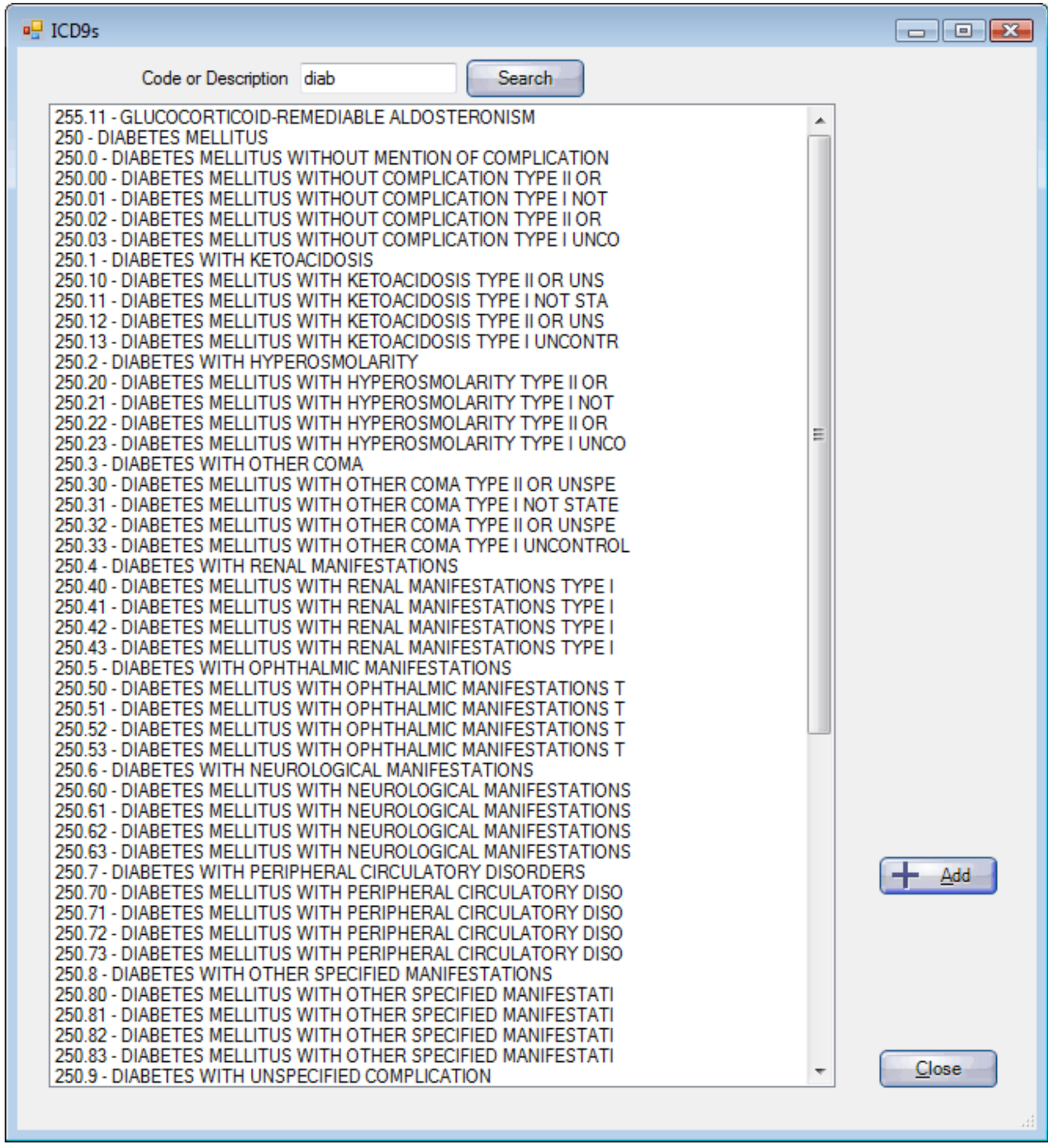
An Electronic Health Record (EHR) is an evolving concept defined as a systematic collection of electronic health information about individual patients or populations



Jensen, Peter B., Lars J. Jensen, and SØren Brunak. "Mining electronic health records: towards better research applications and clinical care." Nature Reviews Genetics (2012).

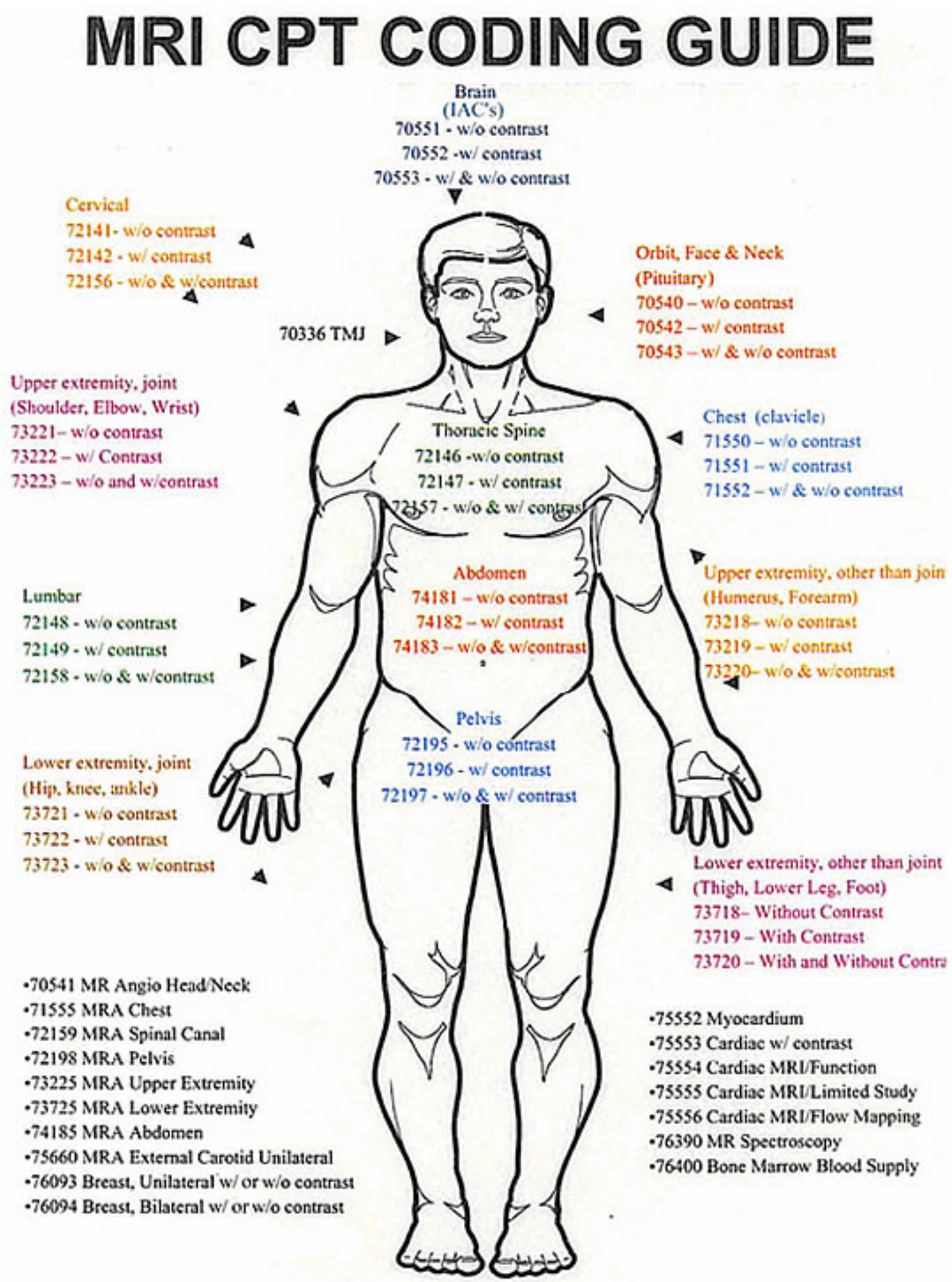
Diagnosis-ICD Codes

- ICD stands for International Classification of Diseases
- ICD is a hierarchical terminology of diseases, signs, symptoms, and procedure codes maintained by the World Health Organization (WHO)
- In US, most people use ICD-9, and the rest of world use ICD-10



Procedure - CPT Codes

- CPT: Current Procedural Terminology (CPT)
- Describes medical, surgical, and diagnostic services and is designed to communicate uniform information about medical services and procedures among physicians, coders, patients, accreditation organizations, and payers for administrative, financial, and analytical purposes

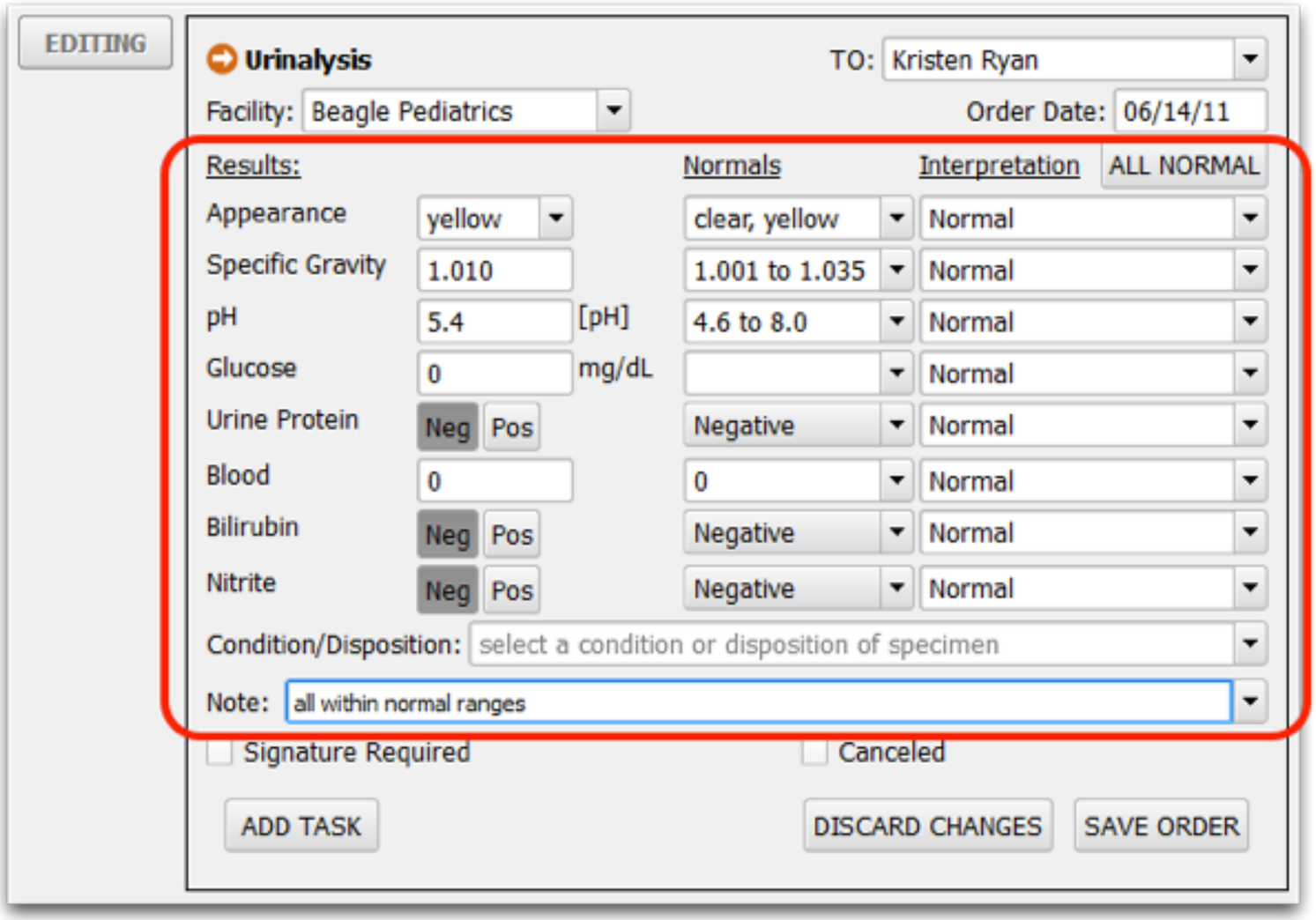


https://accessprofessional.wordpress.com/2012/02/28/mri-coding-guide/

Lab Tests

The standard code for lab is Logical Observation Identifiers Names and Codes (LOINC®)

- Challenges for lab
  - Many lab systems still use local dictionaries to encode labs
  - Diverse numeric scales on different labs
    - Often need to map to normal, low or high ranges in order to be useful for analytics
- Missing data
  - not all patients have all labs
  - The order of a lab test can be predictive, for example, BNP (B-type Natriuretic Peptide) indicates high likelihood of heart failure



Medication

- Standard code is National Drug Code (NDC) by Food and Drug Administration (FDA), which gives a unique identifier for each drug
- Not used universally by EHR systems
- Too specific, drugs with the same ingredients but different brands have different NDC
- RxNorm: a normalized naming system for generic and branded drugs by National Library of Medicine
- Medication data can vary in EHR systems
  - can be in both structured or unstructured forms
  - Availability and completeness of medication data vary
  - Inpatient medication data are complete, but outpatient medication data are not
  - Medication usually only store prescriptions but we are not sure whether patients actually filled those prescriptions

Clinical Notes

Clinical notes contain rich and diverse source of information

- Challenges for handling clinical notes
  - Ungrammatical, short phrases
  - Abbreviations
  - Misspellings
  - Semi-structured information
    - Copy-paste from other structure source
      - Lab results, vital signs
    - Structured template:
      - SOAP notes: Subjective, Objective, Assessment, Plan

Surgery Service, Dr. Jones

Subjective: No further Chest Pain or Shortness of Breath. "Feeling better today." Patient reports headache.	***Objective:*** Afebrile, P 84, R 16, BP 130/82. No acute distress. Neck no JVD, Lungs clear Cor RRR Abd Bowel sounds present, mild RLQ tenderness, less than yesterday. Wounds look clean. Ext without edema
Assessment: Patient is a 37 year old man on post-operative day 2 for laparoscopic appendectomy.	***Plan:*** Recovering well. Advance diet. Continue to monitor labs. Follow-up with Cardiology within three days of discharge for stress testing as an out-patient. Prepare for discharge home tomorrow morning.

http://en.wikipedia.org/wiki/SOAP_note

	ICD codes	CPT codes	Laboratory Data	Medication records	Clinical Documentation
Availability in EHR systems	Near-universal	Near-universal	Near-universal	Variable	Variable
Recall	Medium	Poor	Medium	Inpatient: High Outpatient: Variable	Medium
Precision	Medium	High	High	Inpatient: High Outpatient: Variable	Medium-High
Fragmentation effect	Medium	High	Medium-High	Medium	Low-Medium
Query method	Structured	Structured	Mostly structured	Structured, text queries, and NLP	NLP, text queries, and rarely structured
Strengths	-Easy to query -Serves as a good first pass of disease status	-Easy to query -High precision	-Value depends on test -High data validity	Can have high validity	Best record of what providers thought
Weaknesses	-Disease codes often used for screening when disease not a ctually present -Accuracy hindered by billing realities and clinic workflow	-Most susceptible to missing data errors (e.g., performed at another hospital) -Procedure receipt influenced by patient and payer factors external to disease process	-May need to aggregate different variations of the same data elements -Normal ranges and units may change over time	-Often need to interface inpatient and outpatient records -Medication records from outside providers not present -Medications prescribed not necessary taken	-Difficult to process automatically -Interpretation accuracy depends on assessment method -May suffer from significant "cut and paste" -Not universally available in EHRs -May be self-contradictory
Summary	Essential first element for electronic phenotyping	Helpful addition if relevant	Helpful addition if relevant	Useful for confirmation and a marker of severity	Useful for confirming common diagnoses or for finding rare ones

doi:10.1371/journal.pcbi.1002823.t001

Joshhua C. Denny Chapter 13: Mining Electronic Health Records in the Genomics Era. PLoS Comput Biol. 2012 December; 8(12)

Vector Based Representation

A collection of numbers

Row Vector

$\boldsymbol{v} = [v_1, v_2, \ldots, v_d]$

Dimensionality

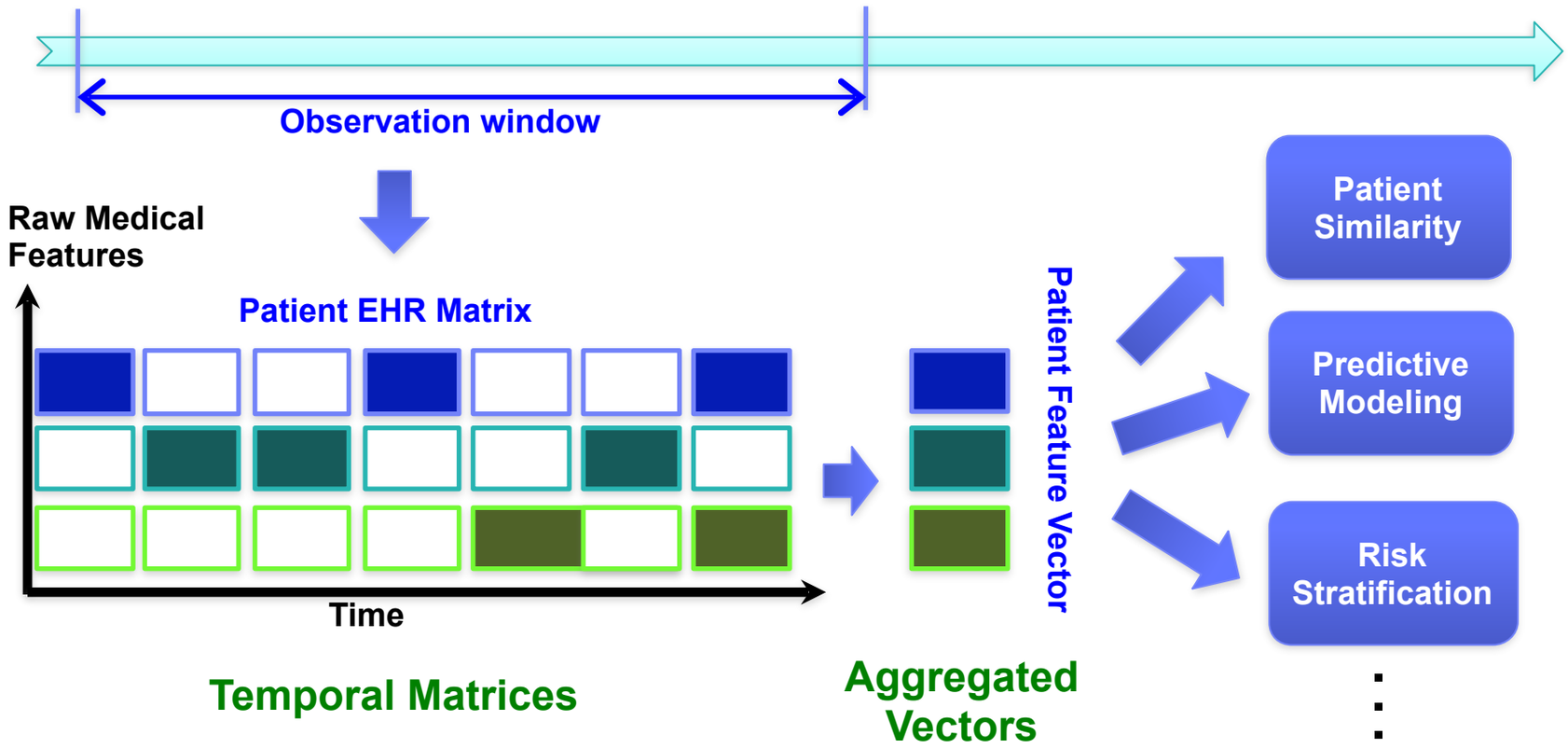
Transpose

Column Vector

$$v' = \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_d \end{bmatrix}$$

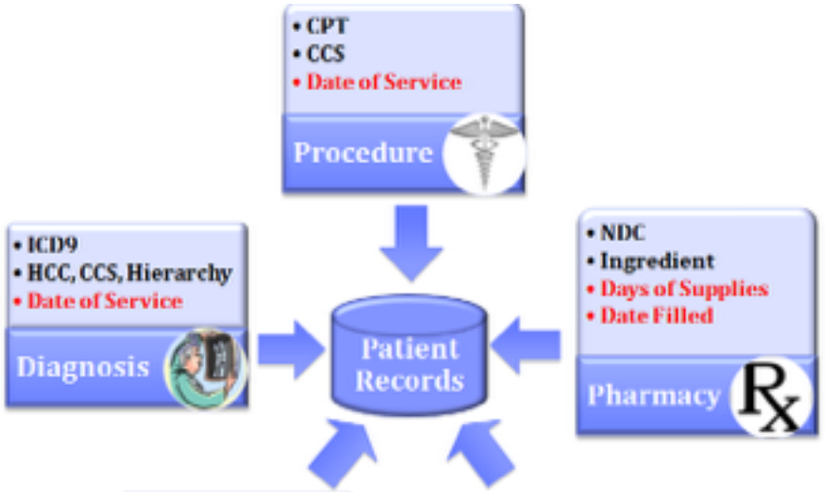
Patient Diagnosis Vector: d is the number of distinct diagnosis code, x_i represents the frequency of the i-th diagnosis code in his/her historical records

Matrix Based Representation



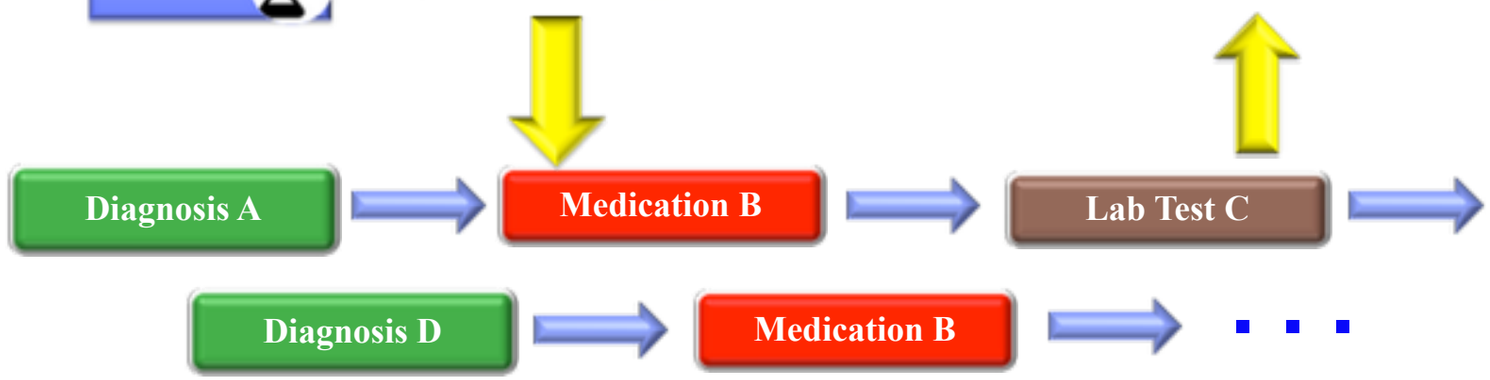
- Jimeng Sun, Fei Wang, Jianying Hu, Shahram Edabollahi: Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explorations 14(1): 16-24 (2012)
- Fei Wang, Jimeng Sun, Shahram Ebadollahi: Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. Statistical Analysis and Data Mining 5(1): 54-69 (2012)
- J.Wu, J. Roy,W. F. Stewart, Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches, Medical Care 48 S106–S113 (2010)

Sequence Based Representation



How to interpret and make use of the sequentiality of the events?

The sequentiality of those events may indicate some impending disease conditions

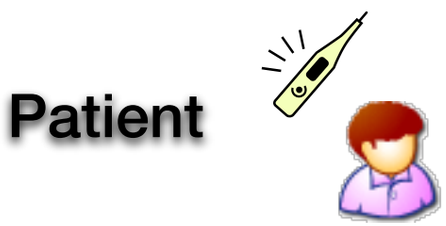


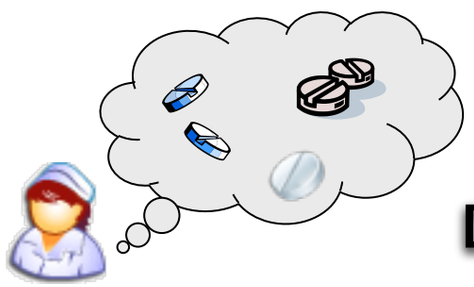
Roadmap

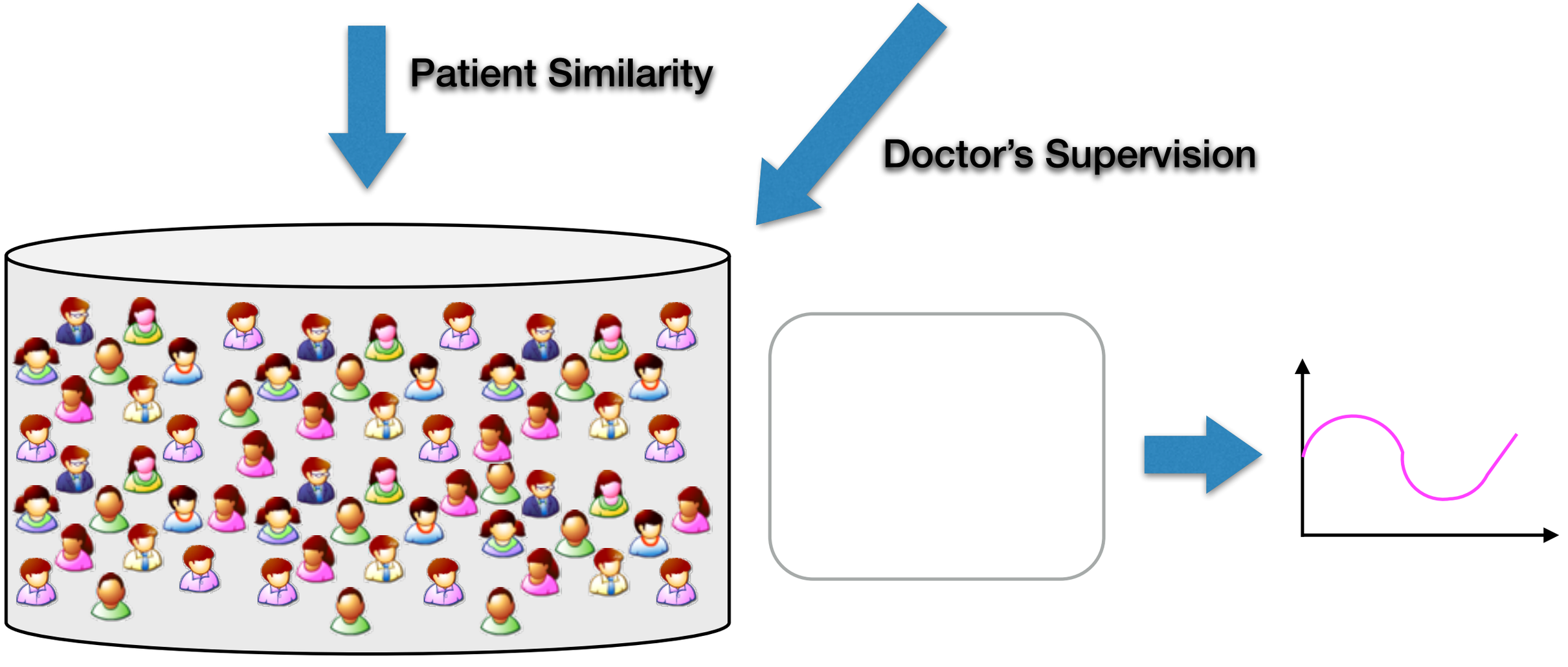
- Background
- Electronic Health Records
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Patient Similarity Problem

Prognostication/Outcome Analysis

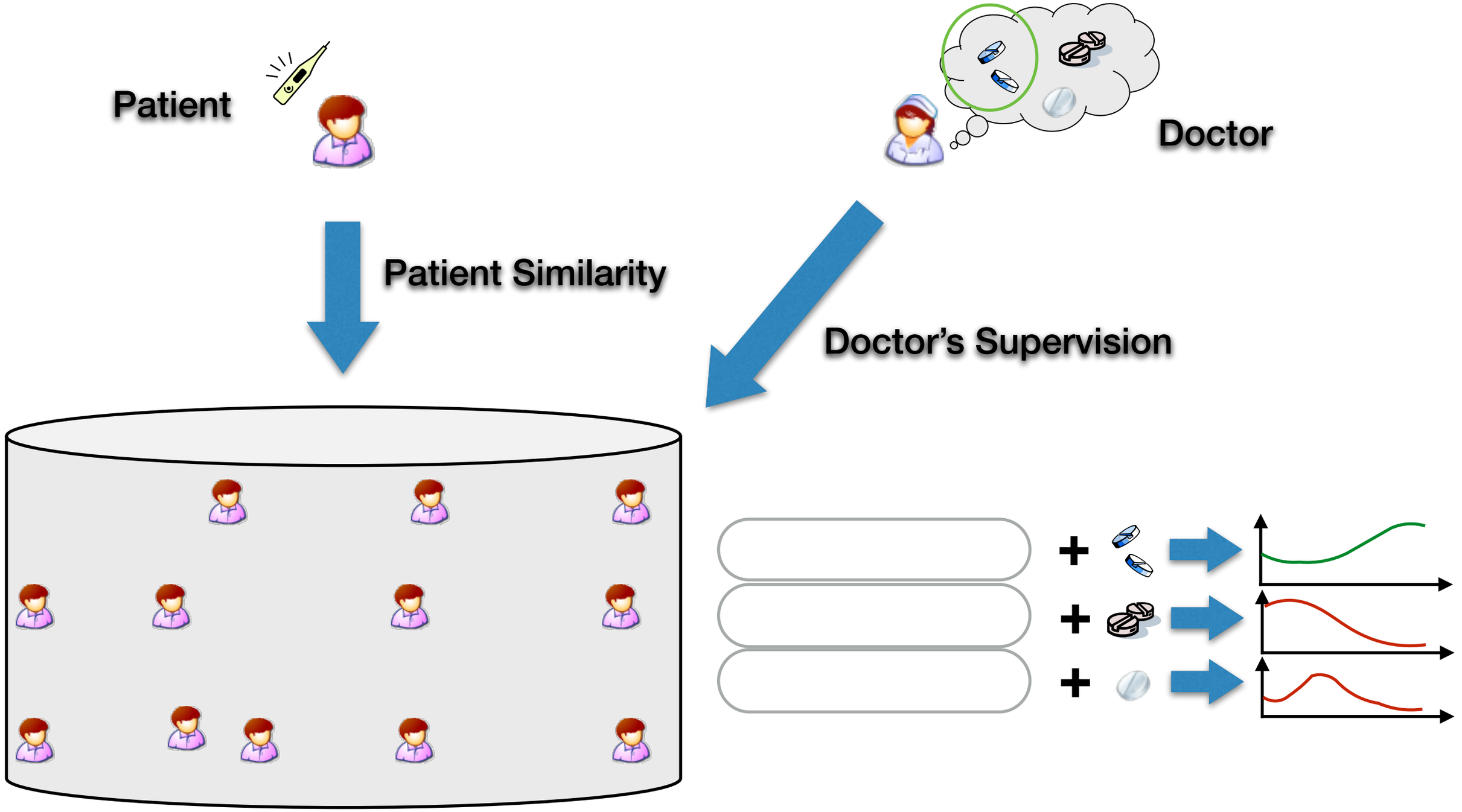




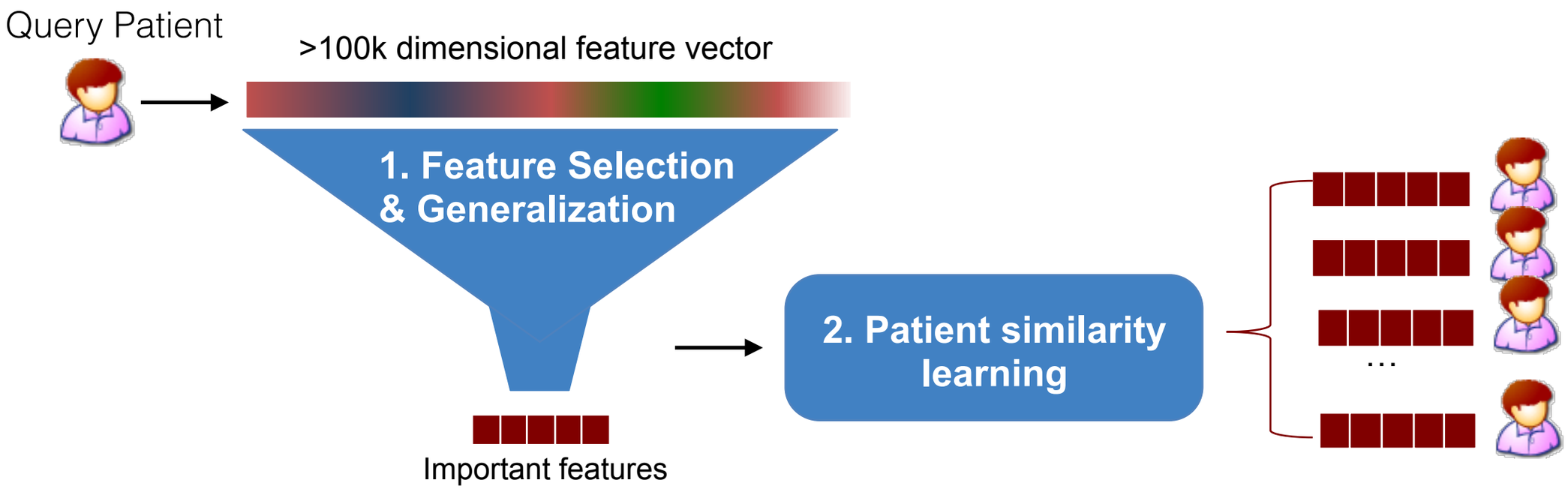


Patient Similarity Problem

Personalized Treatment Recommendation

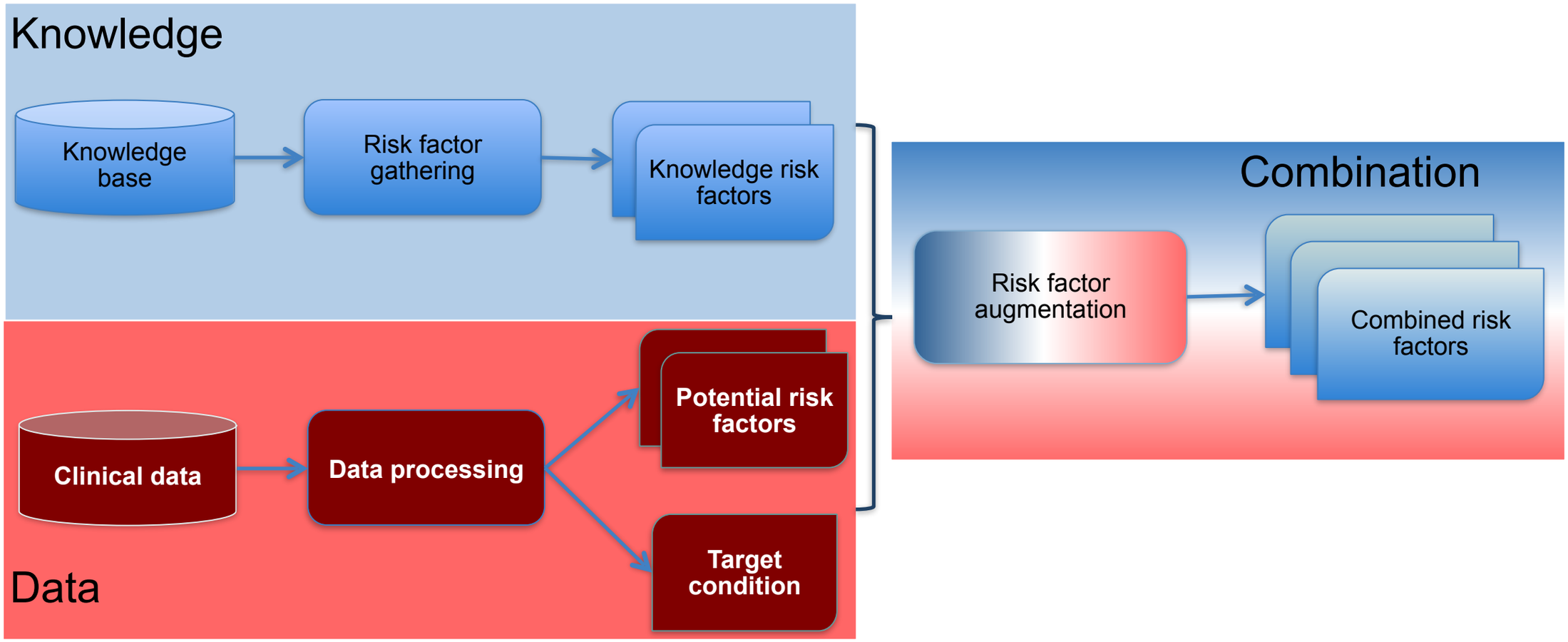


Our Approach



- For a clinical context,
  1. What are important features?
  2. What is the right similarity measure?

Feature Selection



Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Ebadollahi, Steven E. Steinhubl, Zahra Daar, Walter F. Stewart. Combining Knowledge and Data Driven Insights for Identifying Risk Factors using Electronic Health Records. AMIA2012.

Dijun Luo, Fei Wang, Jimeng Sun, Marianthi Markatou, Jianying Hu,Shahram Ebadollahi, SOR: Scalable Orthogonal Regression for Low-Redundancy Feature Selection and its Healthcare Applications. SDM'12

Scalable Orthogonal Regression

- Incorporating knowledge driven risk factors
- Accurate prediction
- Minimal redundancy:
  - Little correlation between the selected data driven risk factors and existing knowledge driven risk factor
  - Little correlation among the additional risk factors from data, to further ensure quality of the additional factors

$$f(\boldsymbol{\alpha}) = \mathcal{L}(\mathbf{y}, \mathbf{X}\boldsymbol{\alpha}) + \frac{\beta}{4}\left[\sum_{ij\in\mathcal{D}}\left(\alpha_i \mathbf{x}_i^T \mathbf{x}_j \alpha_j\right)^2 + \sum_{i\in\mathcal{D}, j\in\mathcal{K}}\left(\alpha_i \mathbf{x}_i^T \mathbf{x}_j \alpha_j\right)^2\right] + \lambda\|\boldsymbol{\alpha}\|_1$$

Model Accuracy

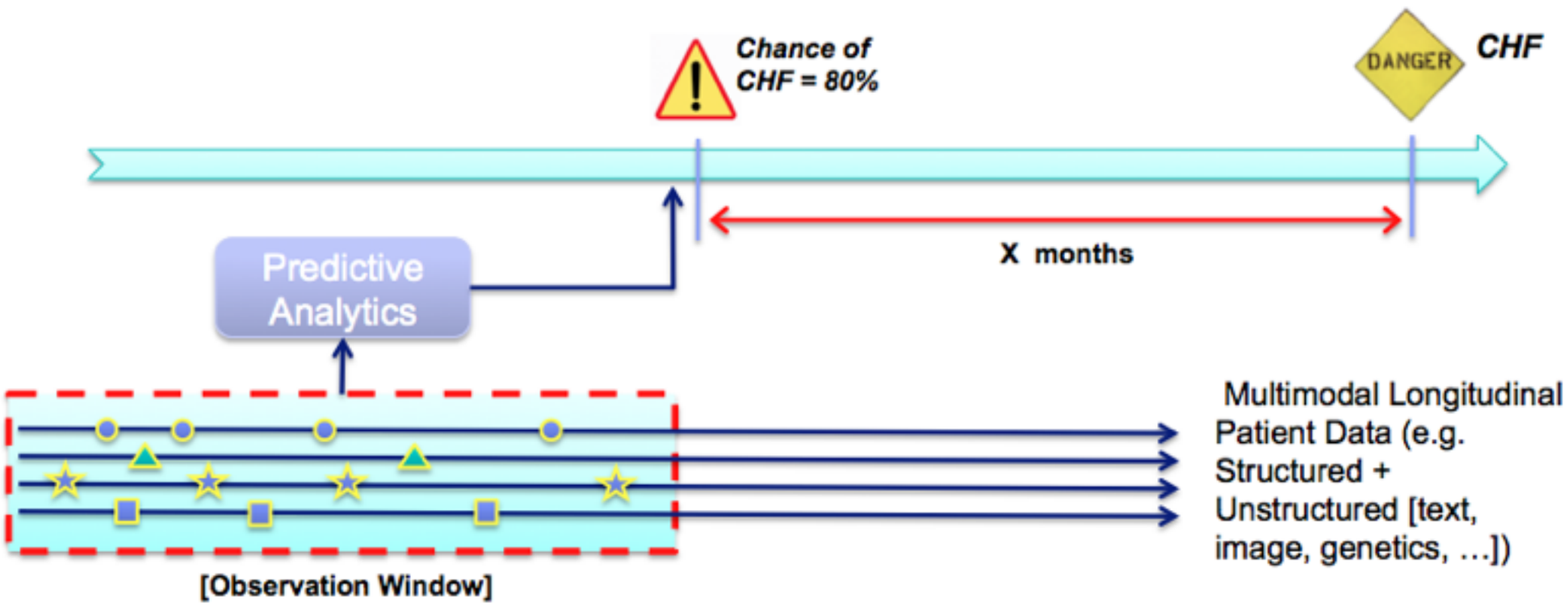
Correlation between data-driven features

Correlation between data- and knowledge-driven features

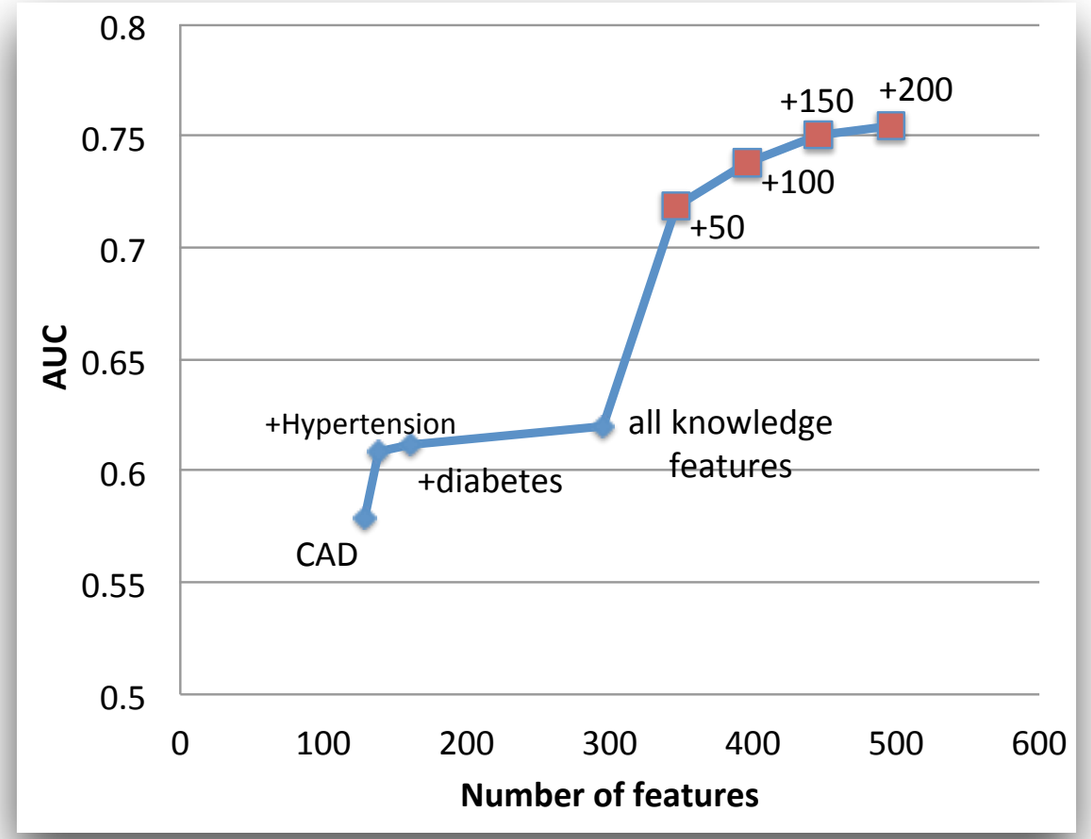
Sparsity Penalty

Early Detection of CHF

- ***Goal***:
  - Build a model for predicting HF onset x months before the HF diagnosis
- ***Data: Longitudinal patient records***
  - Structured data:
    - Demographics, Outpatient diagnoses, Problem List , Vitals, Medication, Labs
  - Unstructured text : encounter notes
- ***Challenge faced by our clinical partners:***
  - How to systematically collect and evaluate many weak and non-specific indicators and identify the ones that combined are truly predictive



CHF Prediction Results



- AUC significantly improves as complementary data driven risk factors are added into existing knowledge based risk factors.
- A significant AUC increase occurs when we add first 50 data driven features

- 4644 case patients, 45,981 control patients
- Over 20k features of different types (diagnoses, demographics, Framingham symptoms, lab results, medication, vital)
- Novel feature selection algorithm enabling integration of knowledge driven and data driven risk factors
- Investigation of different observation windows (30 – 900 days) and prediction windows (1 – 720 days)
- Investigation of multiple classification models (logistic regression, random forest, kNN, cox regression …)

Announcement (10/9/13)

NIH grant on early CHF detection

$2 Million Awarded to Sutter Health, IBM and Geisinger Health System to Study Heart Failure Prediction

Three-year collaboration will develop groundbreaking, big data analytic methods to improve care and reduce costs for treatment of heart disease

Top 10 Features

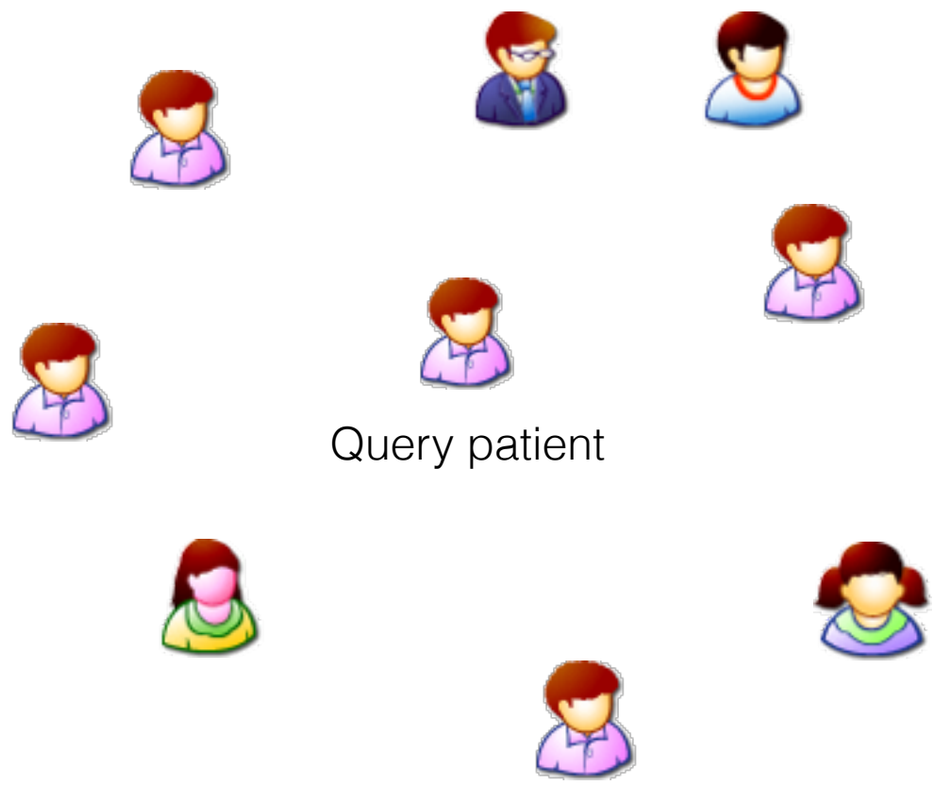
Category
Diagnosis
Medication
Lab
Symptom

Feature	Relevancy to HF
Dyslipidemia	✓
Thiazides-like Diuretics	✓
Antihypertensive Combinations	✓
Aminopenicillins	✓
Bone density regulators	✗
Naturietic Peptide	✓
Rales	✓
Diuretic Combinations	✓
S3Gallop	✓
NSAIDS	✓

- 9 out of 10 are considered relevant to HF
- The data driven features are complementary to the existing knowledge-driven features

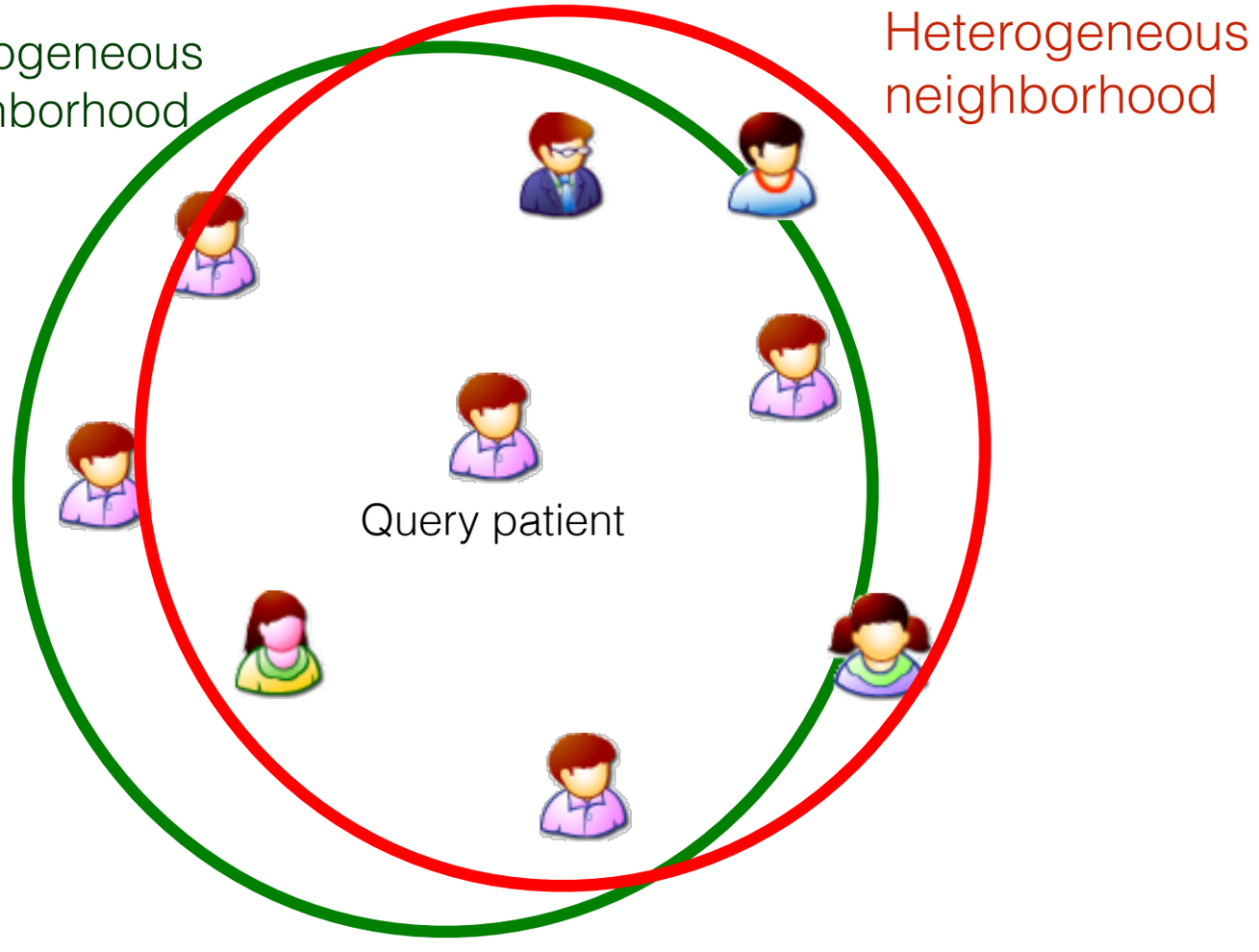
Patient Similarity

Under a specific clinical context



Patient Similarity

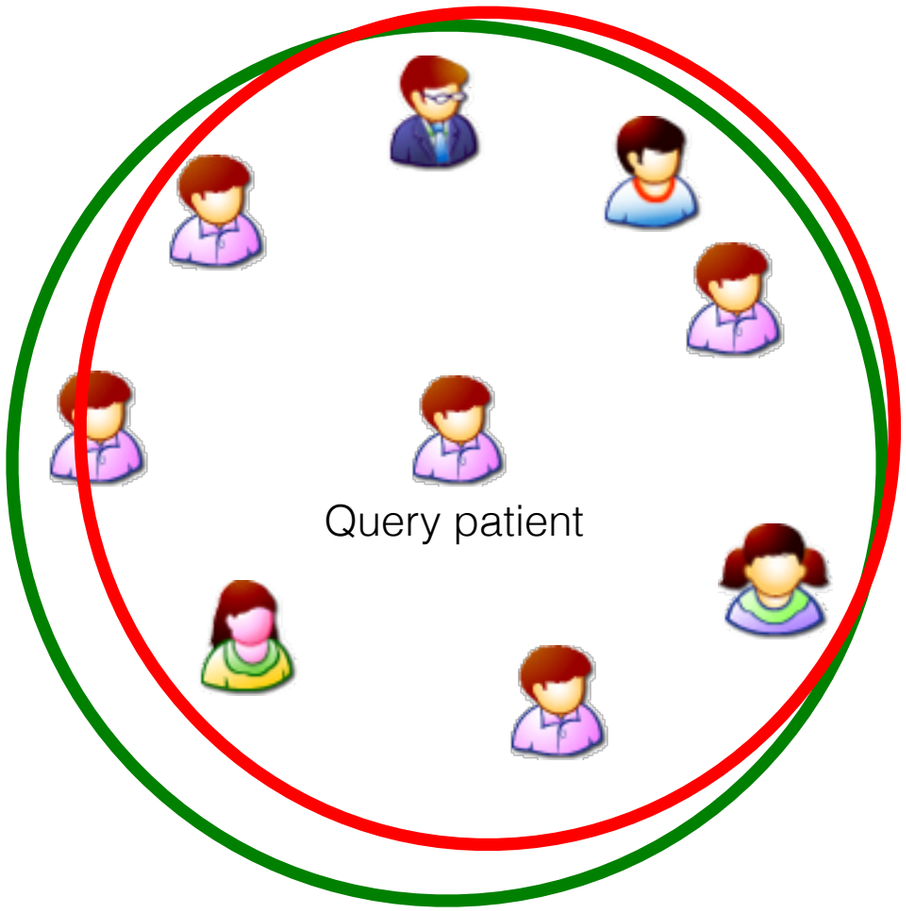
Under a specific clinical context



- Homogeneous neighbors: true positives
- Heterogeneous neighbors: false positives

Patient Similarity

Under a specific clinical context



- Shrink homogeneous neighborhood
- Grow heterogeneous neighborhood

Locally Supervised Metric Learning

Goal: Learn a generalized Mahalanobis distance for a specific clinical context (target label)

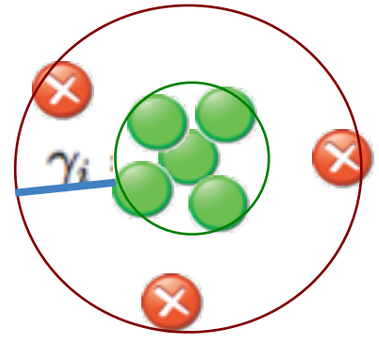
$$d_{\Sigma}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^{\top} \boldsymbol{\Sigma} (\mathbf{x}_i - \mathbf{x}_j)} \qquad \boldsymbol{\Sigma} = \mathbf{W}\mathbf{W}^{\top}$$

Margin for x_i

$$\gamma_i = \sum_{k:\mathbf{x}_k \in \mathcal{N}_i^e} \left\| \mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_k \right\|^2 - \sum_{j:\mathbf{x}_j \in \mathcal{N}_i^o} \left\| \mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j \right\|^2$$

Total distance to heterogeneous neighbors

Total distance to homogeneous neighbors



$\mathcal{N}_i^o$ Homogeneous neighborhood for x_i

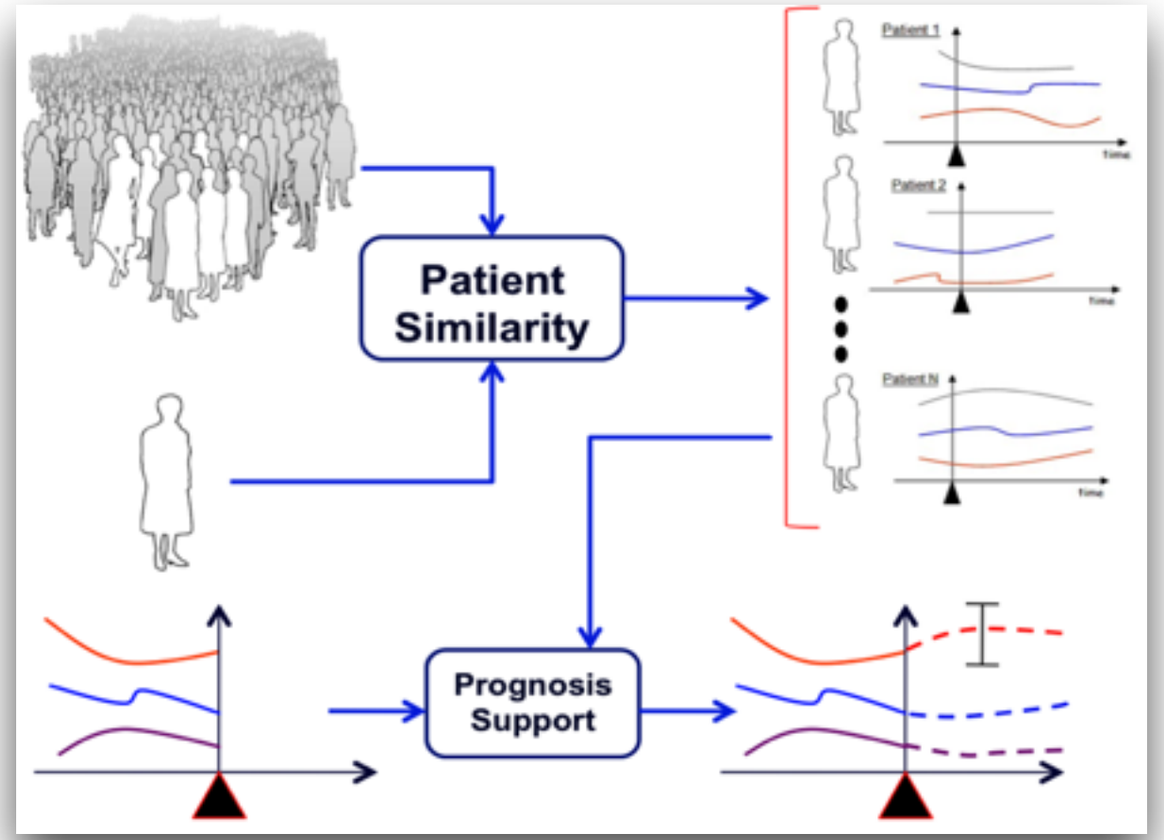
$\mathcal{N}_i^e$ Heterogeneous neighborhood for x_i

Maximize the total margin

$$\gamma = \sum_i \sum_{k:\mathbf{x}_k \in \mathcal{N}_i^e} \mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_k)(\mathbf{x}_i - \mathbf{x}_k)^T \mathbf{W} - \sum_i \sum_{j:\mathbf{x}_j \in \mathcal{N}_i^o} \mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}$$

Fei Wang. Semi-Supervised Metric Learning by Maximizing the Constraint Margin. IEEE TSMC-B. 2011.

Near-Term Prognostication in ICU



Physiological Streams

ABP, systolic ABP, diastolic ABP, Sp02 heart rate

Patients

(H Group) 590 patients experienced at least one occurrence of Acute Hypotensive Episode (AHE),
(C Group) 910 patients did not

Data

2 hours centered around decision time T0
(H Group) T0 is one hour before AHE
(C Group) T0 is randomly chosen

Results

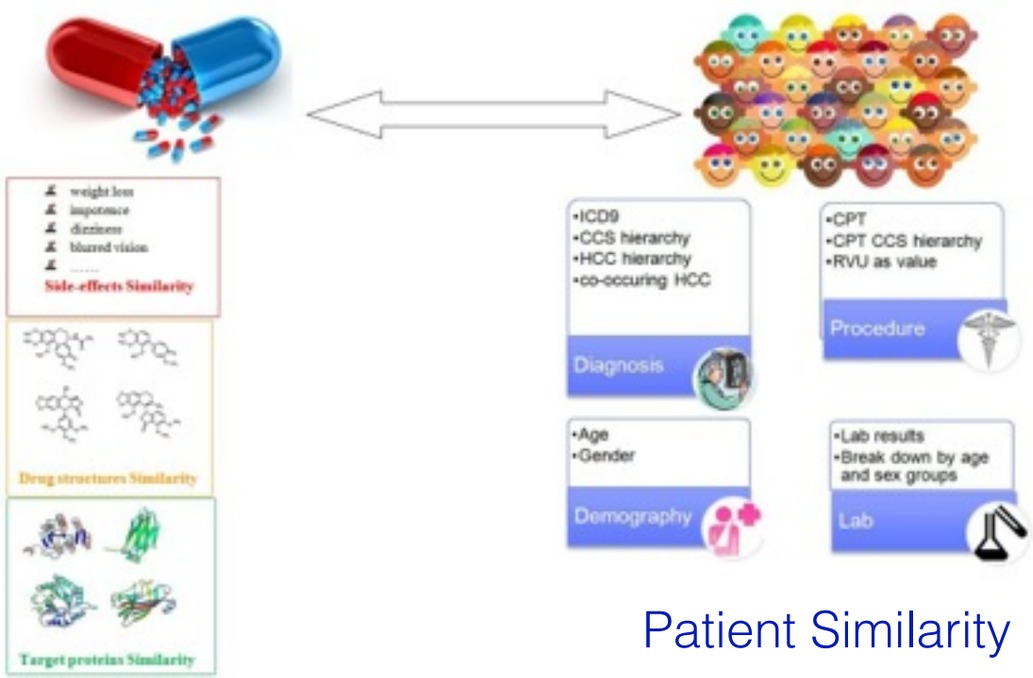
(Challenge09) 0.4982 (LSML) 0.8551

SP02 measures the amount of oxygen affixed to hemoglobin cells within the circulatory system.

(Challenge09) X. Chen, D. Xu, G. Zhang, and R. Mukkamala. Forecasting acute hypotensive episodes in intensive care patients based on peripheral arterial bloom pressure waveform. Computers in Cardiology. 2009.

(LSML) Shahram Ebadollahi, Jimeng Sun, David Gotz, Jianying Hu, Daby Sow, and Chalapathy Neti. Predicting patient's trajectory of physiological data using temporal trends in similar patients: A system for Near-Term prognostics. AMIA 2010. (Best Paper Finalist)

Personalized Medicine

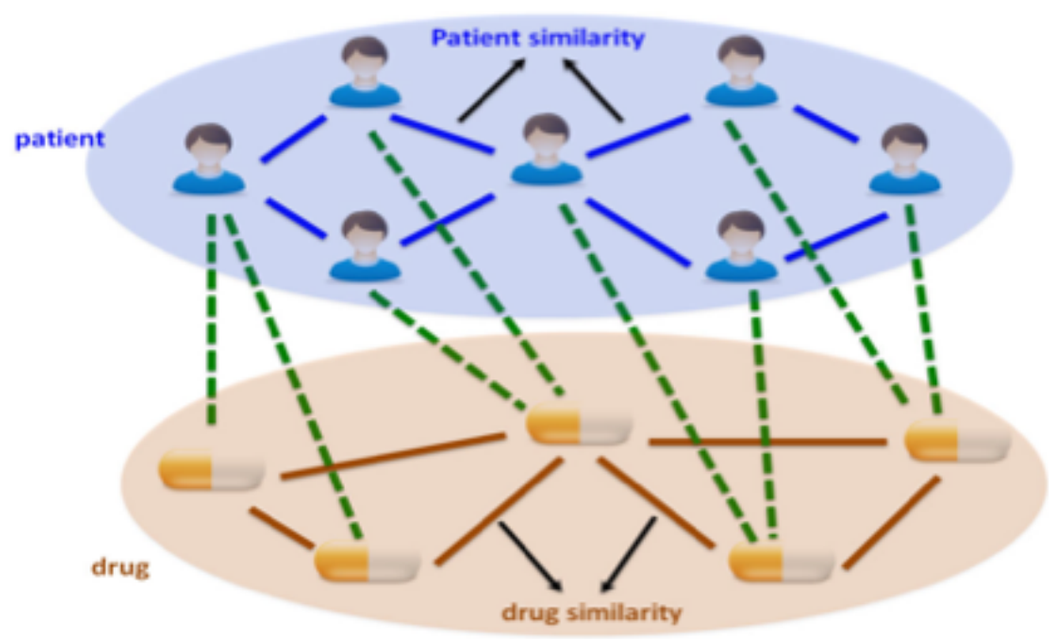


Patient Similarity

Drug Similarity

Personalized Efficacy Prediction

Label Propagation Over Heterogeneous Network

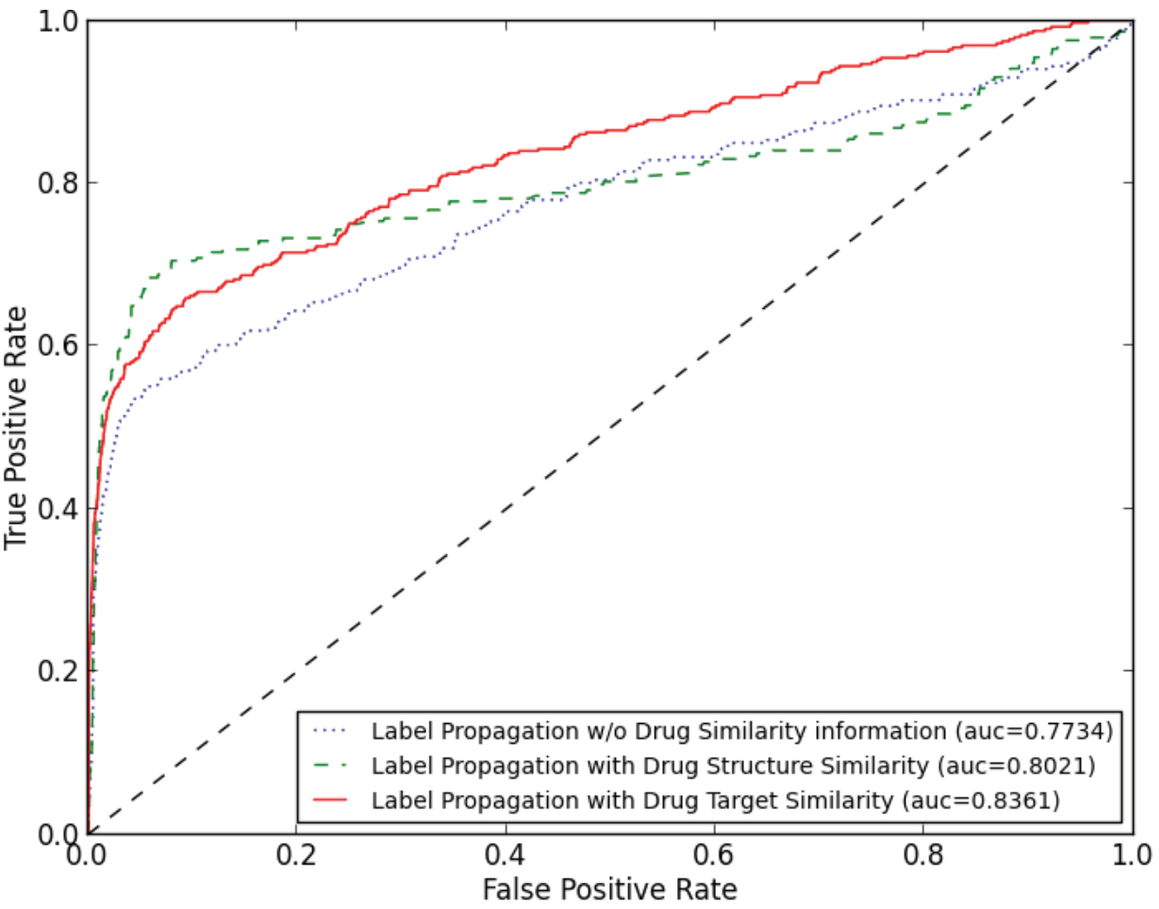


Similarity Network Approach

- Learn from limited number of positive labels
- Propagate from better known patients/drugs to less knows ones

Ping Zhang, Fei Wang, Jianying Hu, and Robert Sorrentino. Towards Personalized Medicine: Leveraging Patient Similarity and Drug Similarity Analytics. AMIA Translational Bioinformatics Summit. 2014. (Finalist for Marco Romani Distinguished Paper Award)

Personalized Medicine

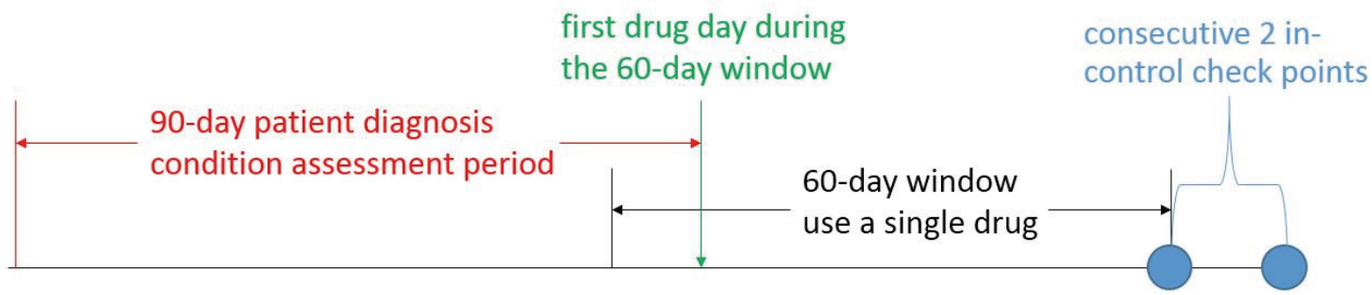


Condition: Dyslipidemia

Criterion

Low-Density Lipoprotein (LDL) level is below 130 mg/dL, is considered to be "well-controlled"

Evaluation Period



Data

1219 Distinct Patients

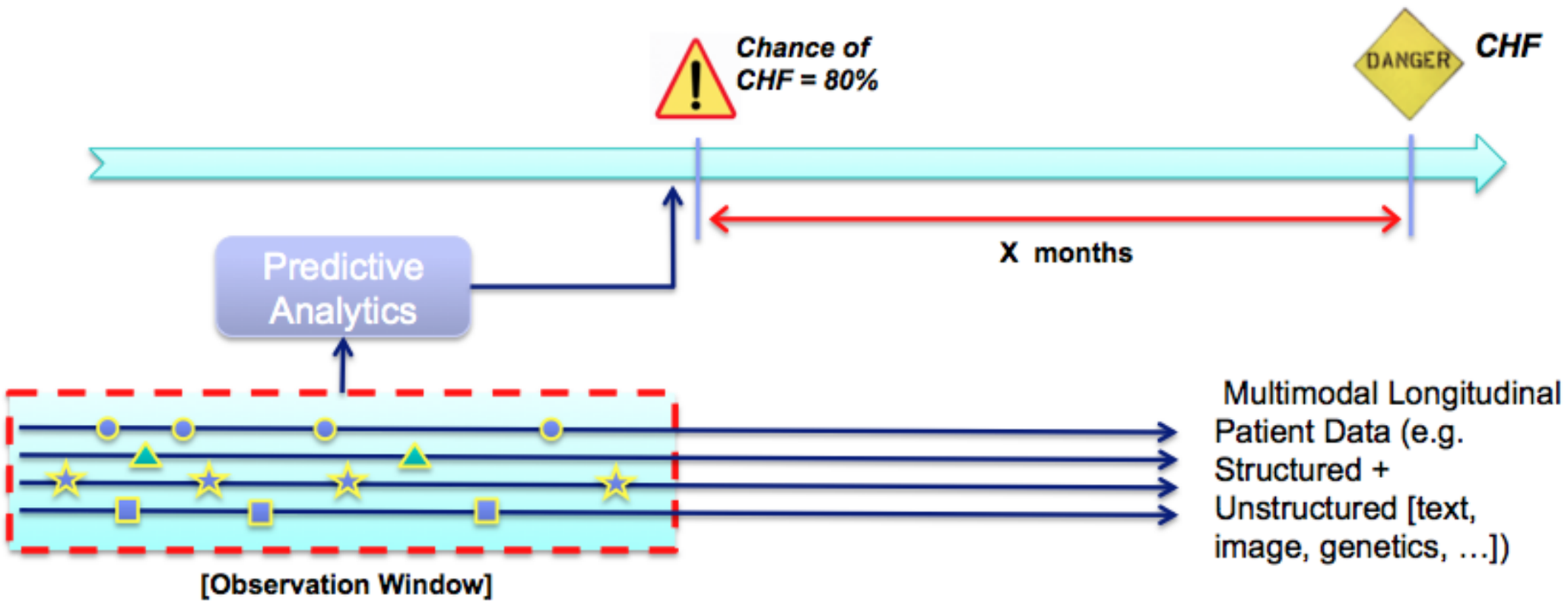
4 Drugs: Atorvastatin, Lovastatin, Pravastatin, Simvastatin

Ping Zhang, Fei Wang, Jianying Hu, and Robert Sorrentino. Towards Personalized Medicine: Leveraging Patient Similarity and Drug Similarity Analytics. AMIA Translational Bioinformatics Summit. 2014. (Finalist for Marco Romani Distinguished Paper Award)

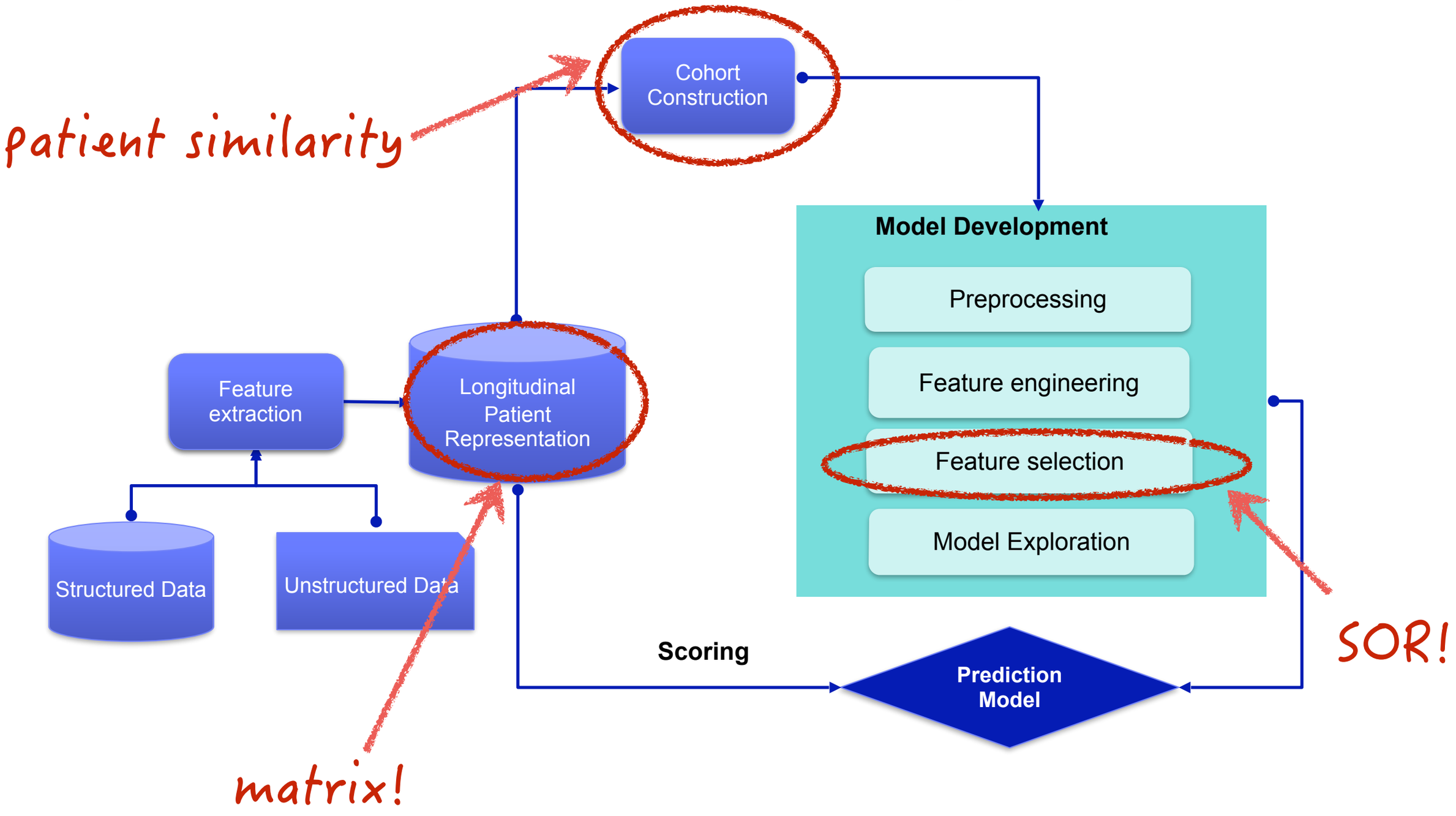
Roadmap

- Background
- Electronic Health Records
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Predictive Modeling Scenario Revisited



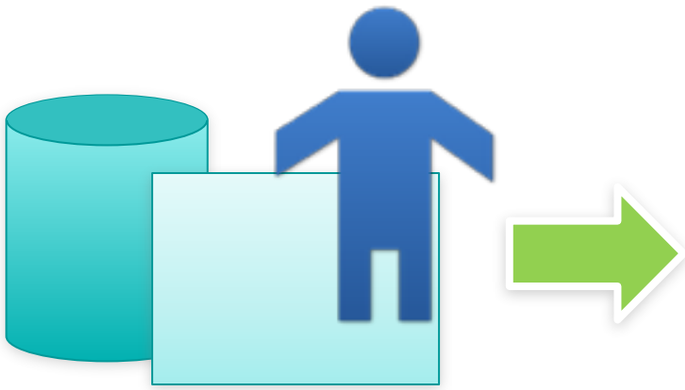
Predictive Modeling Pipeline



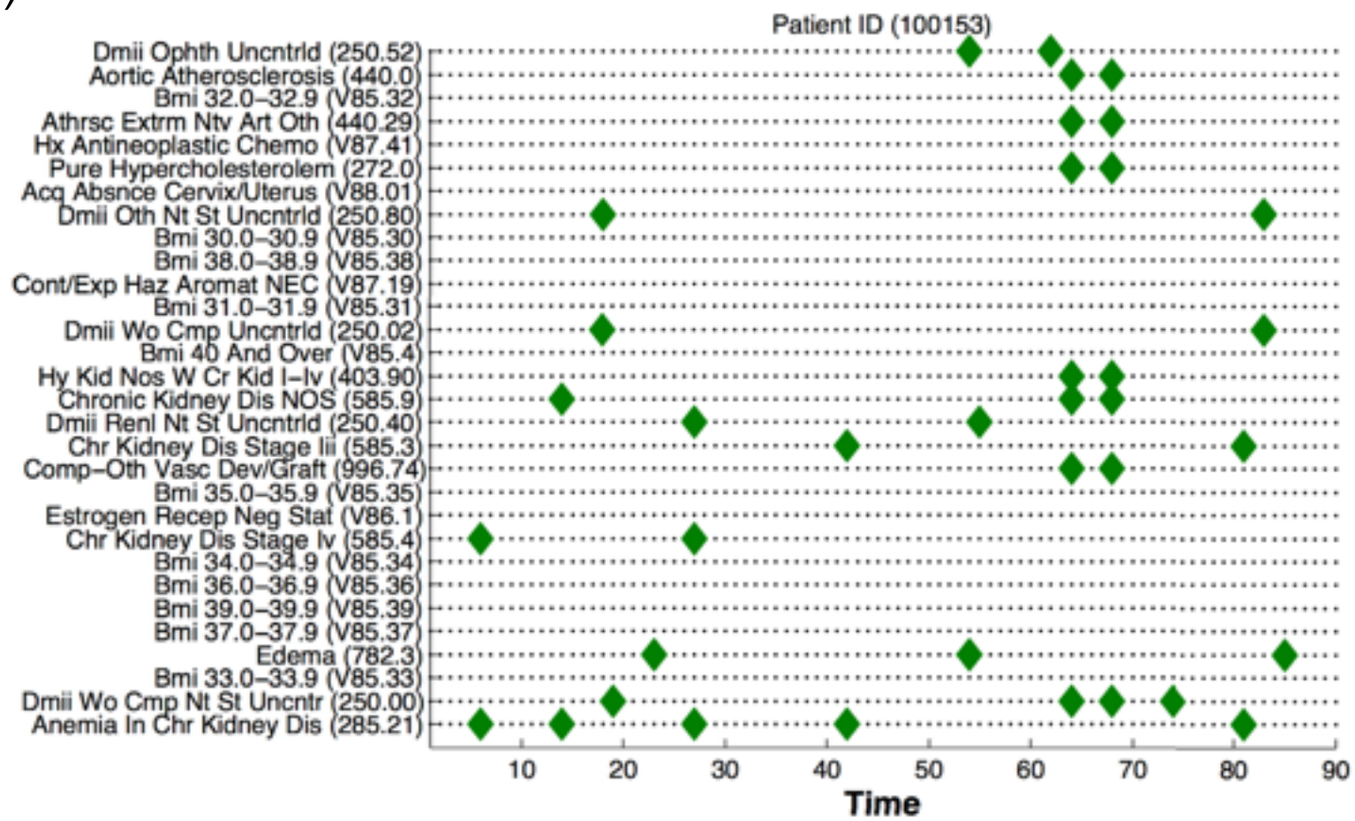
EMR Sparsity

- Longitudinal patient matrices.
- Challenges
  - Different start time, end time, and length of records.
  - Extremely sparse (low-density)

Diagnosis Codes
CHF Cohort Density
0.0034%



Patient
Longitudinal
EMR Data



F. Wang et al. Towards Heterogeneous Temporal Clinical Event Pattern Discovery:A Convolutional Approach. *KDD 2012*.

Is Sparsity Good?

Sparsity may cause problems in phenotyping patients.

Patient 1	…	t	t+1	t+2	…
Heart Failure		X		X	
Head Injury			X		
…					

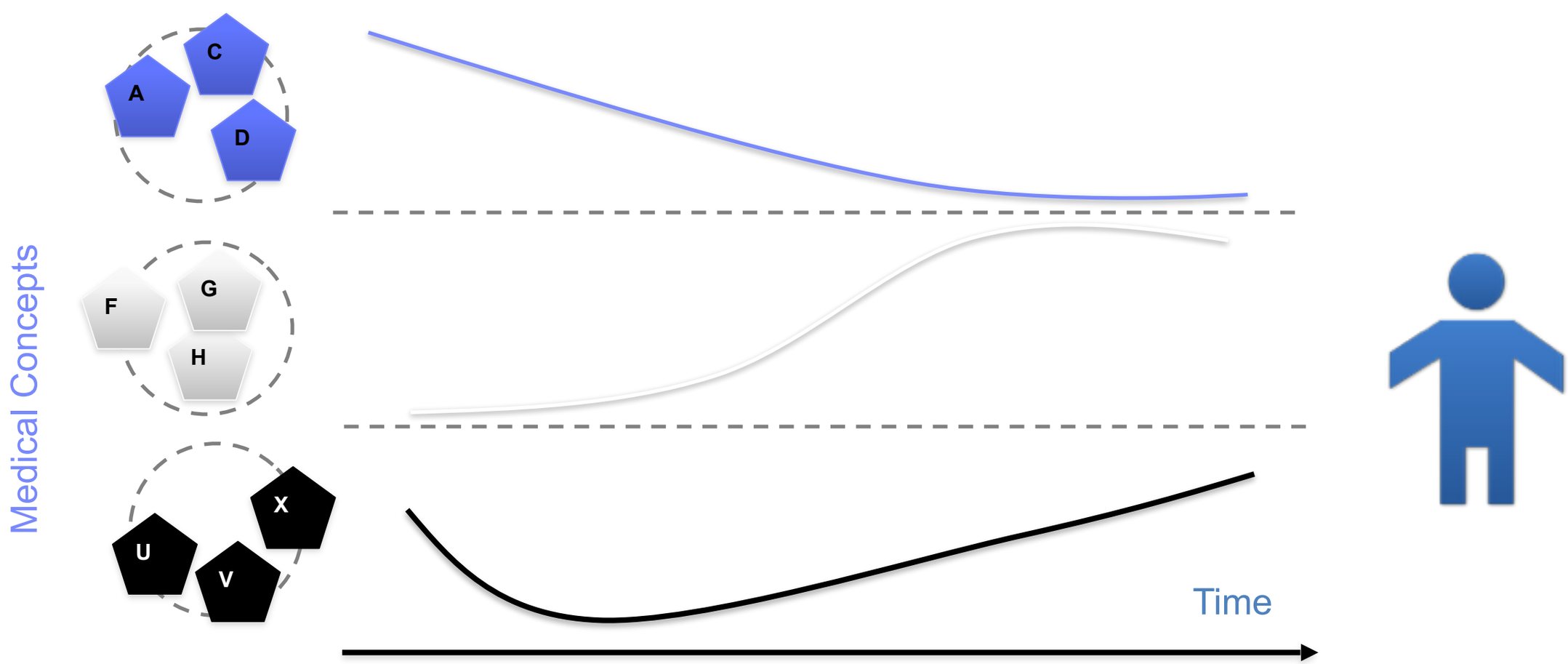
Patient 2	…	t	t+1	t+2	…
Heart Failure		X	X	X	
Head Injury					
…					

Questions 1
Is heart failure cured at this time?

Question 2
Does patient 1 have less risk when predicting future heart failure than patient 2?

Medical Concepts

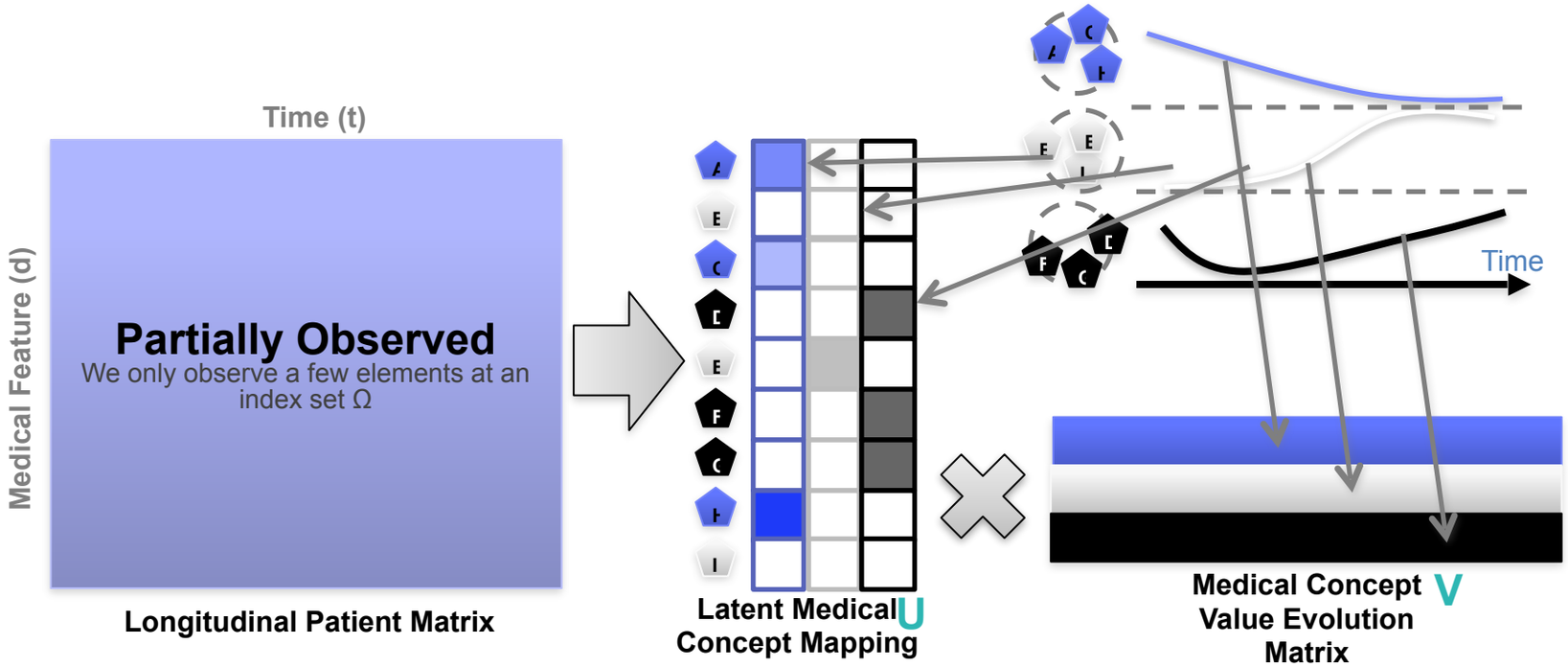
- In the disease progression, biomarkers evolve continuously over time.
- We assume that:
  - There exist some high-level ***medical concepts***, which consist of raw medical features.
  - For each patient, the medical concepts ***evolve smoothly***.



Jack Jr, Clifford R., et al. "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade." The Lancet Neurology 9.1 (2010

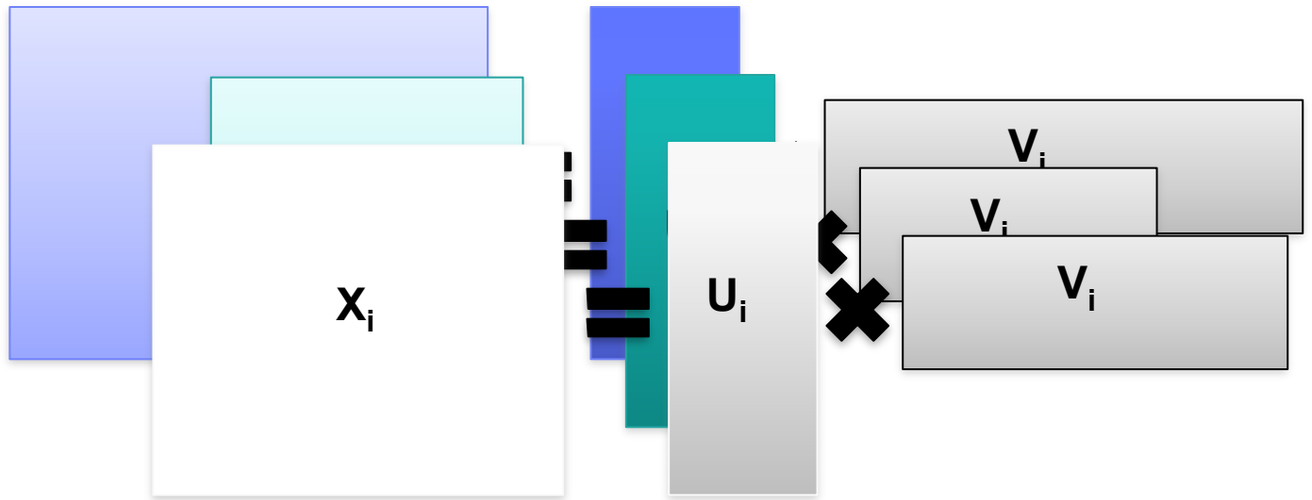
Patient Record Densifier

- Assume the ***full*** longitudinal patient matrix can be approximated by a low rank matrix
  - Macro-phenotype mapping matrix **U**: ***Sparse, Non-negative***
  - Concept value evolution matrix **V**: ***Temporal Smoothness***



Individual Basis Approach

- Assume that each patient has different medical concepts from other patients



- Formulation

$$\min_{\{S_i\},\{U_i\},\{V_i\}} \sum_{i=1}^{n} \frac{1}{2t_i}\|S_i - U_i V_i\|_F^2 + \lambda_1\|U\|_1 + \lambda_2 \sum_{i=1}^{n} \frac{1}{2t_i}\|V_i\|_F^2 + \lambda_3 \sum_{i=1}^{n} \frac{1}{2t_i}\|V_i R_i\|_F^2$$
$$\text{subject to: } \mathcal{P}_{\Omega_i}(S_i) = \mathcal{P}_{\Omega_i}(X_i), U_i \geq 0, \forall i$$

Matrix Completion
Completion via matrix factorization. Enforce a low rank factorization U_i, V_i and encourage the values of $U_i V_i$ at the observed location to be close to the observed ones.

Meaningful Medical Concepts
Medical concepts involve non-negative components of medical features.

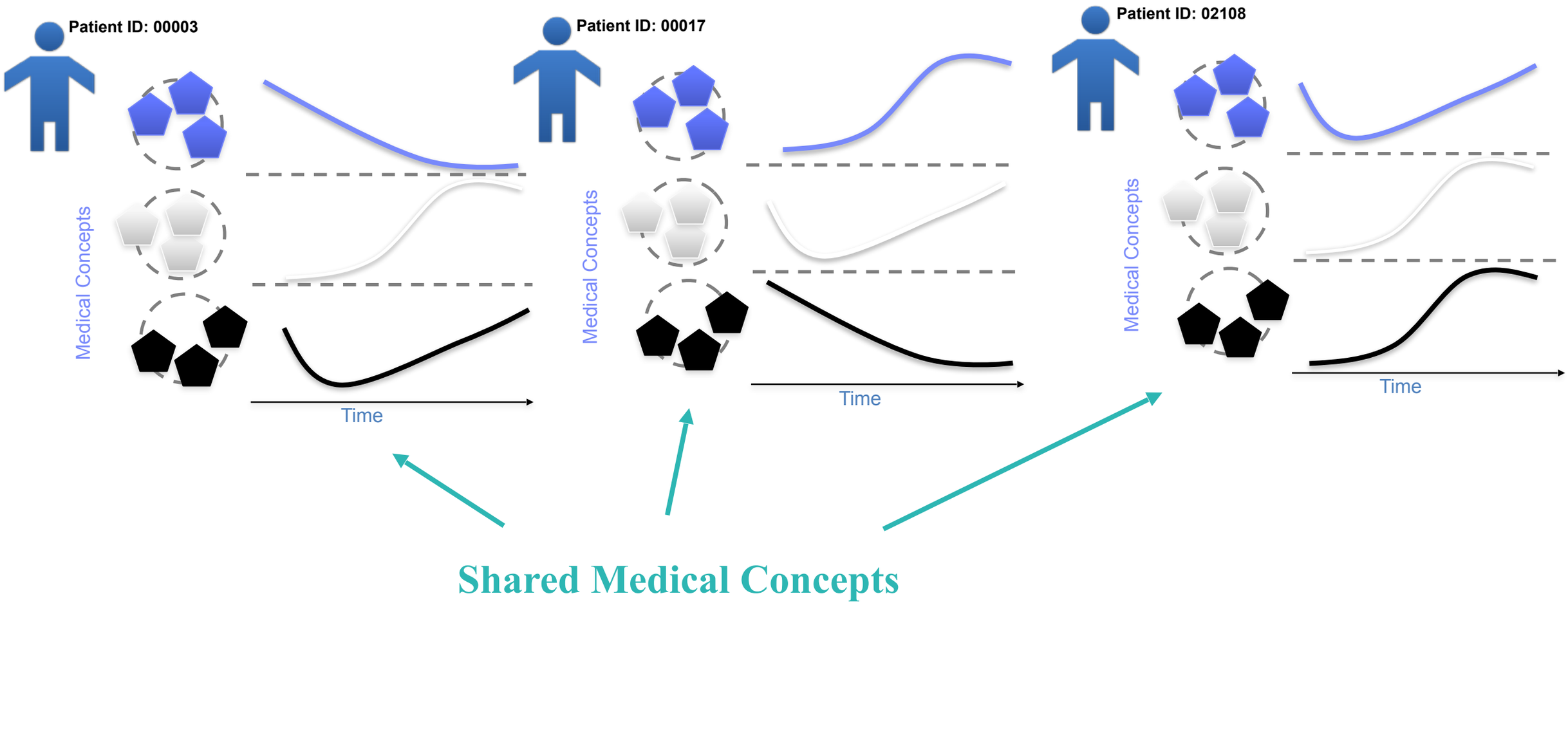
Sparsity
Controllable sparsity that encourages a few medical features in each concept.

Overfitting Control
Prevent V_i from overfitting.

Temporal Smoothness
Couple the columns of V_i and force them to close to each other.

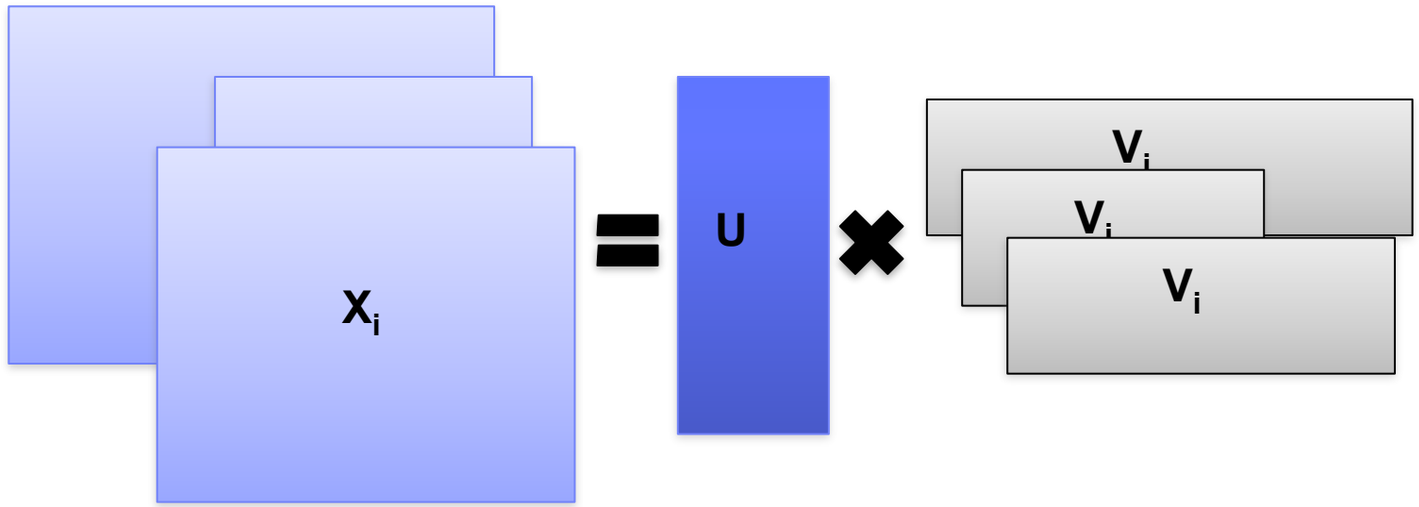
Shared Basis Approach

- If patients are similar to each other, then the patients may share the same set of medical concepts, while each of them may evolve differently.



Shared Basis Approach

- Assume that each patient has shared medical concepts from other patients



- Formulation

Shared Medical Concept Mapping

$$\min_{\{S_i\},U,\{V_i\}} \sum_{i=1}^{n} \frac{1}{2t_i}\|S_i - UV_i\|_F^2 + \lambda_1\|U\|_1 + \lambda_2 \sum_{i=1}^{n} \frac{1}{2t_i}\|V_i\|_F^2 + \lambda_3 \sum_{i=1}^{n} \frac{1}{2t_i}\|V_i R_i\|_F^2$$

$$\text{subject to:} \quad \mathcal{P}_{\Omega_i}(S_i) = \mathcal{P}_{\Omega_i}(X_i), U \geq 0$$

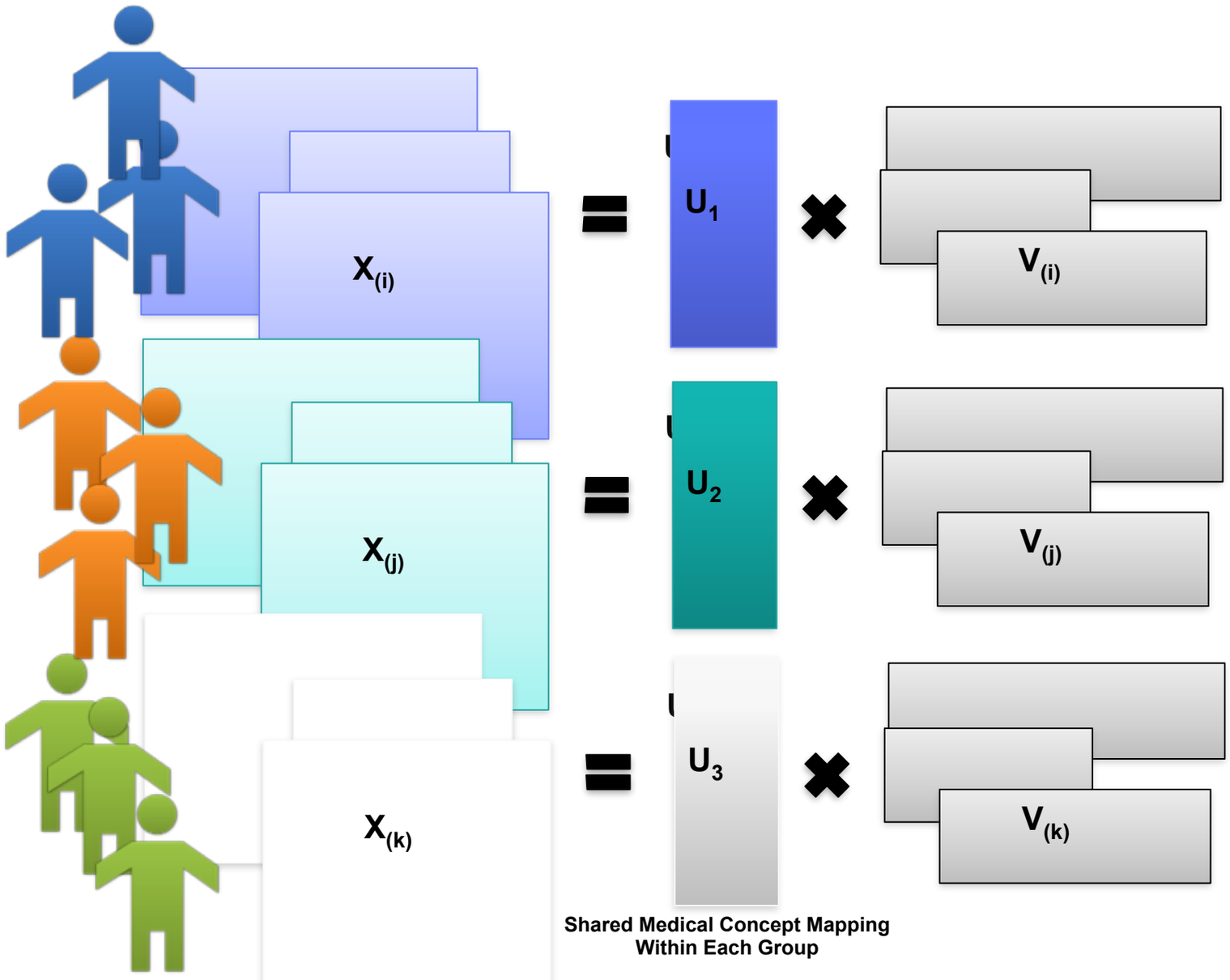
Clustered Basis Approach

- Patients may exhibit some groups
  - Inside each group: homogeneous, share medical concepts.
  - Outside group: heterogeneous
  - Transfer knowledge within each group

- Each group may consist of patients of similar health condition:
  - Affected by the same set of diseases or share comorbidity.
  - Similar in terms of blood type, geographic, age, sex.
  - Leverage external domain knowledge.

Clustered Basis Approach

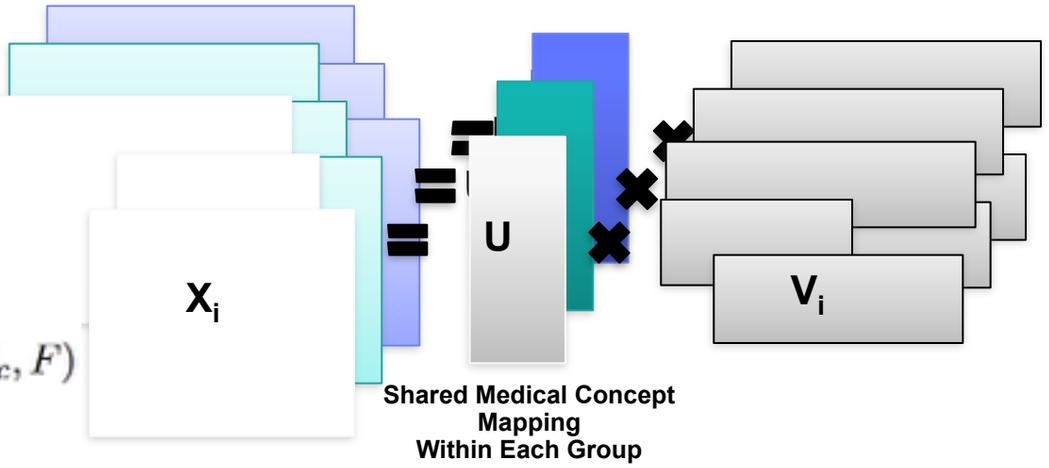
Simultaneously clustering patients and densification



Clustered Basis Approach

- Model-based patient clustering in Pacifier
  - Spectral K-Means objective on medical concepts

$$\min_{F:F^TF=I_p} \frac{1}{2k}\sum_{c=1}^{k}\left(\mathrm{tr}(U_c^TU_c)-\mathrm{tr}(F^TU_c^TU_cF)\right) := \min_{F:F^TF=I_p} \frac{1}{2k}\sum_{c=1}^{k}\mathcal{C}(U_c,F)$$



- Formulation of Pacifier-CBA

Patient Clustering

$$\min_{F,\{U_{(i)},V_{(i)},S_{(i)}\}} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2t_i}\|S_{(i)}-U_{(i)}V_{(i)}\|_F^2 + \lambda_c\frac{1}{2k}\sum_{c=1}^{k}\mathcal{C}(U_c,F) + \mathcal{T}(\{V_{(i)}\})$$

$$\text{subject to: } \mathcal{P}_{\Omega_{(i)}}(S_{(i)}) = \mathcal{P}_{\Omega_{(i)}}(X_{(i)}),\ \{U_{(i)}\}\in\mathcal{U},\ F^TF=I_p$$

Matrix Completion via Factorization

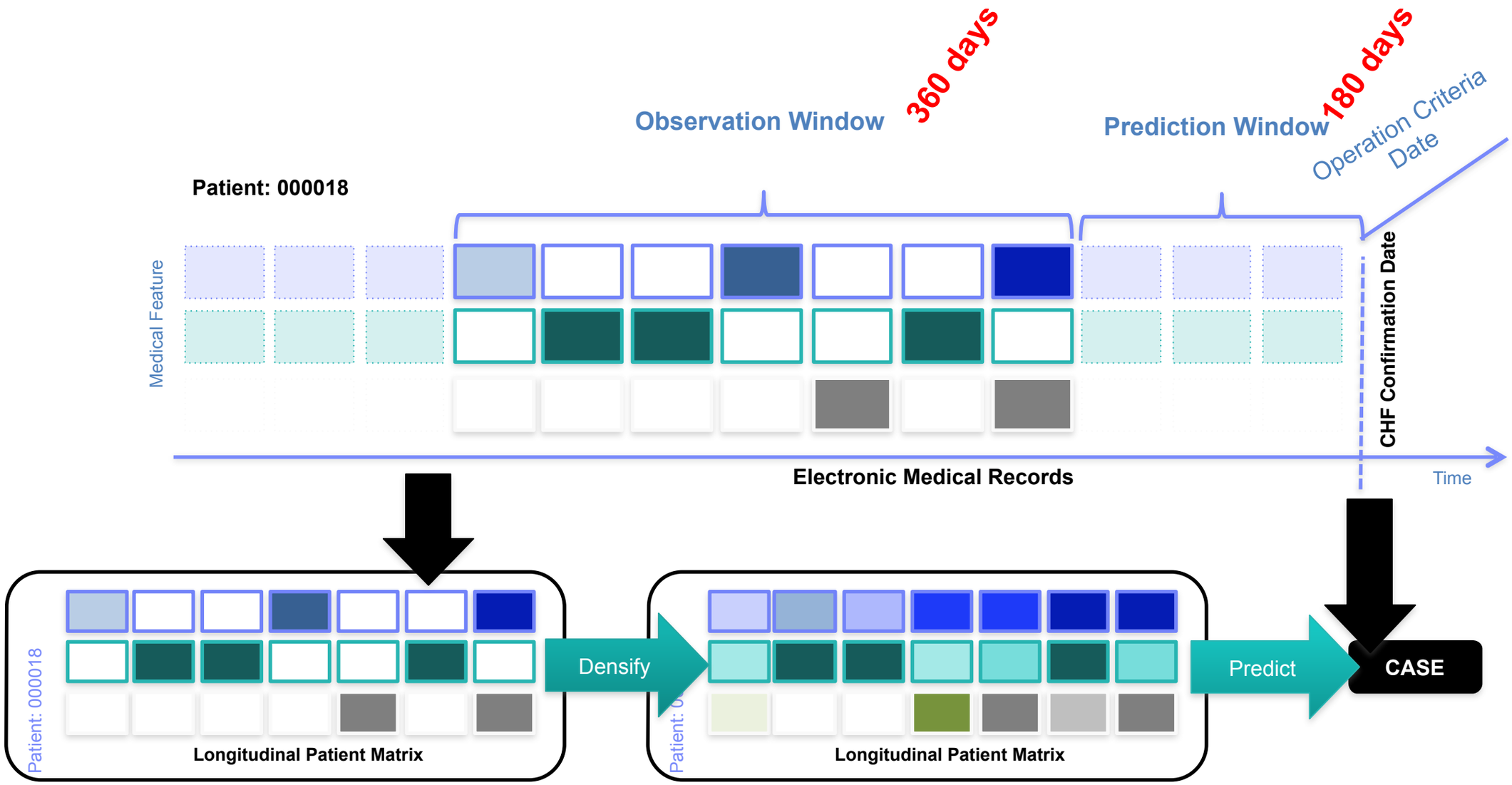
Meaningful Medical Concepts

$$\mathcal{U} := \{U | U \geq 0, \|U\|_1 \leq \tau\}$$

Temporal Smoothness

$$\mathcal{T}(\{V_{(i)}\}) = \lambda_2\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2t_i}\|V_{(i)}\|_F^2 + \lambda_3\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2t_i}\|V_{(i)}R_{(i)}\|_F^2.$$

Application in CHF Onset Prediction



Experimental Settings

Densification Configuration

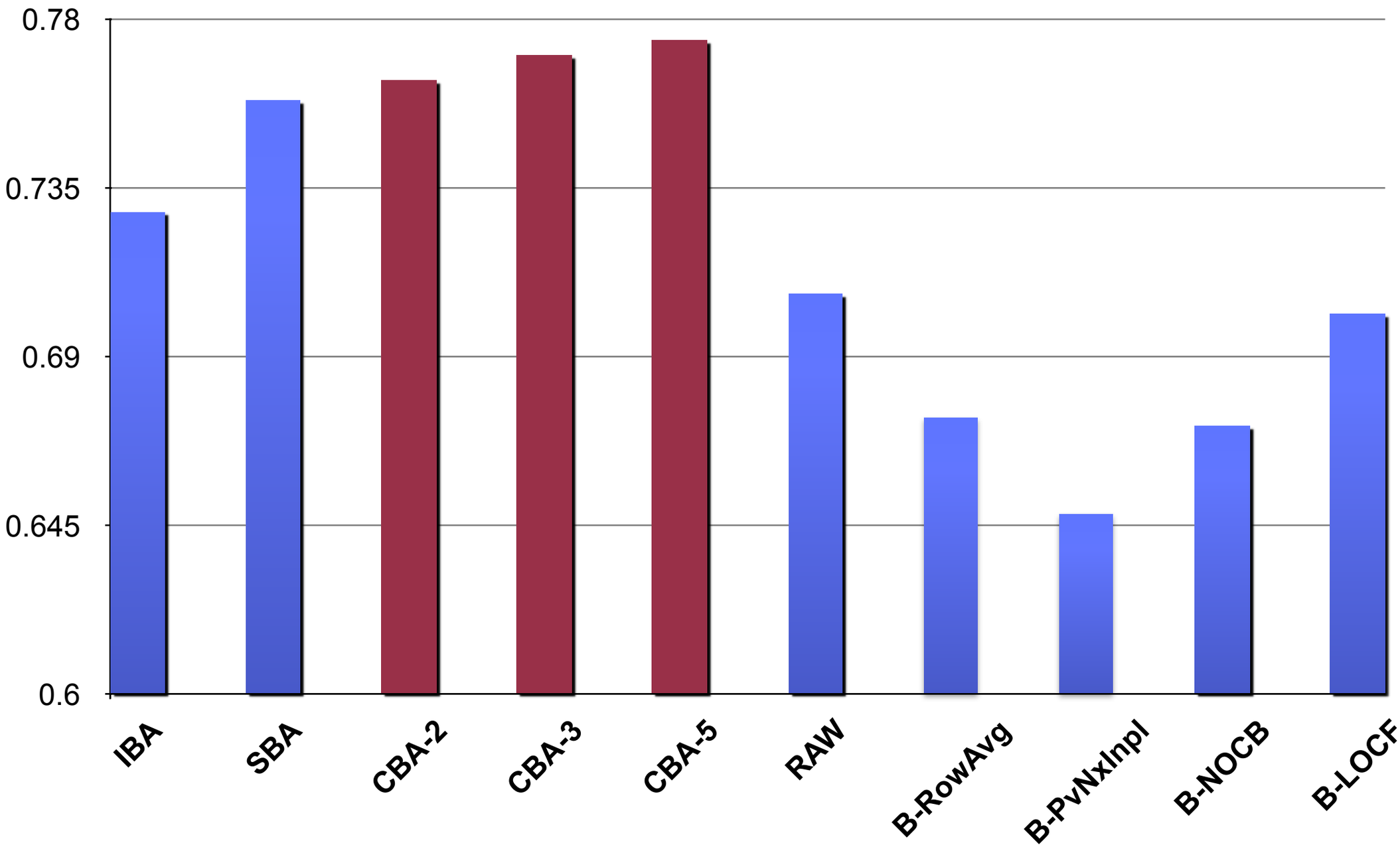
Competing methods
- **Baseline - Zero Imputation (RAW)**
- **Basic Imputation methods**
  - **Row Average (RowAvg)**
  - **Next Occurrence Carry Backward (NOCB)**
  - **Last Occurrence Carry Forward (LOCF)**
  - **Interpolation of Previous and Next (PvNxInpl)**

Parameter selection
- **10-fold cross-validation**

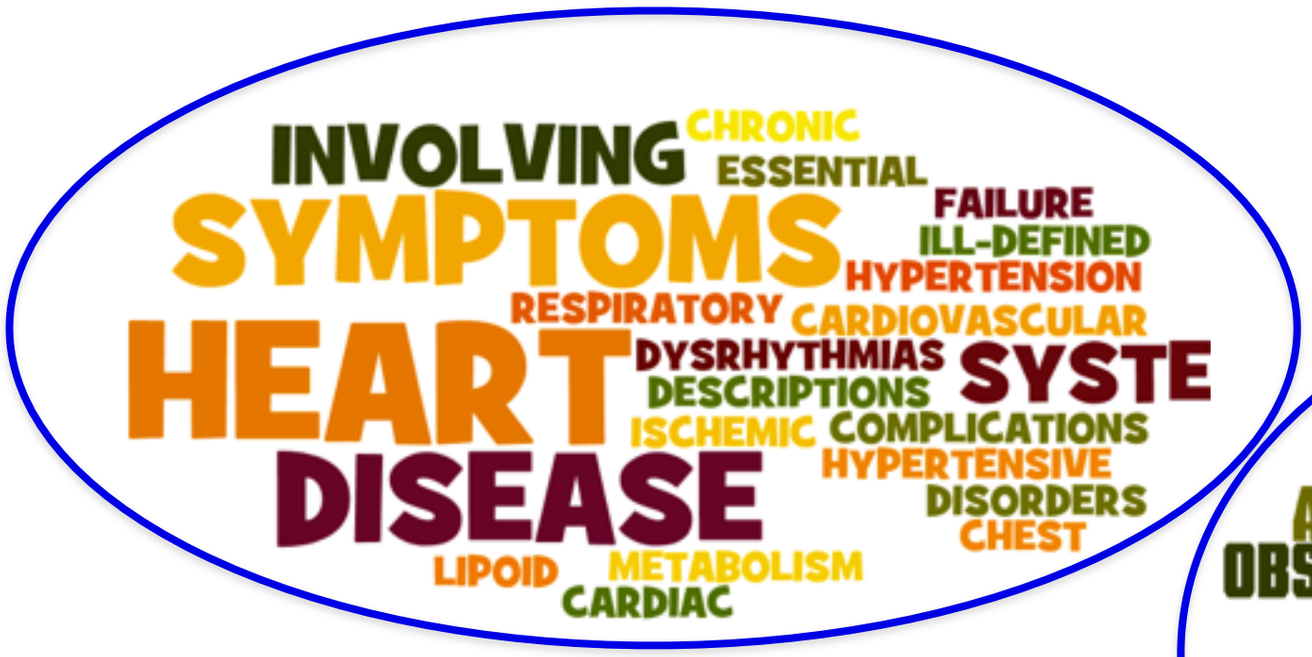
Classification Configuration

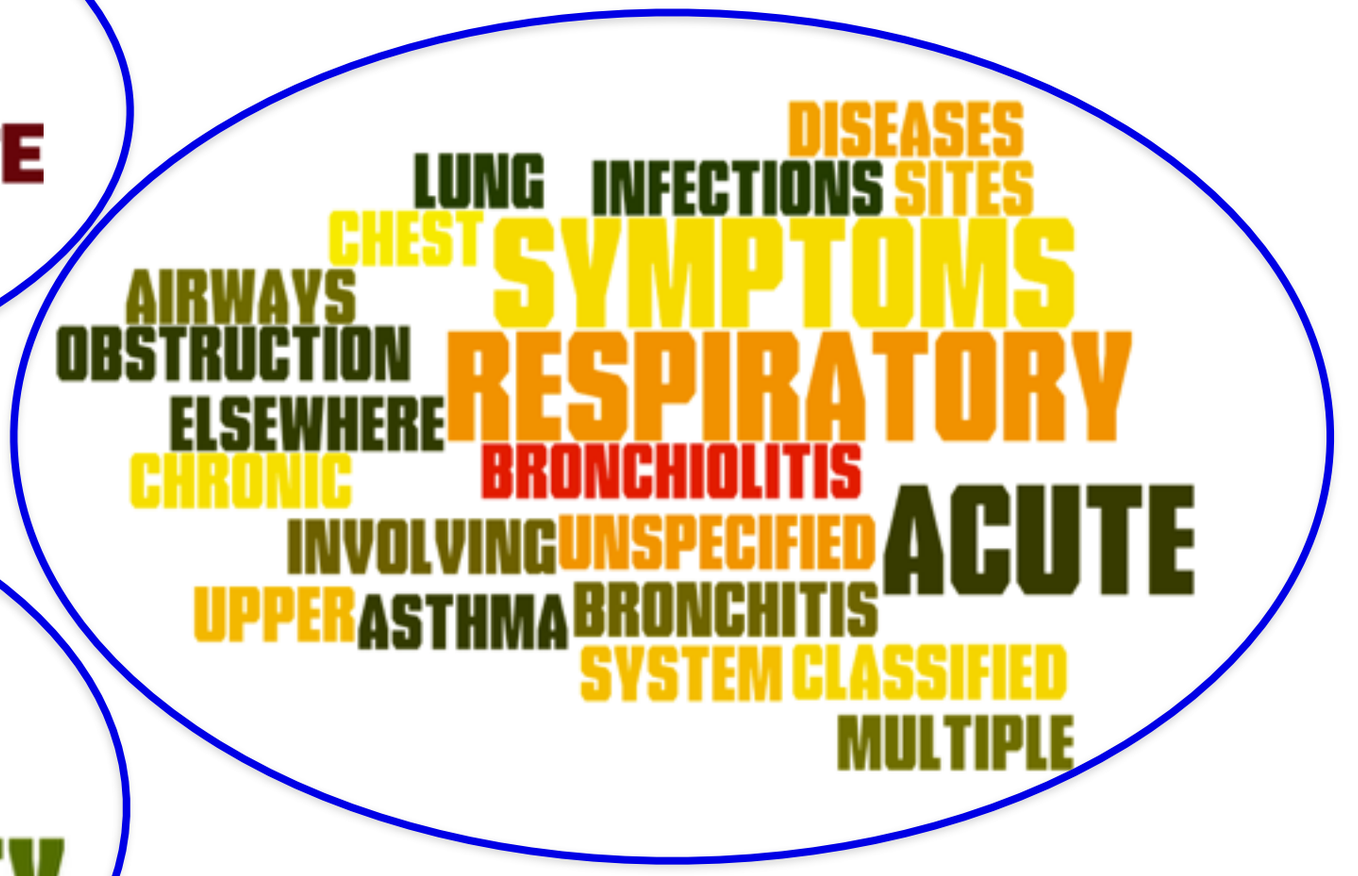
Training: 90% Percent
Classifier: sparse logistic regression
Model selection scheme: 5-fold cross validation
Metric: AUC

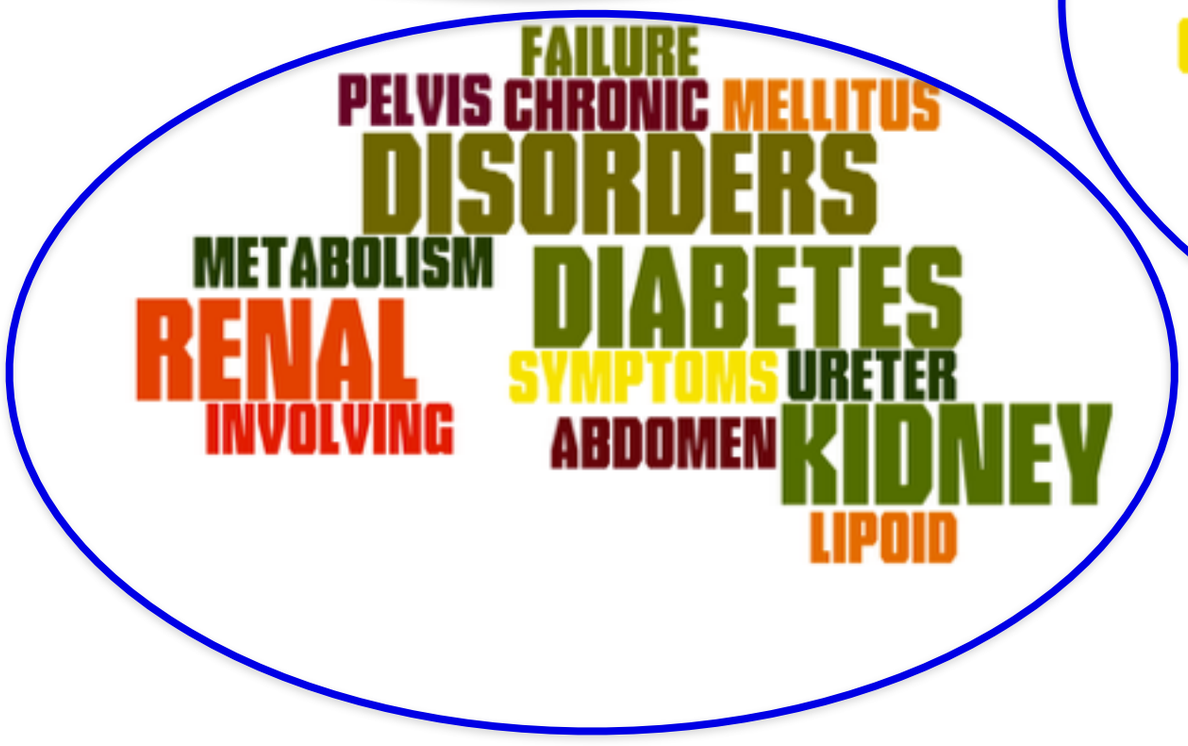
Performance



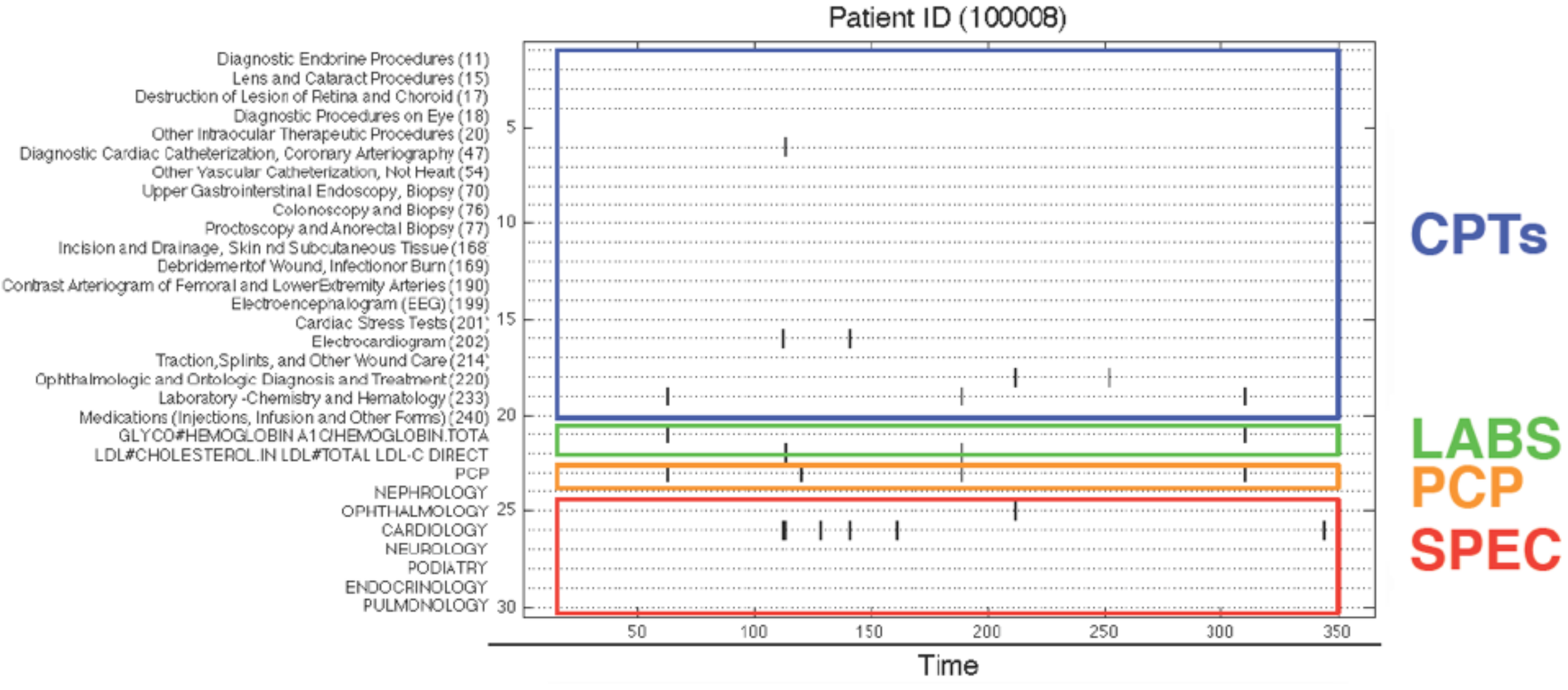
Learned Medical Concepts



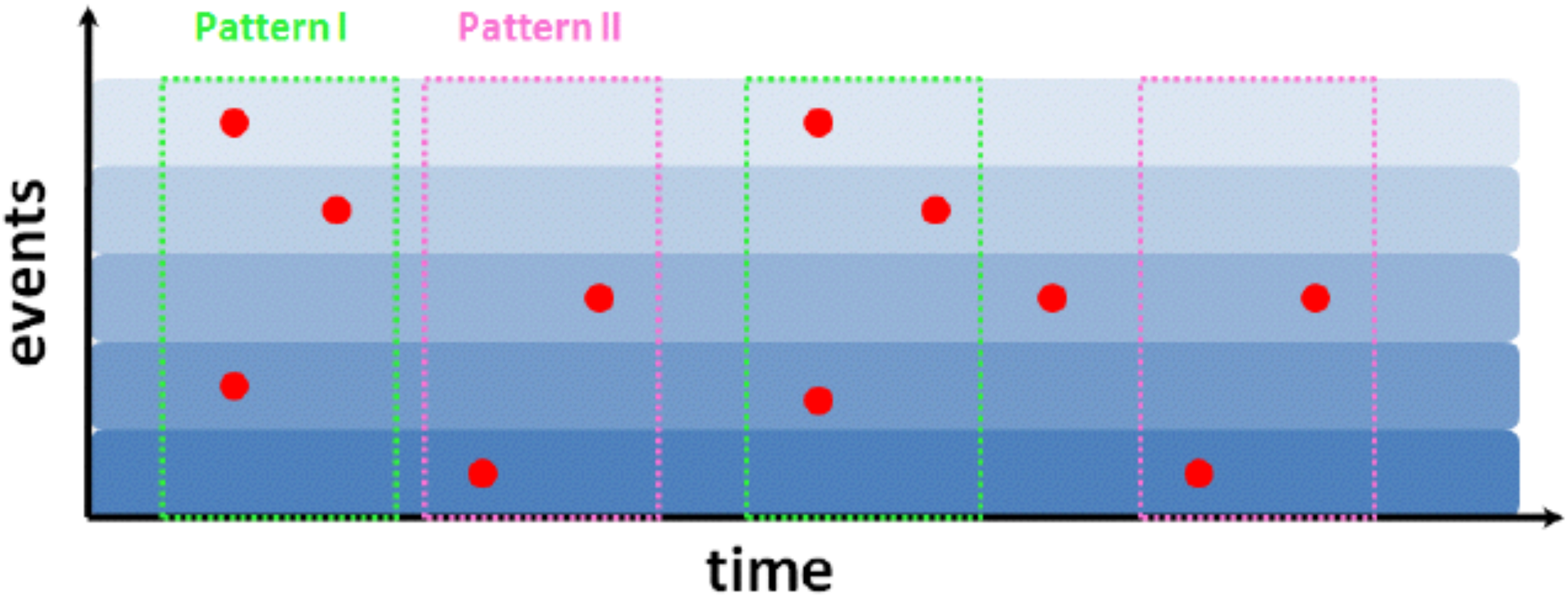




Matrix Representation of EMR Revisited



Temporal Patterns



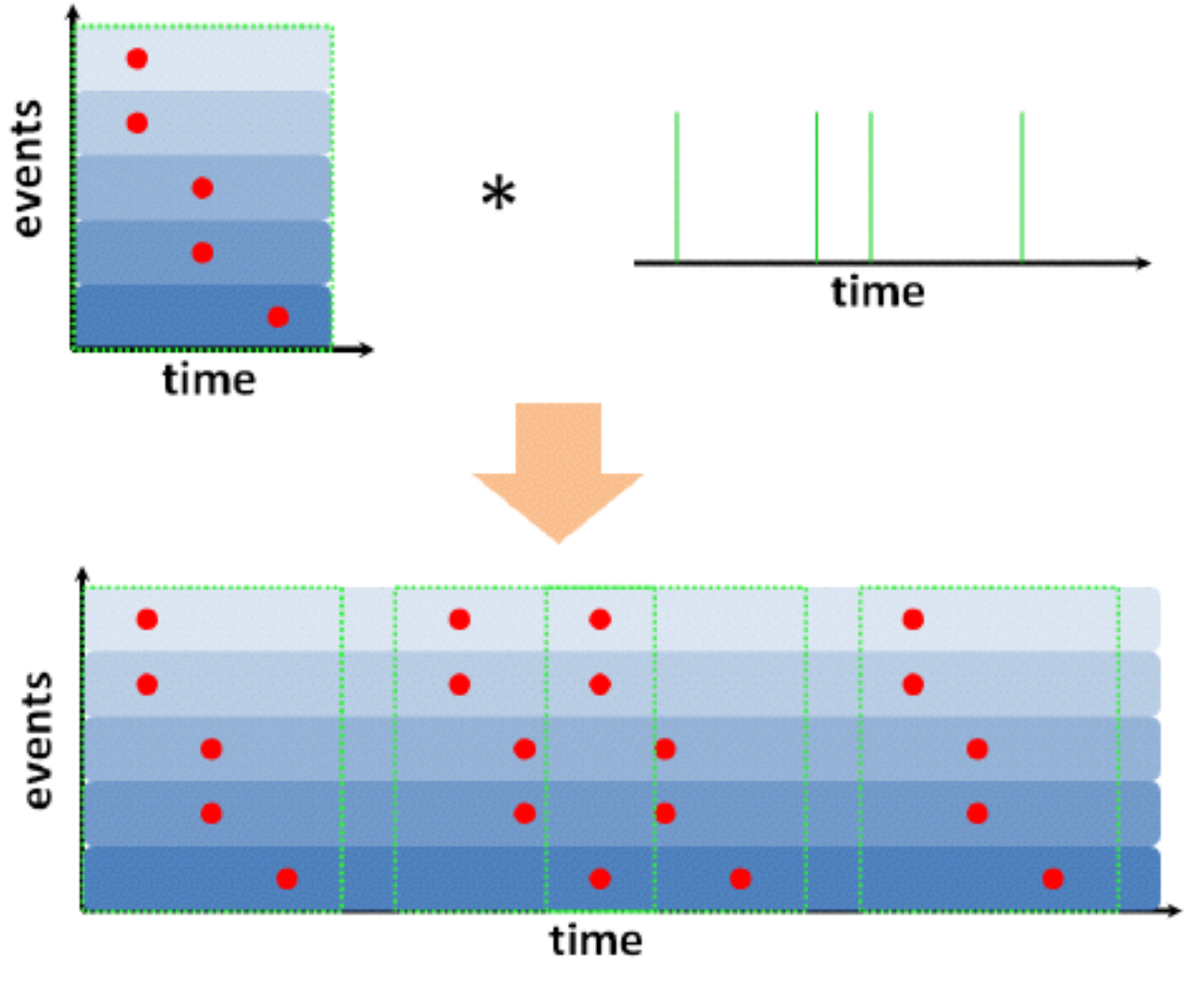
F. Wang et al. Towards Heterogeneous Temporal Clinical Event Pattern Discovery: A Convolutional Approach. KDD 2012.

One-Side Convolution

Definition (One-Side Convolution). *The one-side convolution of* $\mathbf{F} \in \mathbb{R}^{n \times m}$ *and* $\mathbf{g} \in \mathbb{R}^{t \times 1}$ *is an* $n \times t$ *matrix with*

$$(\mathbf{F} * \mathbf{g})_{ij} = \sum_{k=1}^{t} g_{j-k+1} F_{ik}$$

Note that $g_j = 0$ *if* $j \leqslant 0$ *or* $j > t$, *and* $F_{ik} = 0$ *if* $k > m$.



One-Side Convolutional NMF

$$\begin{aligned} \min_{\mathcal{F},\mathcal{G}} \quad & \mathcal{J} \\ s.t. \quad & \forall r = 1, \cdots, R; c = 1, \cdots, C \\ & \mathbf{F}^{(r)} \geqslant 0, \mathbf{g}_c^{(r)} \geqslant 0 \end{aligned}$$

$$\begin{aligned} F_{ik}^{(r)} &\leftarrow F_{ik}^{(r)} \left(\frac{\sum_{c=1}^{C} \sum_{j=1}^{t} A_{c_{ij}}^{\beta-1} X_{c_{ij}} Y_{c_{ij}}^{\beta-2} g_{c_{j-k+1}}^{(r)}}{\sum_{c=1}^{C} \sum_{j=1}^{t} A_{c_{ij}} Y_{c_{ij}}^{\beta-1} g_{c_{j-k+1}}^{(r)} + \lambda_1} \right)^{\eta(\beta)} \\ g_{c_k}^{(r)} &\leftarrow g_c^{(r)} \left(\frac{\sum_{i=1}^{n} \sum_{j=1}^{t} A_{c_{ij}}^{\beta-1} X_{c_{ij}} Y_{c_{ij}}^{\beta-2} F_{i,j-k+1}^{(r)}}{\sum_{i=1}^{n} \sum_{j=1}^{t} A_{c_{ij}} Y_{c_{ij}}^{\beta-1} F_{i,j-k+1}^{(r)} + \lambda_2} \right)^{\eta(\beta)} \end{aligned}$$

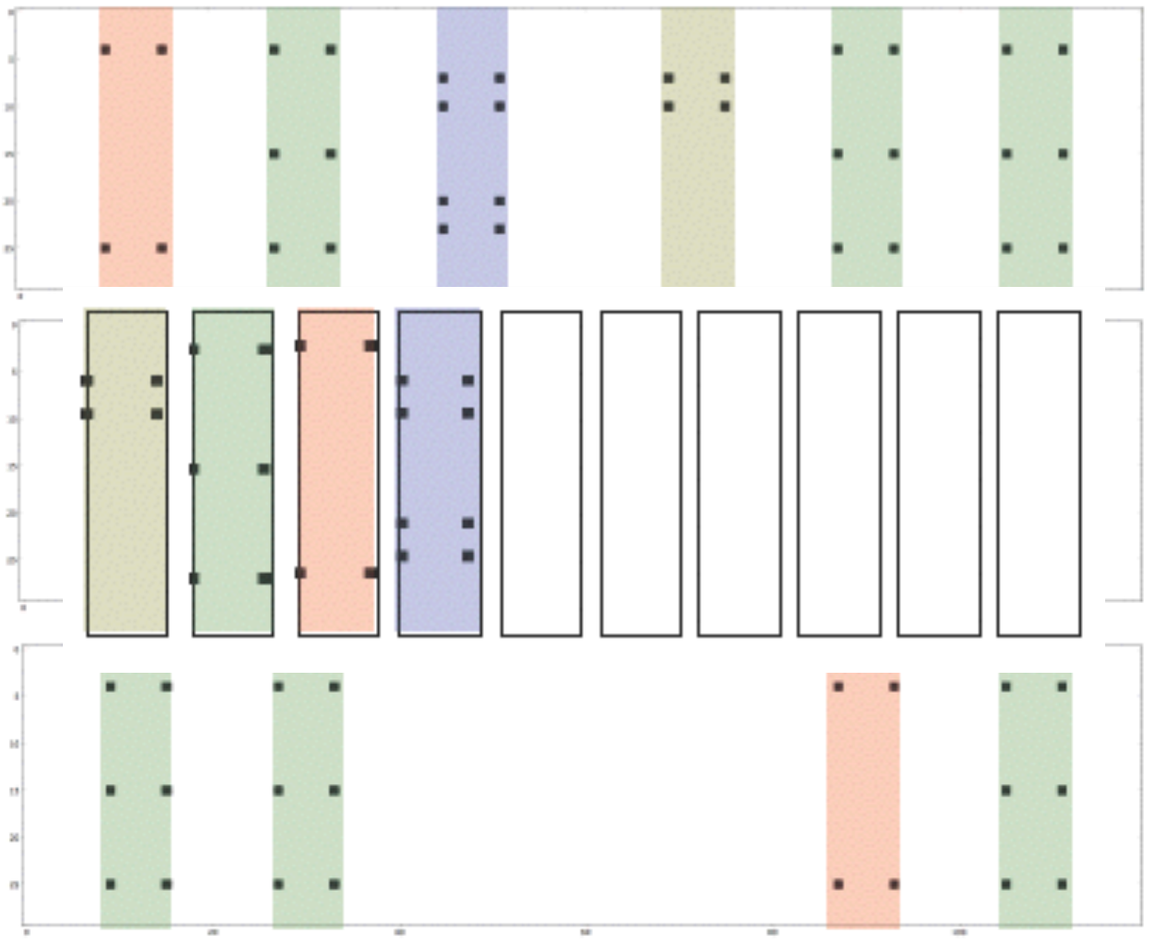
$$\mathcal{J} = \sum_{c=1}^{C} d_\beta \left(\mathbf{A}_c \odot \mathbf{X}_c, \mathbf{A}_c \odot \left(\sum_{r=1}^{R} \mathbf{F}^{(r)} * \mathbf{g}_c^{(r)} \right) \right) + \lambda_1 \sum_{r=1}^{R} \|\mathbf{F}^{(r)}\|_1 + \lambda_2 \sum_{c=1}^{C} \sum_{r=1}^{R} \|\mathbf{g}_c^{(r)}\|_1$$

$$\eta(\beta) = \begin{cases} \frac{1}{2-\beta}, & \beta < 1 \\ 1, & 1 \leqslant \beta \leqslant 2 \\ \frac{1}{\beta-1}, & \beta > 2 \end{cases}$$

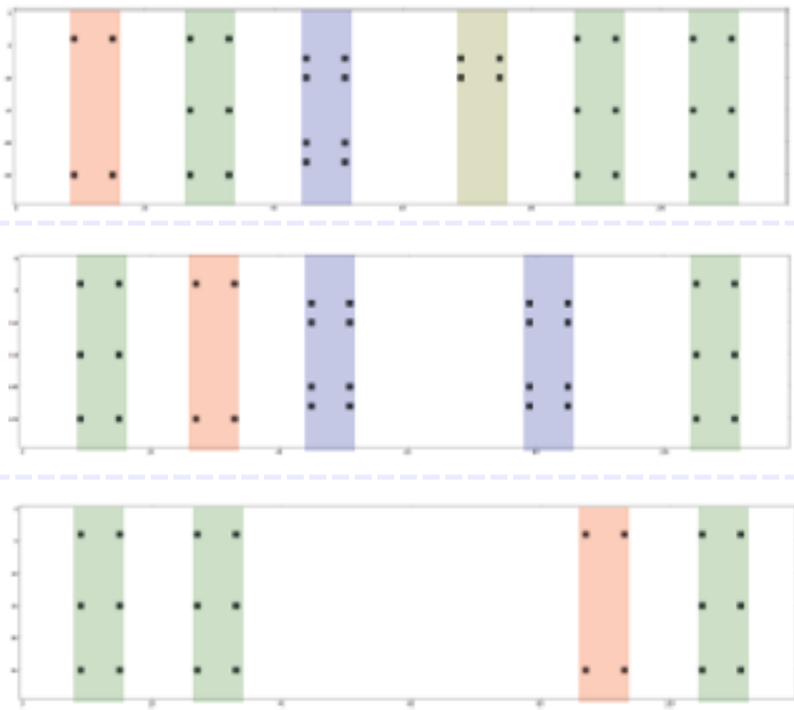
Definition (β-divergence) *The β-divergence between two matrices $\mathbf{A}$ and $\mathbf{B}$ with the same size is*

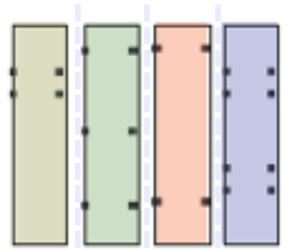
$$d_\beta(\mathbf{A}, \mathbf{B}) = \frac{1}{\beta(\beta-1)} \sum\nolimits_{ij} \left(A_{ij}^\beta + (\beta-1) B_{ij}^\beta - \beta A_{ij} B_{ij}^{\beta-1} \right)$$

Synthetic Example



Bag-of-Pattern Representation



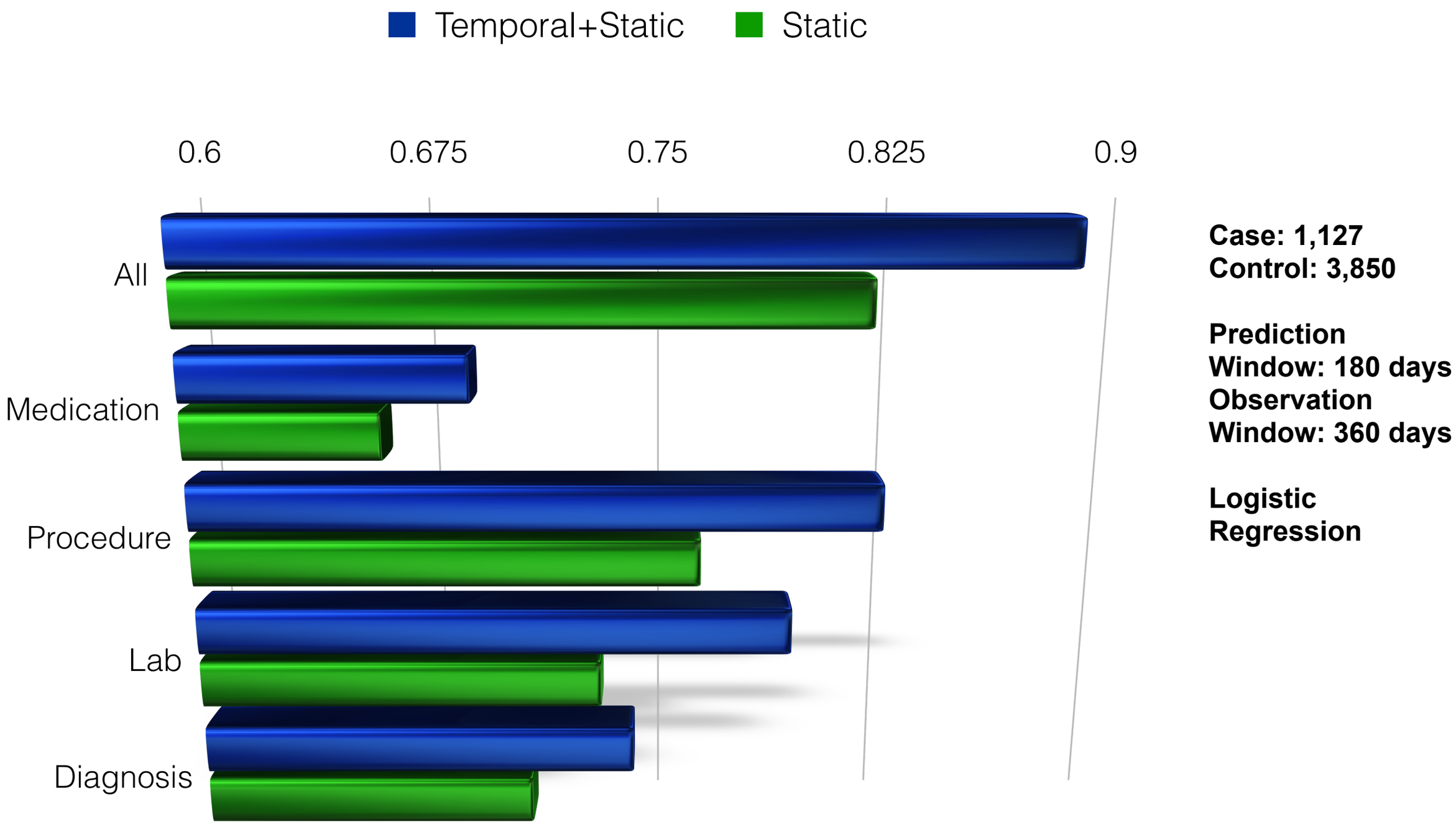


[1 3 1 1]

[0 2 1 2]

[0 3 1 0]

CHF Onset Prediction



Case: 1,127
Control: 3,850

Prediction Window: 180 days
Observation Window: 360 days

Logistic Regression

Feature Interactions in EMR

PID	TIME	MEDICATION	DIAGNOSIS
1	01/01/98	Loop Diuretics	Heart Failure
1	02/01/98	ACE Inhibitors	Hypertension
1	05/01/98	Beta Blockers	Hypertension
1	08/01/98	Cardiotonics	Heart Failure

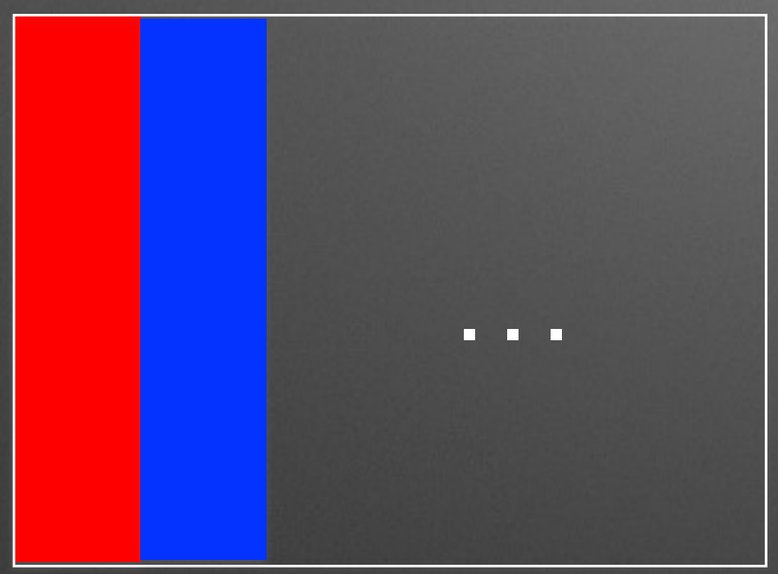
Logistic Regression Revisited

Data Matrix: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$

Decision Function: $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$

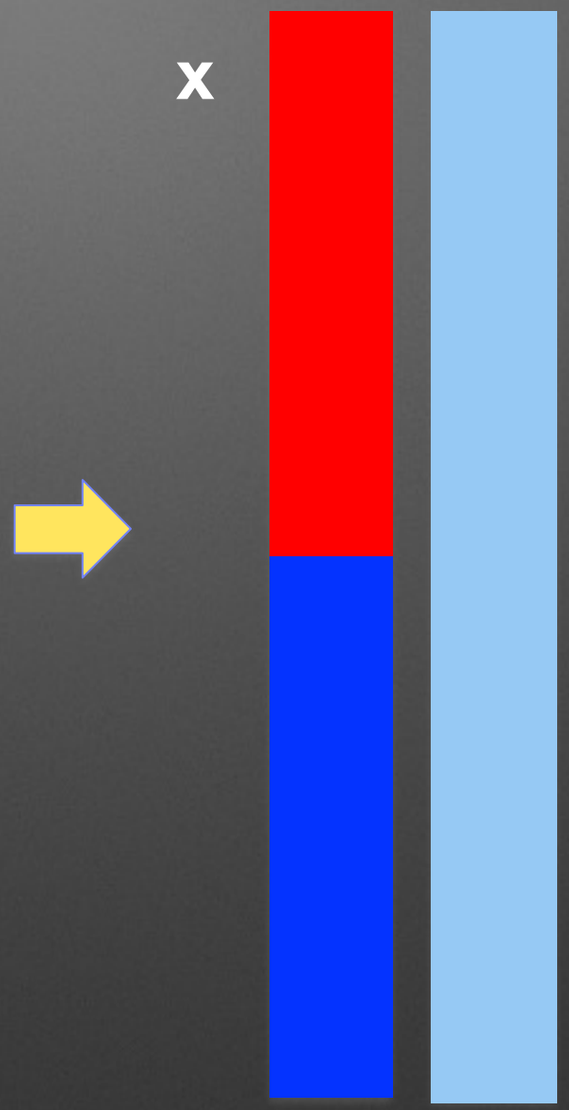
Loss Function: $\ell_{org}(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^{n} \log \left[1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b)) \right]$

Traditional Logistic Regression



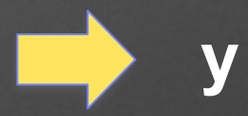
. . .

x



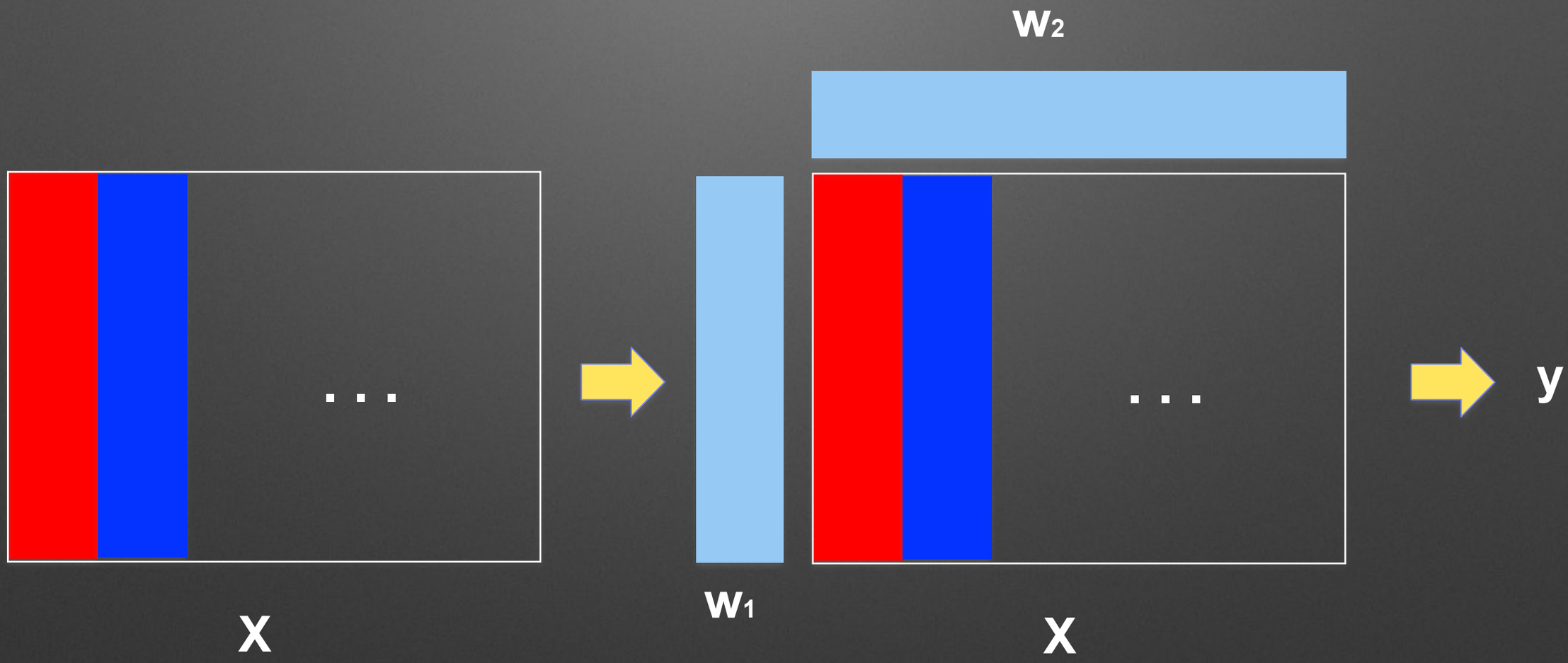
x

w



y

High-Order Logistic Regression



Multi-Linear Sparse Logistic Regression

Decision Function:

$$f_{(\mathcal{W},b)}(\mathcal{X}^i) = \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K + b$$

$$\mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K = \sum_{i_1=1}^{d_1} \sum_{i_2=1}^{d_2} \cdots \sum_{i_K=1}^{d_K} w^1_{i_1} w^2_{i_2} \cdots w^K_{i_K} X^i_{i_1 i_2 \cdots i_K}$$

Loss Function:

$$\min_{\mathcal{W}} \mathcal{J}(\mathcal{W}, b) = \ell(\mathcal{W}, b) + \mathcal{R}(\mathcal{W})$$

$$\begin{aligned} \ell(\mathcal{W}, b) &= \frac{1}{n} \sum_{i=1}^{n} \ell(\mathcal{X}_i, y_i, \mathcal{W}) \qquad (5) \\ &= \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K + b) \\ &= \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_{(\mathcal{W},b)}(\mathcal{X}^i)) \end{aligned}$$

$$\ell_l(\mathcal{W}, b) = \log[1 + \exp(-y_i f_{(\mathcal{W},b)}(\mathcal{X}^i))]$$

$$\begin{aligned} \mathcal{R}(\mathcal{W}) &= \mathcal{R}_1(\mathcal{W}) + \mathcal{R}_2(\mathcal{W}) \\ &= \sum_{k=1}^{K} \lambda_k \|\mathbf{w}^k\|_1 + \frac{1}{2} \sum_{k=1}^{K} \mu_k \|\mathbf{w}^k\|_2^2 \end{aligned}$$

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.

Block Coordinate Descent

$$(\mathcal{W}_{(0)}, b_{(0)})$$

$$\begin{aligned} \mathcal{W}_{(t)}^{1\sim(k-1)} &= \left\{\mathbf{w}_{(t)}^{1}, \mathbf{w}_{(t)}^{2}, \cdots, \mathbf{w}_{(t)}^{k-1}\right\} \\ \mathcal{W}_{(t-1)}^{(k+1)\sim K} &= \left\{\mathbf{w}_{(t-1)}^{k+1}, \mathbf{w}_{(t-1)}^{k+2}, \cdots, \mathbf{w}_{(t-1)}^{K}\right\} \end{aligned}$$

$$(\mathbf{w}_{(t)}^{k}, b_{(t)}) = \arg\min_{(\mathbf{w},b)} \left[\ell(\mathcal{W}_{(t)}^{1\sim(k-1)}, \mathbf{w}, \mathcal{W}_{(t-1)}^{(k+1)\sim K}, b) + \lambda_k \|\mathbf{w}\|_1 + \frac{\mu_k}{2}\|\mathbf{w}\|_2^2\right]$$

Algorithm 1 Block Coordinate Descent Procedure

Require: Data set $\{\mathcal{X}_i, y_i\}_{i=1}^{n}$, Regularization parameters $\{\lambda_k, \mu_k\}_{k=1}^{K}$

1: **Initialization**: $(\mathcal{W}_{(0)}, b_{(0)})$, $t = 0$
2: **while** Not Converge **do**
3: **for** $k = 1 : K$ **do**
4: Update $(\mathbf{w}_{(t)}^{k}, b_{(t,k)})$ by solving problem (11)
5: $t = t + 1$
6: **end for**
7: **end while**

Block Proximal Gradient

Algorithm 2 Block Proximal Gradient Descent for Multilinear Sparse Logistic Regression

Require: Data set $\{\mathcal{X}_i, y_i\}_{i=1}^n$, Regularization parameters $\{\lambda_k, \mu_k\}_{k=1}^K$, $r_0 = 1$, $\delta_\omega < 1$

Initialization: $(\mathcal{W}_{(0)}, b_{(0)})$, $t = 1$

while Not Converge **do**

  for $k = 1 : K$ **do**

    Compute $\tau_{(t)}^k$ with $\tau_{(t)}^k = \frac{\sqrt{2}}{n} \sum_{i=1}^n \left(\left\| \nabla_{\mathbf{w}^k}^{(t,k)} f_{(\mathcal{W},b)}(\mathcal{X}^i) \right\|_2 + 1 \right)^2$

    Compute $\omega_{(t)}^k$ with $\omega_{(t)}^k = \min \left(\omega_{(t)}, \delta_\omega \sqrt{\frac{\tau_{(t-1)}^k}{\tau_{(t)}^k}} \right)$

    Compute $\widetilde{\mathbf{w}}_{(t)}^k$ with $\widetilde{\mathbf{w}}_{(t)}^k = \mathbf{w}_{(t-1)}^k + \omega_{(t)}^k (\mathbf{w}_{(t-1)}^k - \mathbf{w}_{(t-2)}^k)$

    Update $\mathbf{w}_{(t)}^k$ by $\mathbf{w}_{(t)}^k = \mathcal{S}_{\alpha_{(t)}^k} \left(\frac{\tau_{(t)}^k \widetilde{\mathbf{w}}_{(t)}^k - \nabla_{\mathbf{w}^k} \ell_{(t)}^k (\widetilde{\mathbf{w}}_{(t)}^k, b_{(t,k-1)})}{\tau_{(t)}^k + \mu_k} \right)$

    Compute $\widetilde{b}_{(t,k)}$ with $\widetilde{b}_{(t,k)} = b_{(t,k-1)} + \omega_{(t)}^k (b_{(t,k-1)} - b_{(t,k-2)})$

    Update $b_{(t,k)}$ by $b_{(t,k)} = \widetilde{b}_{t,k} - \frac{1}{\tau_{(t)}^k} \nabla_b \ell_{(t)}^k (\mathbf{w}_{(t)}^k, \widetilde{b}_{(t,k)})$

  end for

  if $\ell(\mathcal{W}_{(t-1)}, b_{(t-1,K)}) \leqslant \ell(\mathcal{W}_{(t)}, b_{(t,K)})$ **then**

    Reupdate $\mathbf{w}_{(t)}^k$ and $b_{(t,k)}$ using $\mathbf{w}_{(t)}^k = \mathcal{S}_{\alpha_{(t)}^k} \left(\frac{\tau_{(t)}^k \widetilde{\mathbf{w}}_{(t)}^k - \nabla_{\mathbf{w}^k} \ell_{(t)}^k (\widetilde{\mathbf{w}}_{(t)}^k, b_{(t,k-1)})}{\tau_{(t)}^k + \mu_k} \right)$ and $b_{(t,k)} = \widetilde{b}_{t,k} - \frac{1}{\tau_{(t)}^k} \nabla_b \ell_{(t)}^k (\mathbf{w}_{(t)}^k, \widetilde{b}_{(t,k)})$, with $\widetilde{\mathbf{w}}_{(t)}^k = \mathbf{w}_{(t-1)}^k$ and $\widetilde{b}_{(t,k)} = b_{(t,k-1)}$

  end if

  $t = t + 1$

end while

EMR Data

Medication Claim (Drug+Diagnosis)

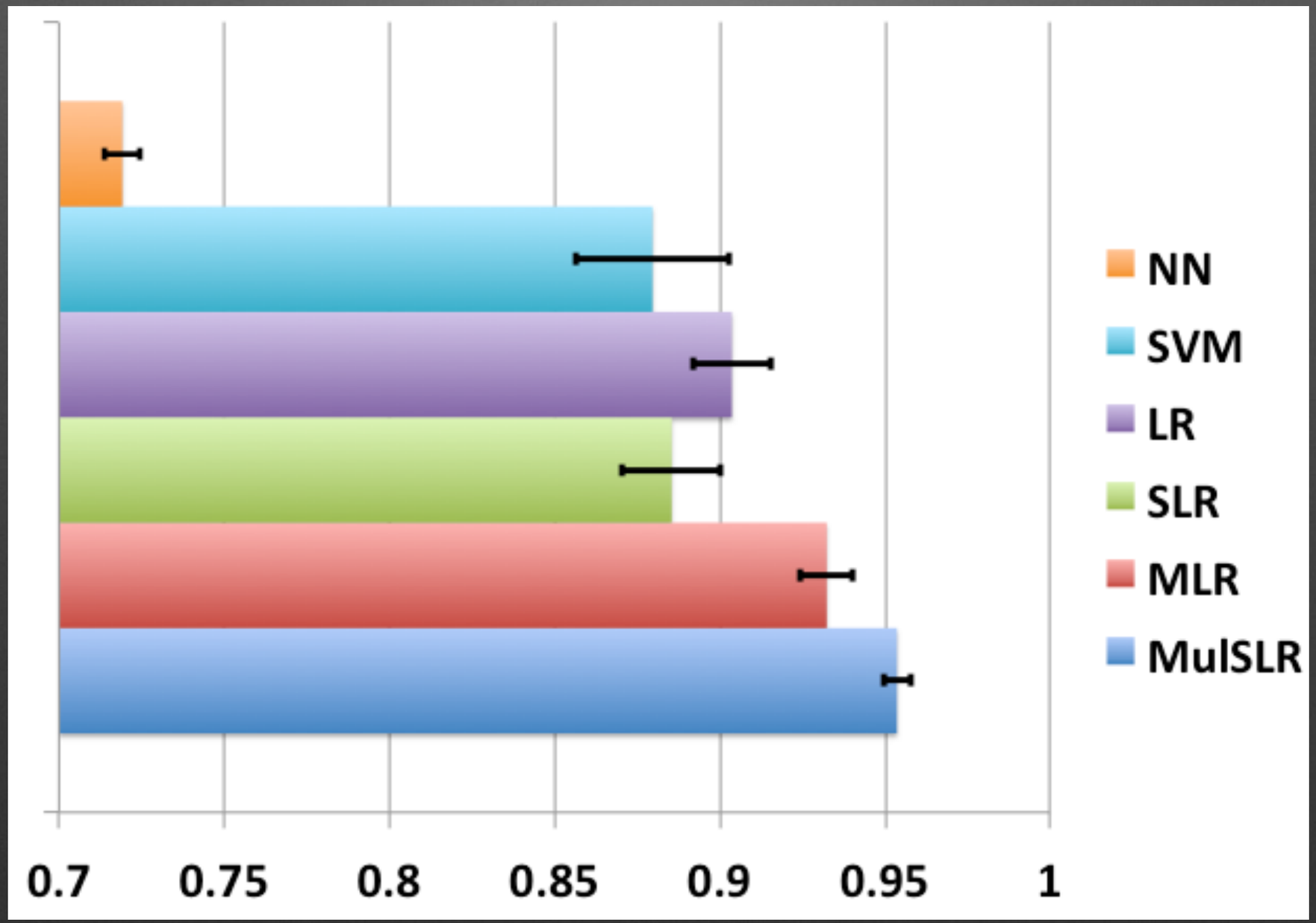
USP Class
92

Hierarchical Condition Code (HCC)
195

1000 CHF Case Patients

2000 Group Matched Control Patients

EMR Data



EHR Data

Diagnosis	
Heart Disease	Congestive Heart Failure
	Acute Myocardial Infarction
	Specified Heart Arrhythmias
	Ischemic or Unspecified Stroke
Hypertension	Hypertension
	Hypertensive Heart Disease
Lung Disease	Fibrosis of Lung and Other Chronic Lung Disorders
	Asthma
	Chronic Obstructive Pulmonary Disease (COPD)
Kidney Disease	Chronic Kidney Disease, Very Severe (Stage 5)
	Chronic Kidney Disease, Mild or Unspecified (Stage 1-2 or Unspecified)

Medication
Antihyperlipidemic
Antihypertensive
Beta Blockers
Calcium Blockers
Cardiotonics
Cardiovascular
Corticosteriods
Diuretics
General Anesthetics
Gout
Vaccines

Wang, Fei, et al. "Clinical Risk Prediction with Multilinear Sparse Logistic Regression." KDD 2014

Roadmap

- Background
- Electronic Health Records
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

What is clinical pathway?

(1980's Karen Zander & Kathleen Bower, New England Medical Center, Boston)

Clinical pathways are designed to provide optimal patient care for groups of patients

- A tool to incorporate local and national guidelines into everyday practice
- Better Service, Less Cost

What is Clinical Pathway

Eligibility & exclusion criteria

pathway title

patient information

actions & advices

Bronchial lung cancer clinical pathway

Suitable for: First-listed diagnosis is bronchial lung cancer (ICD-10: C34; D02.2)

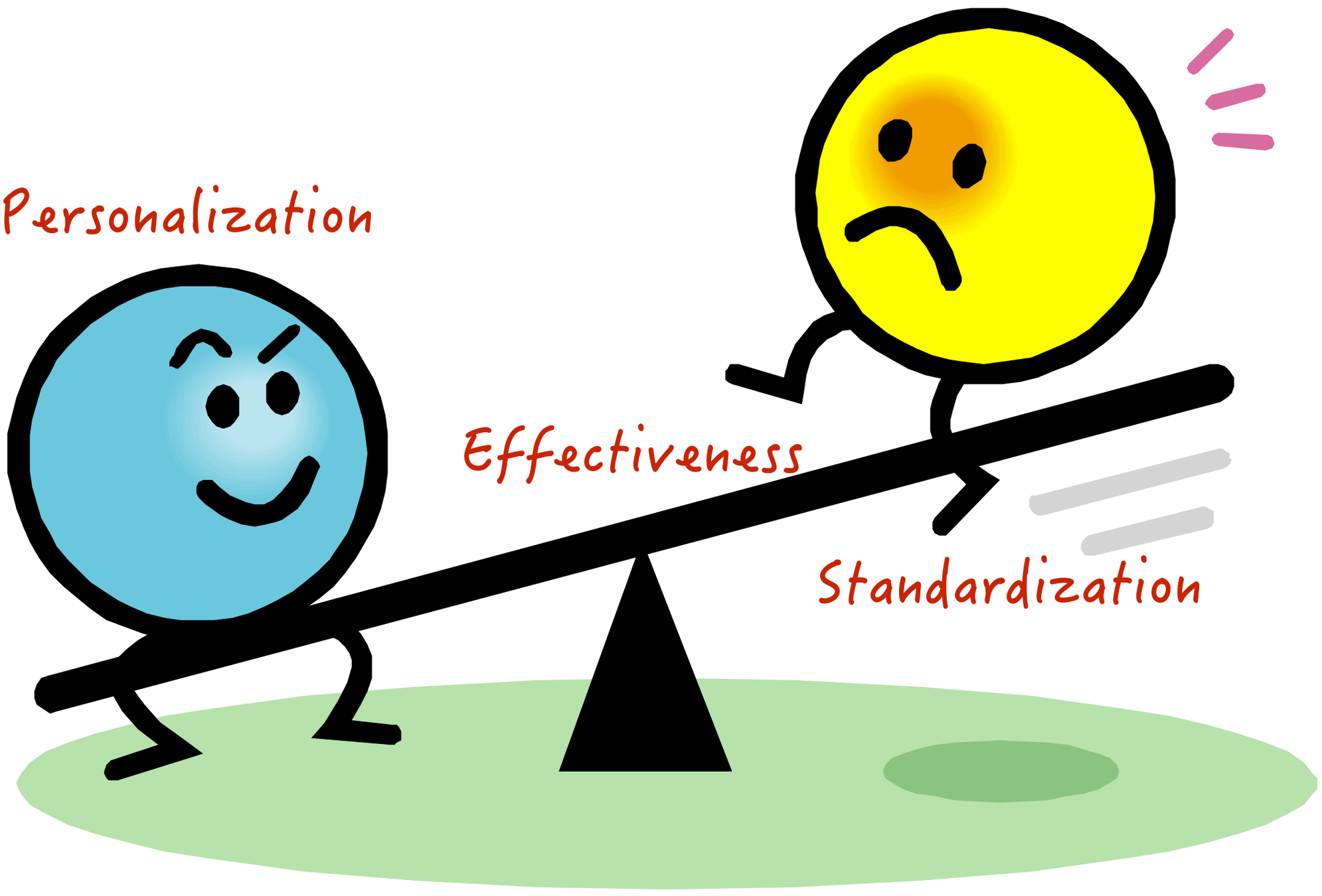
Local excision of pulmonary / Lobectomy / Pneumonectomy + Systematic lymph

Patient Name: _____ Gender: _____ Age: _____

Outpatient service NO: _____ Hospitalization NO: _____

Time	Day 1	Day 2-6 (preoperative day)	Day 4-7 (operative day)
Major diagnosis and treatment	☐ Medical history inquiry and physical examination ☐ Write patient record ☐ Issue laboratory orders and check request form ☐ Attending rounds ☐ Initially set the treatment plan	☐ Higher authority physician rounds ☐ Preoperative preparation ☐ Clinical stage and preoperative evaluation ☐ Preoperative discussion and surgical planning ☐ Complete the consultation within relevant sections according to patient condition ☐ Resident completes medical records including the course of the disease and preoperative log summary, the superior physician records, ☐ Sign the informed consent procedure, goods agreement at their own expense, blood transfusion consent, consent authorization	☐ Indwelling catheter before surgery ☐ Surgery ☐ Surgeon completes the operation record ☐ Resident completes postoperative course ☐ Higher authority physician rounds ☐ Observation of vital signs ☐ Account of illness and postoperative precautions to patients and their families
Doctor's major advice	**Long-term medical advice:** ☐ Thoracic surgery Secondary care ☐ Normal diet **Temporary medical advice:** ☐ Blood, urine, stool routine examination ☐ Coagulation, blood type, liver and kidney function, electrolytes examination, infectious disease screening, tumor markers check ☐ Lung function, arterial blood gas analysis, ECG, echocardiography ☐ Sputum cytology, bronchoscopy + biopsy ☐ Imaging: lateral chest X-ray, chest CT, abdominal ultrasound or CT, whole body bone scan, brain MRI or CT ☐ When necessary: PET-CT or SPECT, mediastinoscopy, 24-hour ambulatory ECG, percutaneous lung biopsy, etc.	**Long-term medical advice:** ☐ Atomizing inhalation **Temporary medical advice:** ☐ Schedule for tomorrow under general anesthesia: ◎Local excision of pulmonary ◎Lobectomy ◎Pneumonectomy ◎Thoracotomy surgery ☐ No food or water intake 6 hours before surgery ☐ Enema the night before surgery ☐ Preoperative skin preparation ☐ Preparation of blood transfusion ☐ Sedative drugs (where appropriate) ☐ Preparation of antibacterial drugs in surgery ☐ Other special advices	**Long-term medical advice:** ☐ General thoracic surgery postoperative care ☐ Premium or first level nursing ☐ Liquid food intake 6 hours after clear-headed ☐ Oxygen inhalation ☐ Body temperature, ECG, blood pressure, respiration, pulse, blood oxygen saturation monitoring ☐ Record the amount of chest drainage ☐ Continued catheterization, record 24-hour intake and output ☐ Atomizing inhalation ☐ Prophylactic antibiotics ☐ Analgesic **Temporary medical advice:** ☐ Other special advices
Nursing care	☐ Introduce the ward environment, facilities and equipment ☐ Admission nursing assessment ☐ Aid smoking cessation	☐ Education, preoperative skin preparation ☐ Inform of no water and food intake ☐ Respiratory exercises	☐ Observe changes in condition ☐ Postoperative care of psychological and life ☐ Maintain patency of airway

Intelligent Clinical Pathway



Create | Inject | Identify Variation | Synthesize | Transform | Deploy

Configuration

- Global Diabetes II Care Pathway Project
- Care Pathway Project for Diabetes II with Nephropathy
  - Diabetes II model with nephropathy
  - Guideline analy

Identify variation in pathway model

Select risk type:

Nephropathy

- Cardiovascular
- Eye damage
- Nerve damage

Select data source: ☐ Select all

■PKUPH diabetes II	http://data.crl.ibm.com/pdat

Run | Cancel

Palette

- Basic BPMN Elements
  - Text Annotation
  - Group
  - Connector
  - Data Object
  - Data Store
  - Process Data Input
  - Process Data Output
  - Lane
  - Task
  - Call Activity
  - SubProcess
- Events
  - Start Event
  - Intermediate Catch Event
  - Intermediate Throw Event
  - Boundary Event
  - End Event
- Gateway Elements
  - Exclusive Gateway
  - Inclusive Gateway
  - Parallel Gateway

Graphical view | Source view

Create | Inject | Identify Variation | Synthesize | Transform | Deploy | Configuration

- Global Diabetes II Care Pathway Project
- Care Pathway Project for Diabetes II with Nephropathy
  - Diabetes II model with nephropathy
  - Guideline analy

Identify variation in pathway model

Select risk group:

- (•) Group 1 (high risk) — 231 patients
- ( ) Group 2 (medium risk) — 64 patients

Select variation types:

- ☐ Structural Variation
  - ☐ Activity
    - ■ Insufficiency (Activity in model is missing)
    - ■ Incorrectness (Activity in model is incorrect)
    - ☐ Unnecessity (Activity in model is superfluous)
  - ☐ Transition
- ☐ Semantic Variation
  - ☐ Criteria
  - ☐ Temporal

Filter: Variation Degree [> ▼] [30 ▼]

OK | Cancel

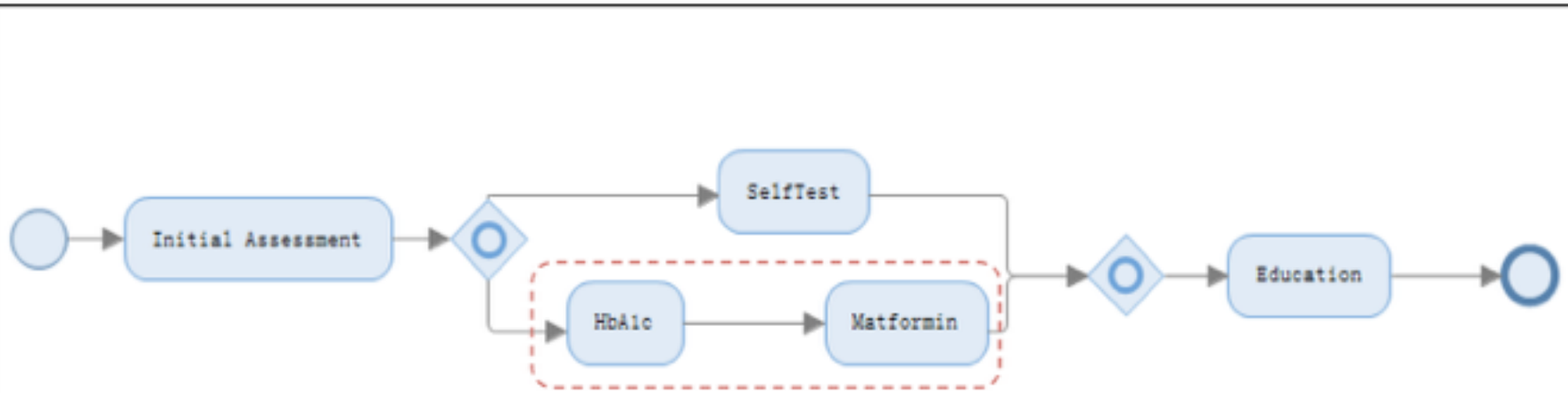
Palette
- Basic BPMN Elements
  - Text Annotation
  - Group
  - Connector
  - Data Object
  - Data Store
  - Process Data Input
  - Process Data Output
  - Lane
  - Task
  - Call Activity
  - SubProcess
- Events
  - Start Event
  - Intermediate Catch Event
  - Intermediate Throw Event
  - Boundary Event
  - End Event
- Gateway Elements
  - Exclusive Gateway
  - Inclusive Gateway
  - Parallel Gateway

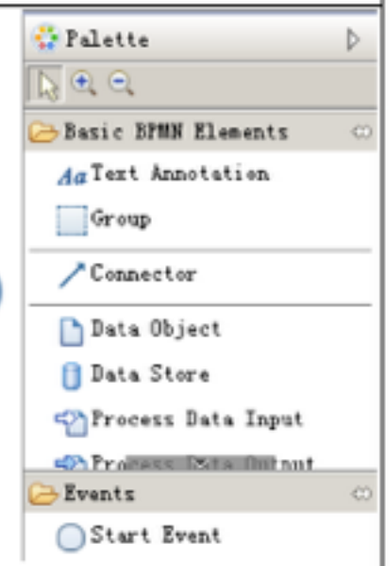
Graphical view | Source view

Create | Inject | Identify Variation | Synthesize | Transform | Deploy

Configuration

- Global Diabetes II Care Pathway Project
- Care Pathway Project for Diabetes II with Nephropathy
  - Diabetes II model with nephropathy
    - Guideline analy
  - G1 Model
    - Variation





Graphical view | Source view

Variation Results

Risk group: Nephropathy Group 1 Filter: Variation Degree >30

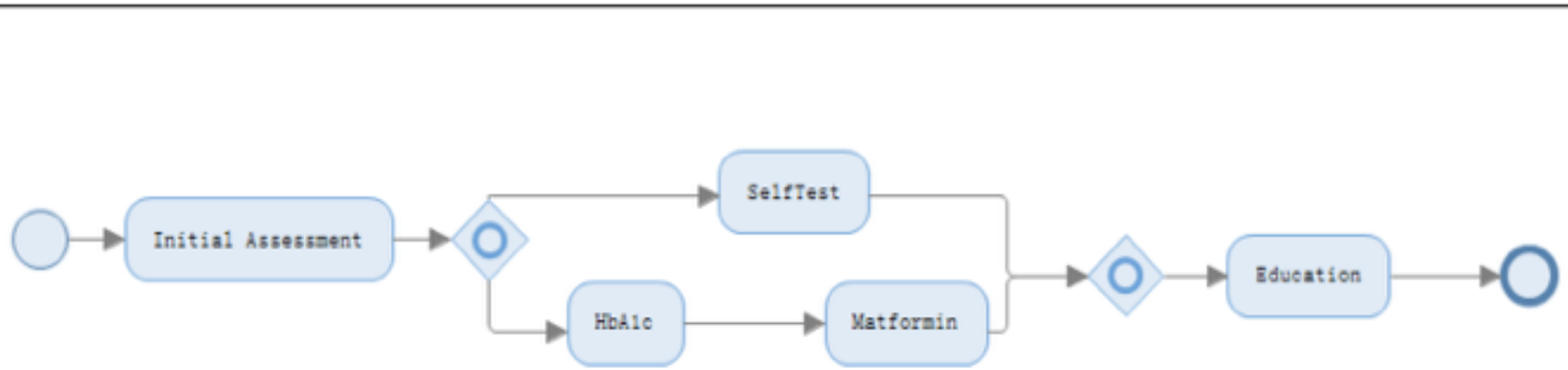
Run All

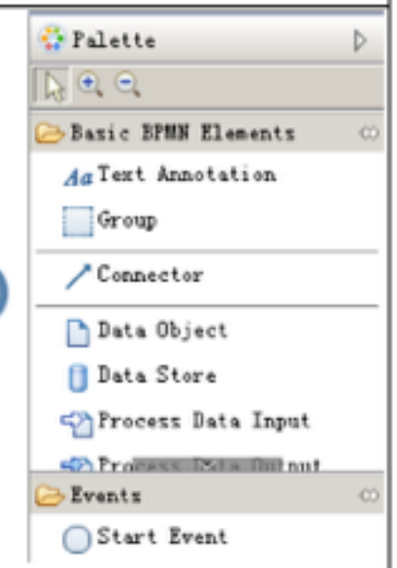
Fragment	Variation Type	Variation Degree	Recommended Analysis Method	
HbA1c → Metformin	·Structural Variation ·Activity ·Insufficiency	63	Frequent Pattern Min Linear Regression ...	Run

Whole model Variation Degree: 41

Create | Inject | Identify Variation | Synthesize | Transform | Deploy | Configuration

- Global Diabetes II Care Pathway Project
- Care Pathway Project for Diabetes II with Nephropathy
  - Diabetes II model with nephropathy
    - Guideline analy
    - G1 Model
      - Variation
      - Explorer

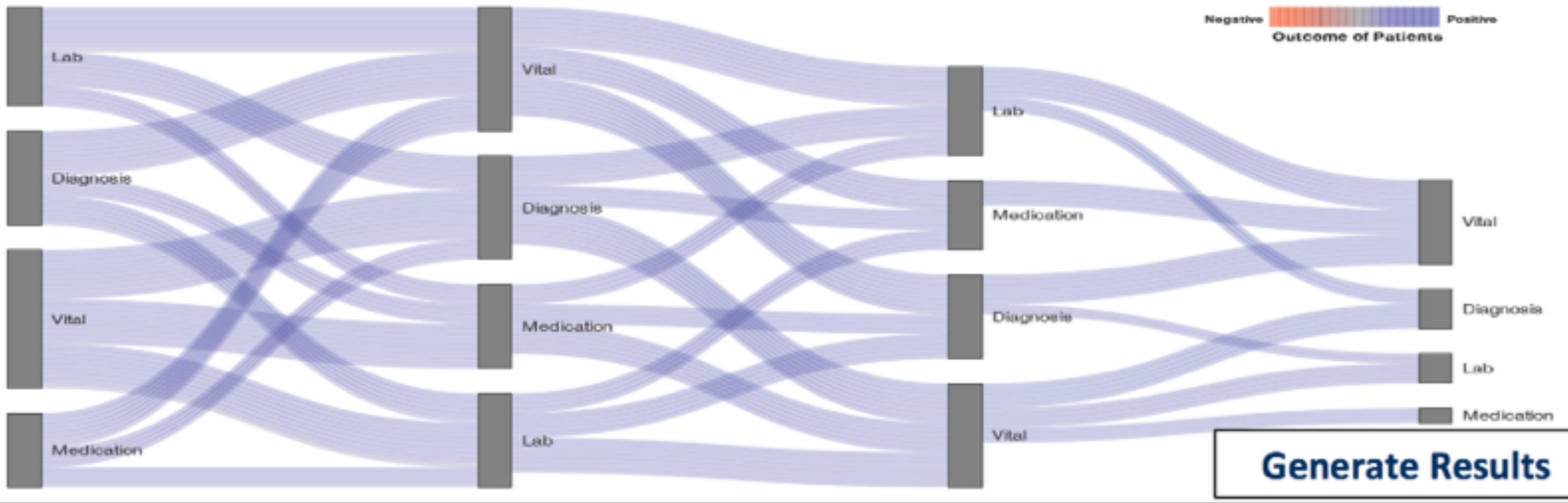




Graphical view | Source view

Interactive Plan Explorer

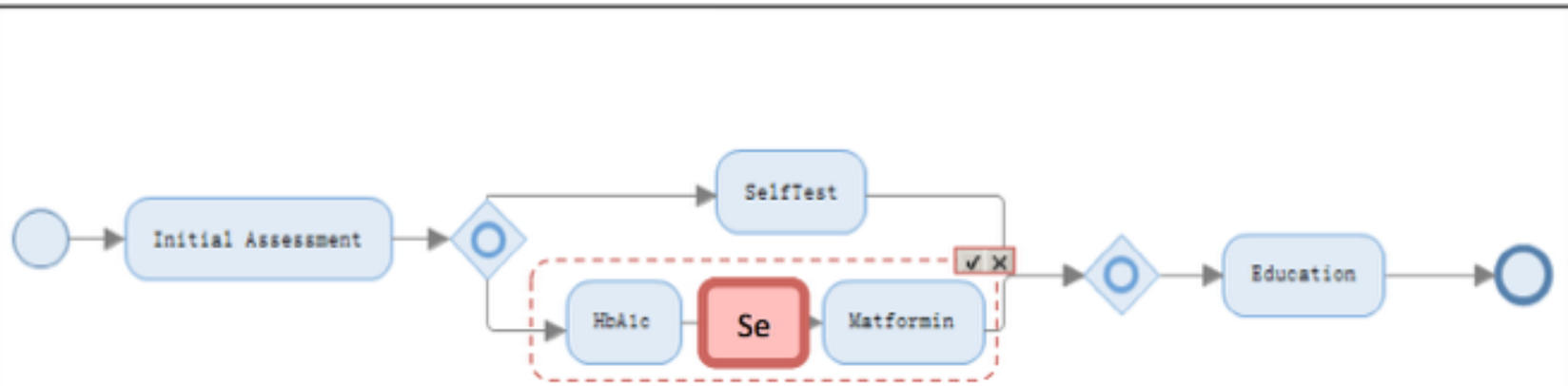
Frequent Pattern Mining: HbA1c → Metformin



Generate Results

Create | Inject | Identify Variation | Synthesize | Transform | Deploy | Configuration

- Global Diabetes II Care Pathway Project
- Care Pathway Project for Diabetes II with Nephropathy
  - Diabetes II model with nephropathy
    - Guideline analy
  - G1 Model
    - Variation
    - Explorer



Graphical view | Source view

Analysis Results
Risk group: Nephropathy Group 1 Filter: Variation Degree >30

Run All

Fragment	Variation Type	Recommended Analysis Method	Analysis Results
HbA1c → Metformin	·Structural Variation ·Activity ·Insufficiency	Frequent Pattern M ▼ Run	HbA1c → Serum creatinine → Metformin HbA1c → Bloc → Metformin

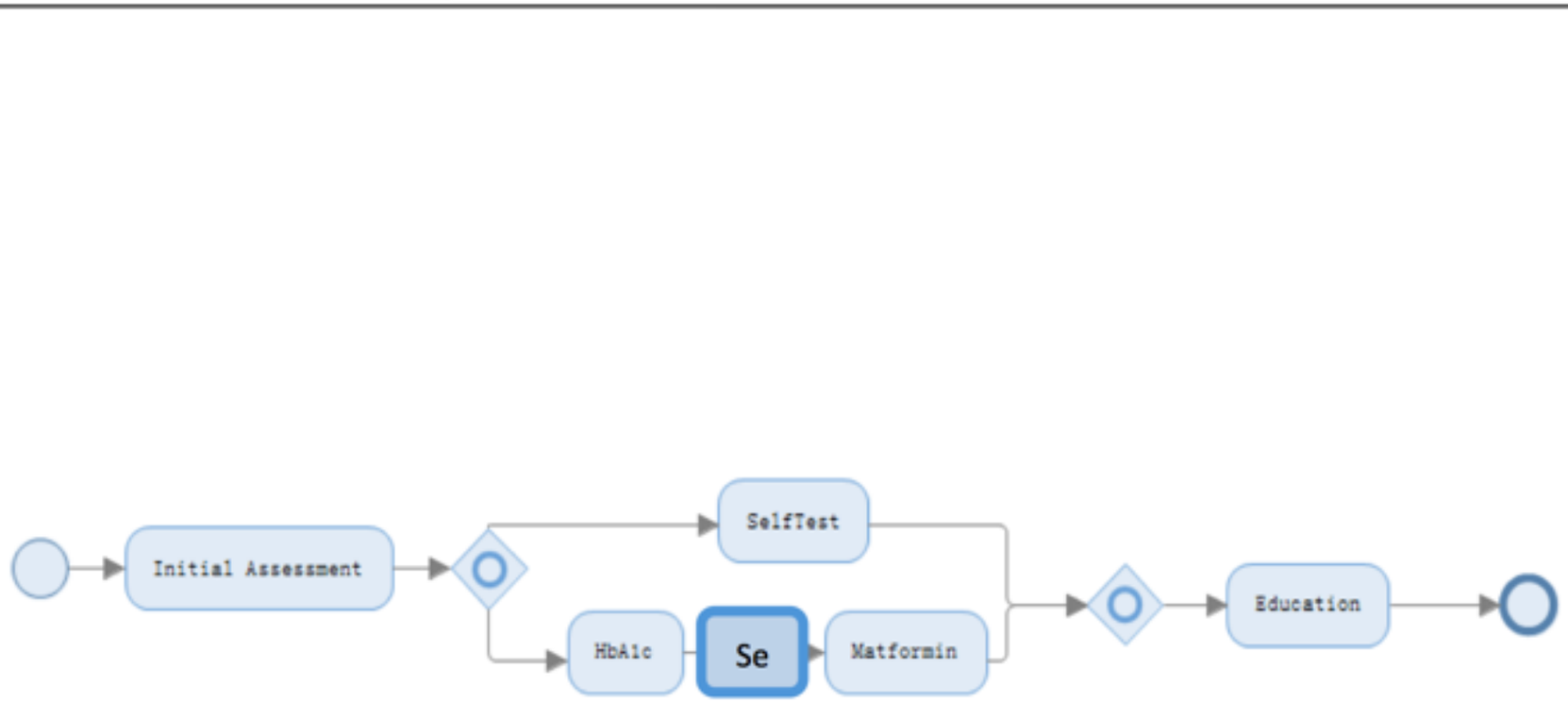
Apply selected
Apply all

Undo

Whole model Variation Degree: 41

Create | Inject | Identify Variation | Synthesize | Transform | Deploy | Configuration

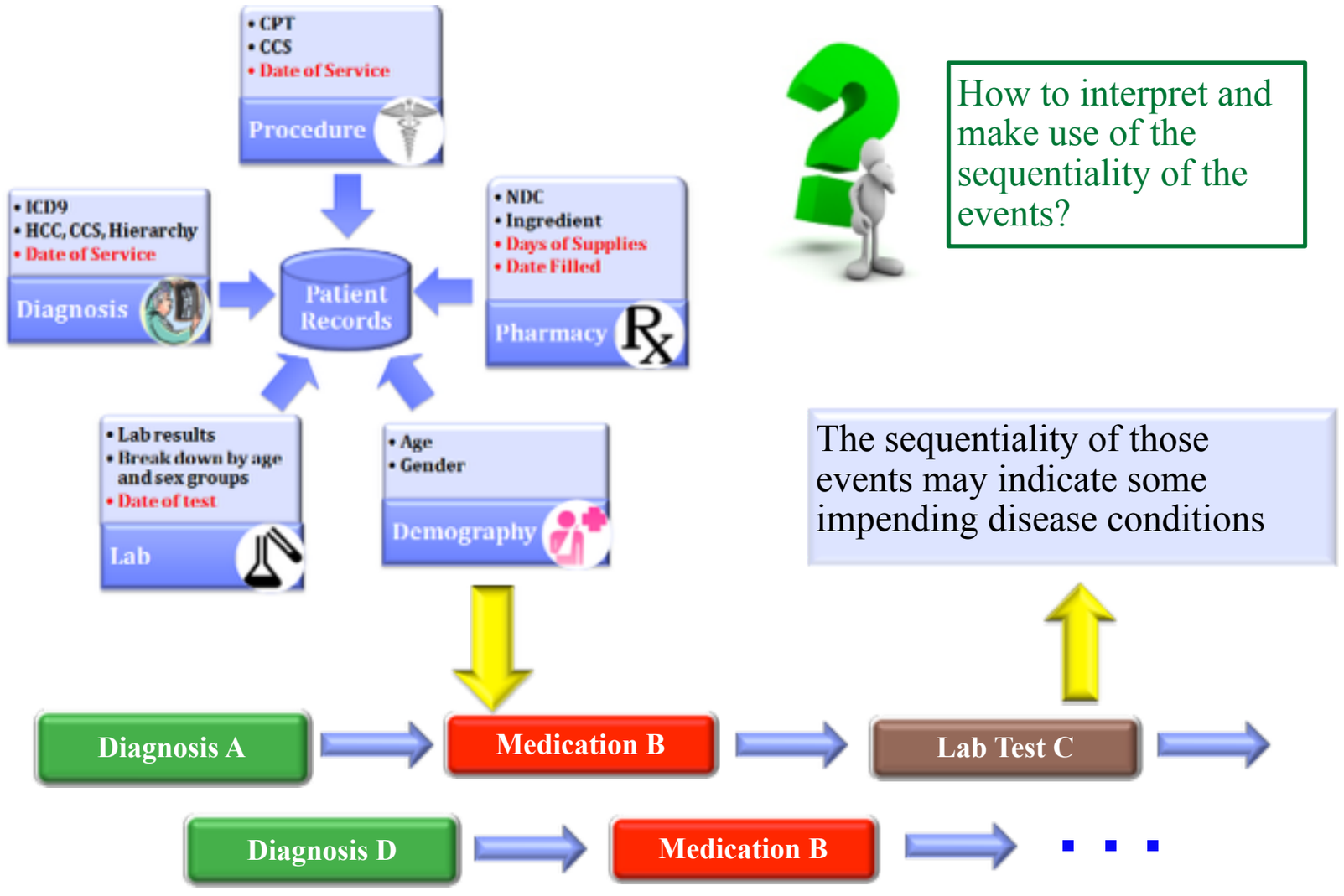
- Global Diabetes II Care Pathway Project
- Care Pathway Project for Diabetes II with Nephropathy
  - Diabetes II model with nephropathy
  - Guideline analy
  - G1 Model
    - Variation
    - Explorer



Palette
Basic BPMN Elements
Text Annotation
Group
Connector
Data Object
Data Store
Process Data Input
Process Data Output
Lane
Task
Call Activity
SubProcess
Events
Start Event
Intermediate Catch Event
Intermediate Throw Event
Boundary Event
End Event
Gateway Elements
Exclusive Gateway
Inclusive Gateway
Parallel Gateway

Graphical view | Source view

Sequence Representation



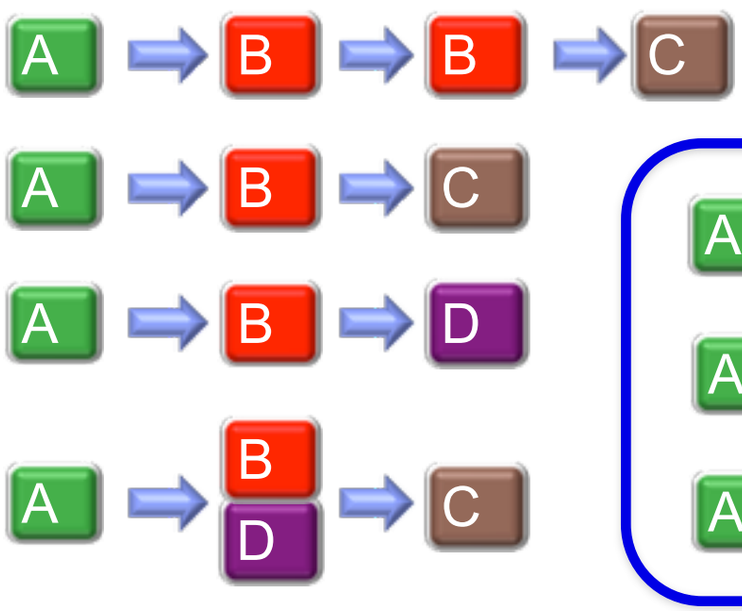
How to interpret and make use of the sequentiality of the events?

The sequentiality of those events may indicate some impending disease conditions

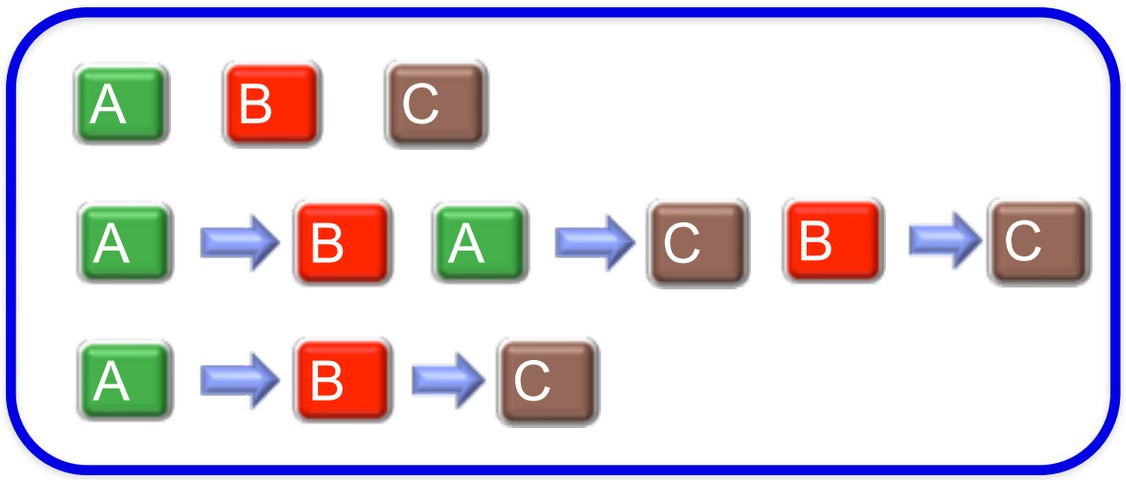
Sequential Pattern Mining

Given a set of sequences and *support* threshold, find the complete set of *frequent* subsequences

Event sequences

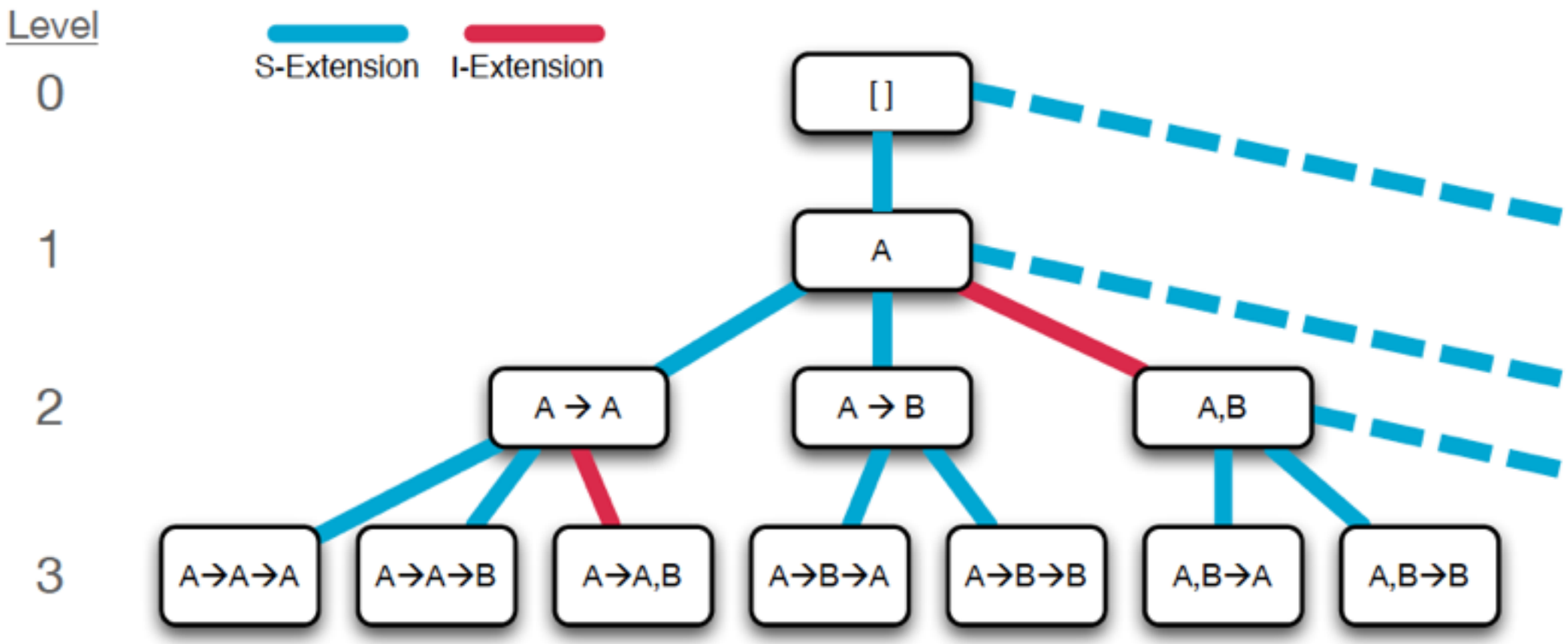


Frequent patterns with support 0.6



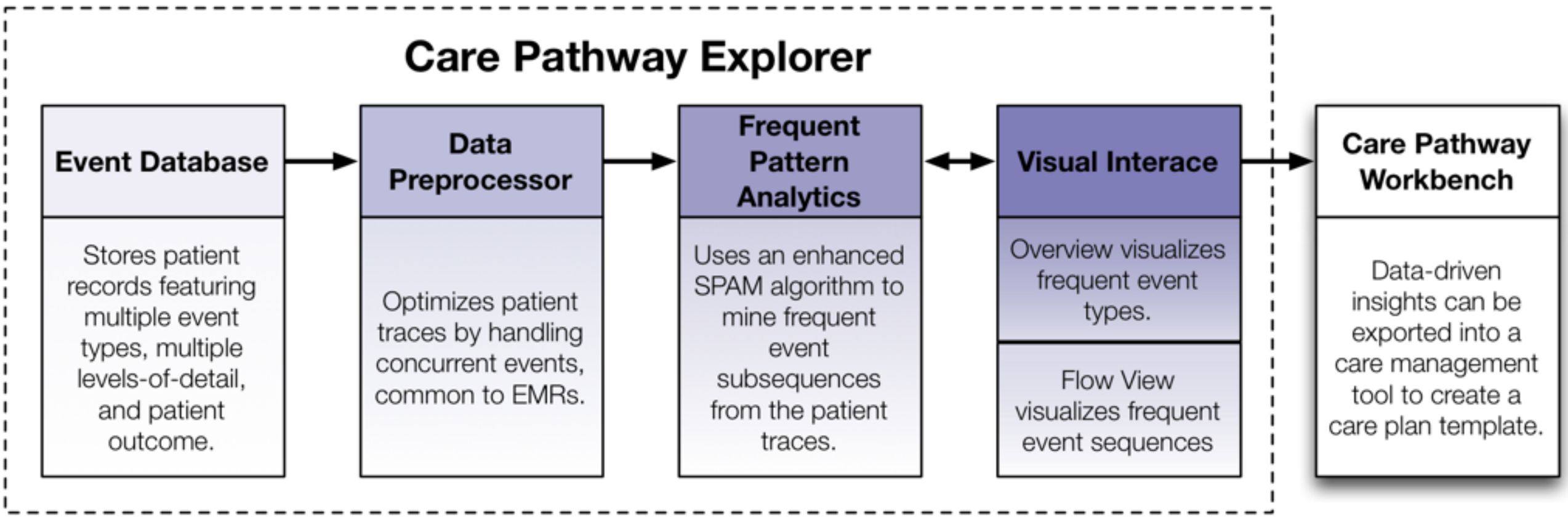
J. Han et al. Frequent pattern mining: current status and future directions. DMKD. 2007.

Pattern Growing

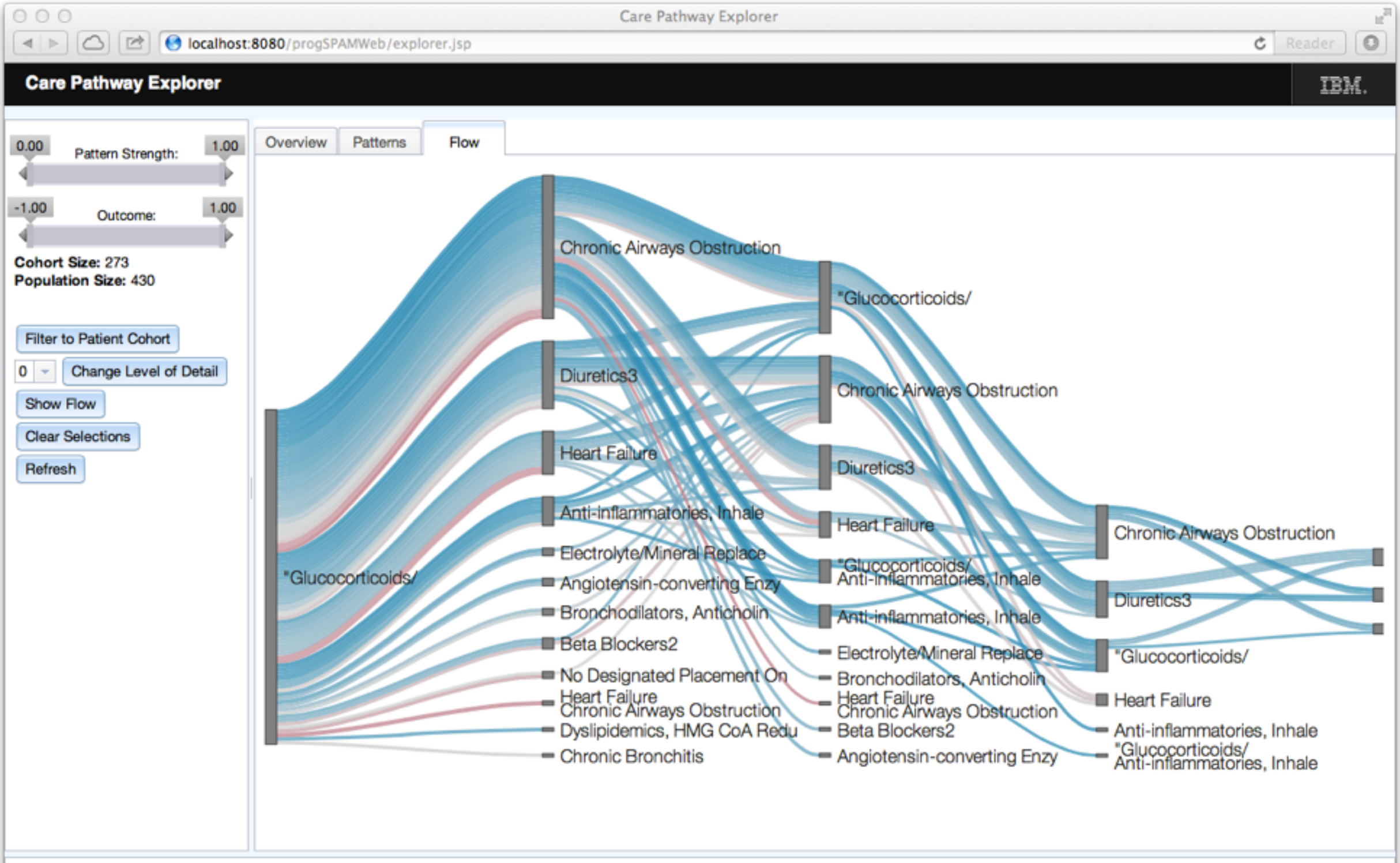


*Perer, Adam, and **Fei Wang**. "Frequence: interactive mining and visualization of temporal frequent event sequences." In Proceedings of the 19th international conference on Intelligent User Interfaces, pp. 153-162. ACM, 2014.*

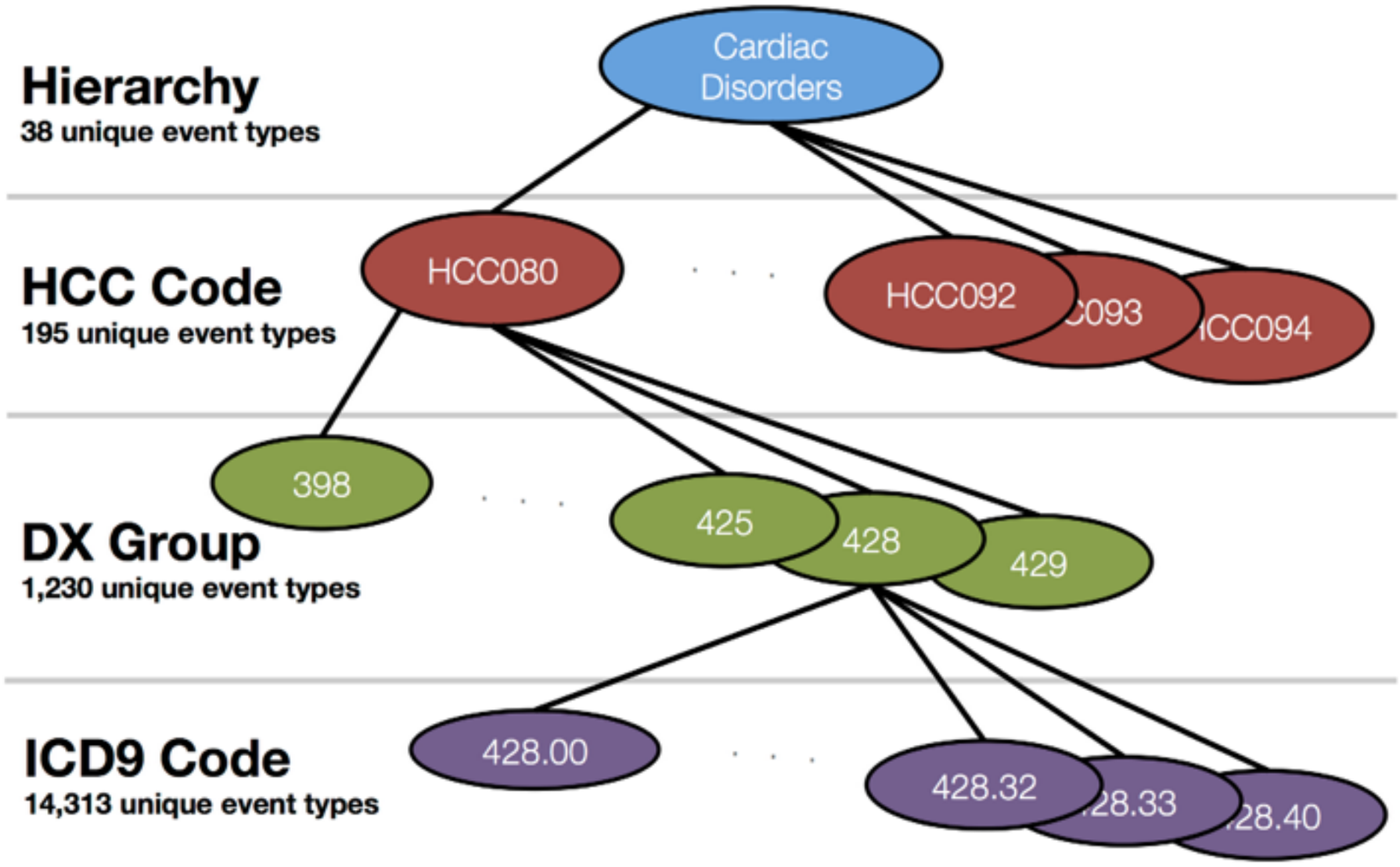
Care Pathway Explorer



User Interface



Hierarchical Exploration



Hierarchical Exploration

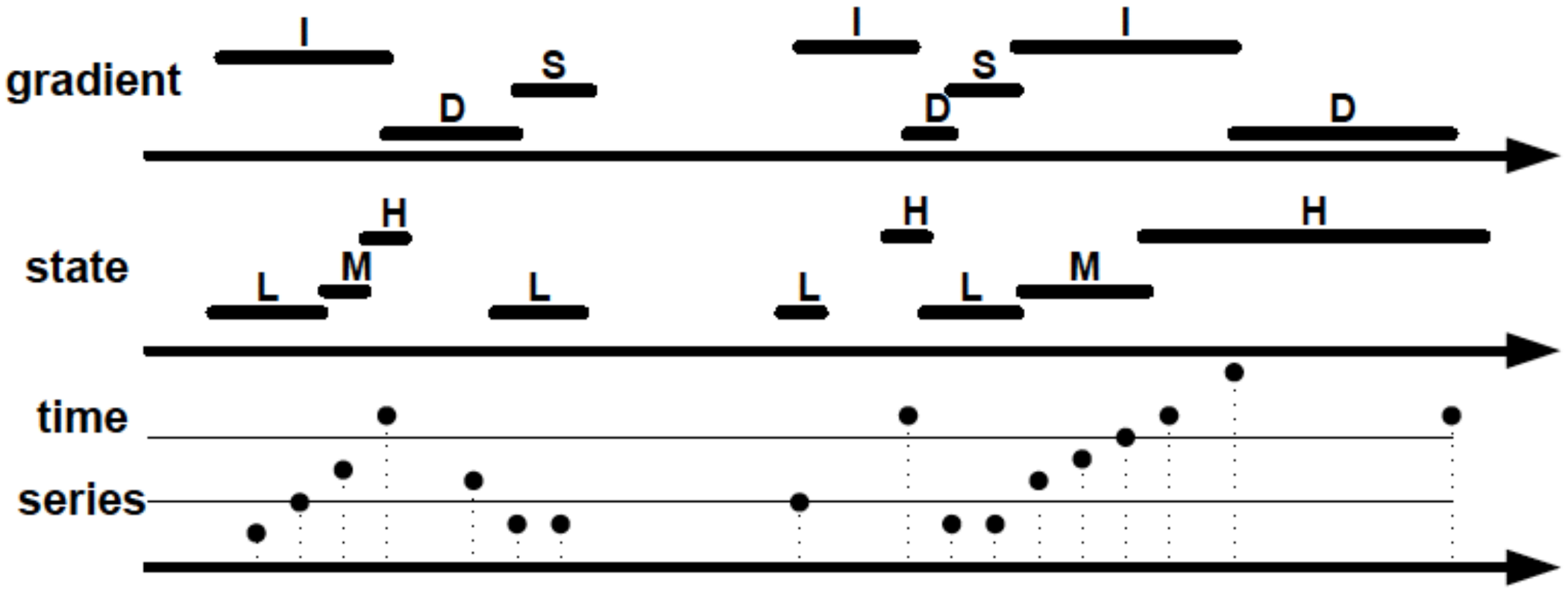
Collecting all concurrent event sets, detecting frequently co-occurred event subsets from them with pre-defined support threshold

Event Package	Events
Metabolic Panel	BUN
	Creatinine
	GFR estimated
	Glucose
	Potassium
	Sodium
Hepatic Panel	HCT
	HGB

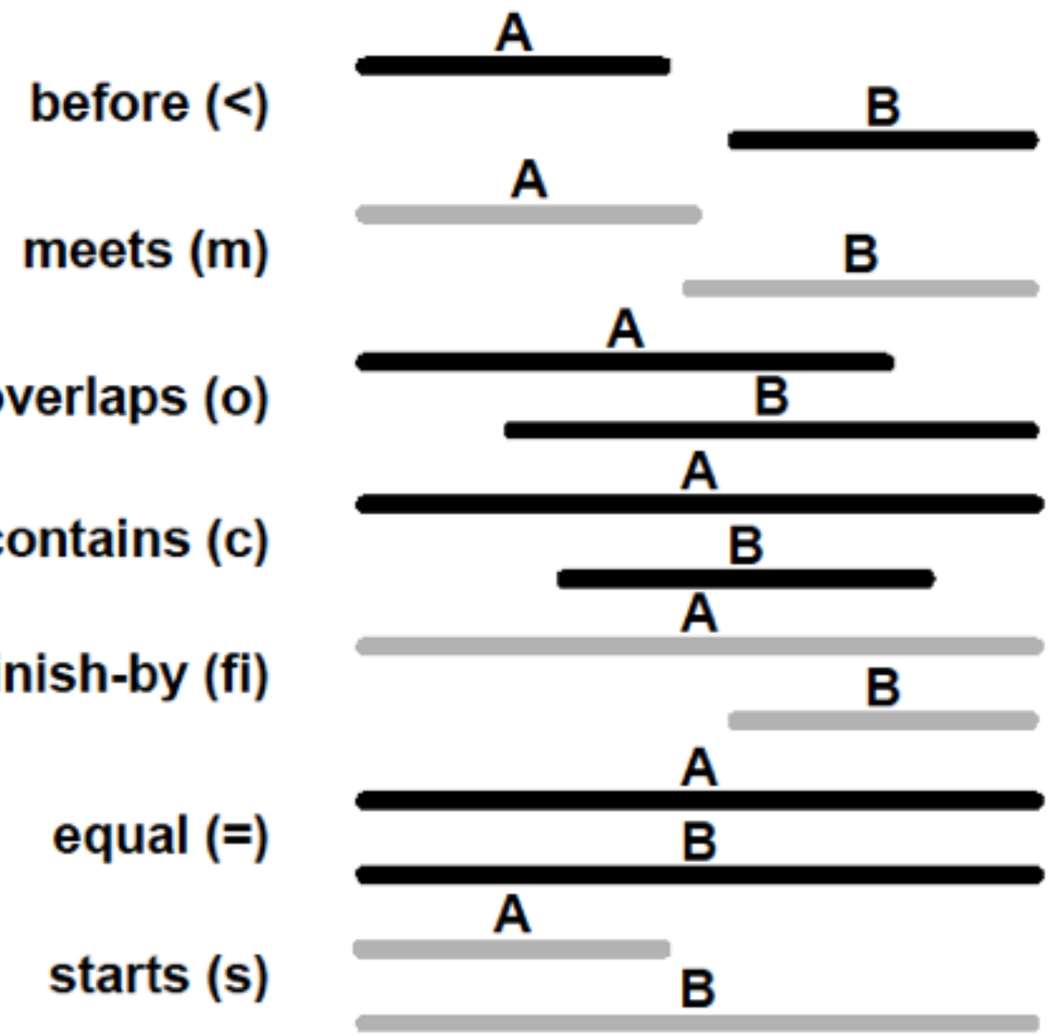
Event Package	Events
Diabetes Related Procedures	SUP-BLOOD GLUCOSE TST STRIP, 50
	SPRING-PWRD DEV FOR LANCET,EACH
	SUP-LANCETS,PER BOX
Diabetes Related Diagnosis	DIABETES MELLITUS
	DISORDERS OF LIPOID METABOLISM

R Agrawal, R Srikant. Fast algorithms for mining association rules. VLDB 1994.

Interval Representation

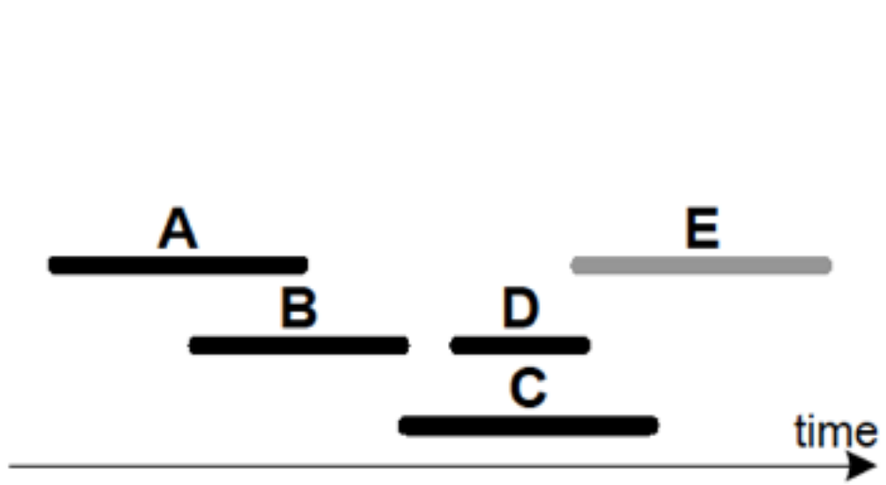


Allen's Event Relations



J. F. Allen. Maintaining knowledge about temporal intervals, Communications of the ACM, 26(11): 832-843, 1983

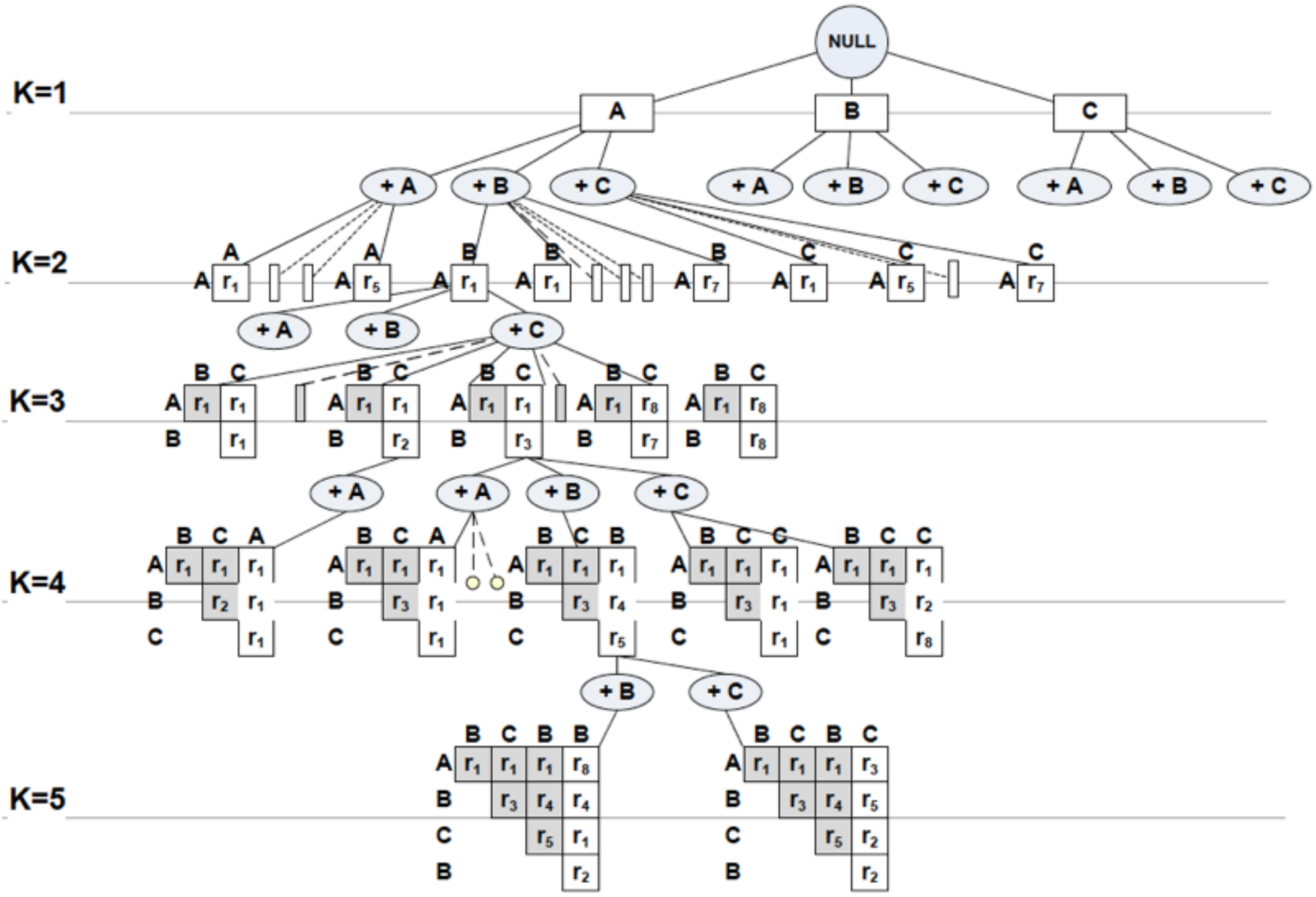
Time Interval Related Pattern



	B	C	D	E
A	o	<	<	<
B	m	<	<	
C	c	o		
D	m			

Moskovitch, Robert, and Yuval Shahar. "Medical temporal-knowledge discovery via temporal abstraction." In AMIA Annual Symposium Proceedings, vol. 2009, p. 452. American Medical Informatics Association, 2009.

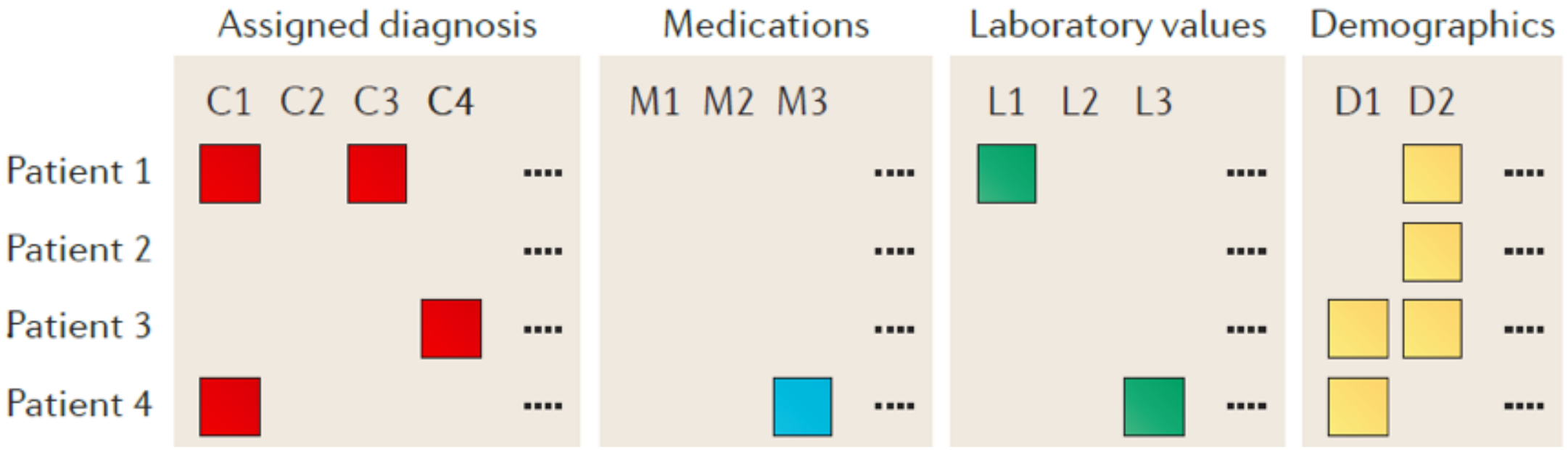
KarmaLego Enumeration Tree



Roadmap

- Background
- Electronic Health Records
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Longitudinal Information is Important



PID	DAY_ID	CLINICAL_EVENT	OP_DATE	ICD9_LONGNAME
000000	74053	305.1	74726	Tobacco Use Disorder
000000	74053	496	74726	Chronic Airway Obstruction, Not Elsewhere Classified
000000	74053	733	74726	Osteoporosis, Unspecified
000000	74053	724.2	74726	Lumbago
000000	74091	733	74726	Osteoporosis, Unspecified
000000	74148	733	74726	Osteoporosis, Unspecified
000000	74148	782.3	74726	Edema
000000	74148	780.79	74726	Other Malaise And Fatigue

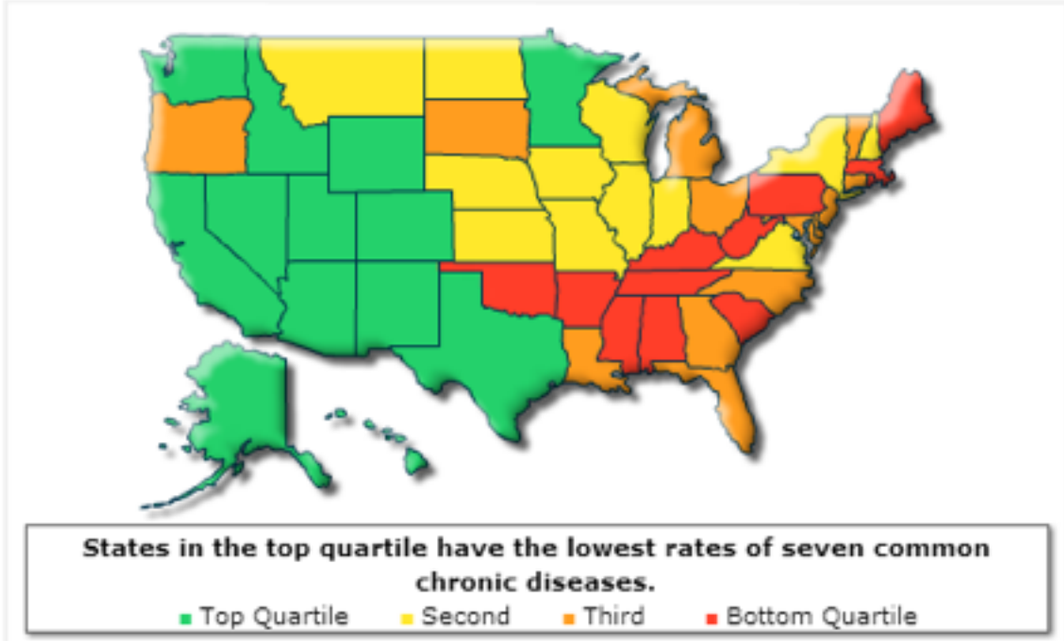
Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 May 2;13(6):395-405.

Chronic Disease

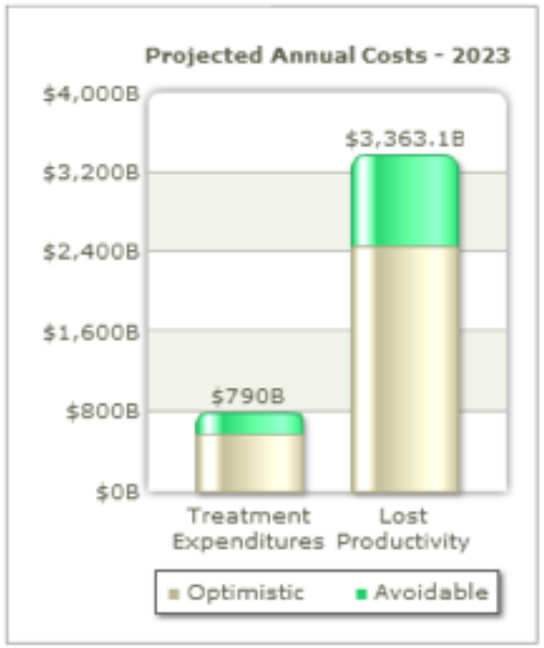
Reported Cases of Common Chronic Diseases 2003	
Millions	(As percent of population)
Cancers:	**10.6 (3.6%)**
Diabetes:	**13.7 (4.7%)**
Heart Disease:	**19.1 (6.6%)**
Hypertension:	**36.8 (12.6%)**
Stroke:	**2.4 (0.8%)**
Mental Disorders:	**30.3 (10.4%)**
Pulmonary Conditions:	**49.2 (16.9%)**
Total Reported Cases:	**162.2 (55.8%)**

United States Economic Impact 2003	
(Annual Costs in billions)	
Treatment Expenditures:	**$277.0B**
Lost Productivity:	**$1,046.7B**
Total Costs:	**$1,323.7B**

Milken Institute Chronic Disease Index



	Projected Annual Costs 2023		
	Current Course	Alternative Future	Costs Avoided
Treatment Expenditures	$790.1B	$572.4B	**$217.6B (27.6%)**
Lost Productivity	$3,363.0B	$2,458.0B	**$905.1B (26.9%)**
Total	$4,130.0B	$2,996.7B	**$1,133.3B (27.4%)**



Real GDP in 2050	
(In billions, 2003 dollars)	
Current Course:	**$32,229.2B**
Alternative Future:	**$37,897.5B**
Potential Gain in GDP:	**$5,668.4B (17.6%)**

http://www.chronicdiseaseimpact.com/

Disease Progression Modeling

- Chronic disease is a global burden
  - Hundreds of millions of people
  - Trillions of dollars spent
  - Loss in life expectancy
  - Loss in quality of life
- Chronic disease management
  - Stabilize disease progression
  - Avoid exacerbation
  - Reduce symptoms and comorbidities
- Disease progression modeling is instrumental
  - Comprehensive characterization of the progression stages
  - Identify the progression trajectory of individual patients
  - Provide decision support for early intervention

- Chronic Obstructive Pulmonary Disease (COPD)
  - Impacts low-income population
  - Key risk factors: smoking and air pollution
  - Causes systematic illness

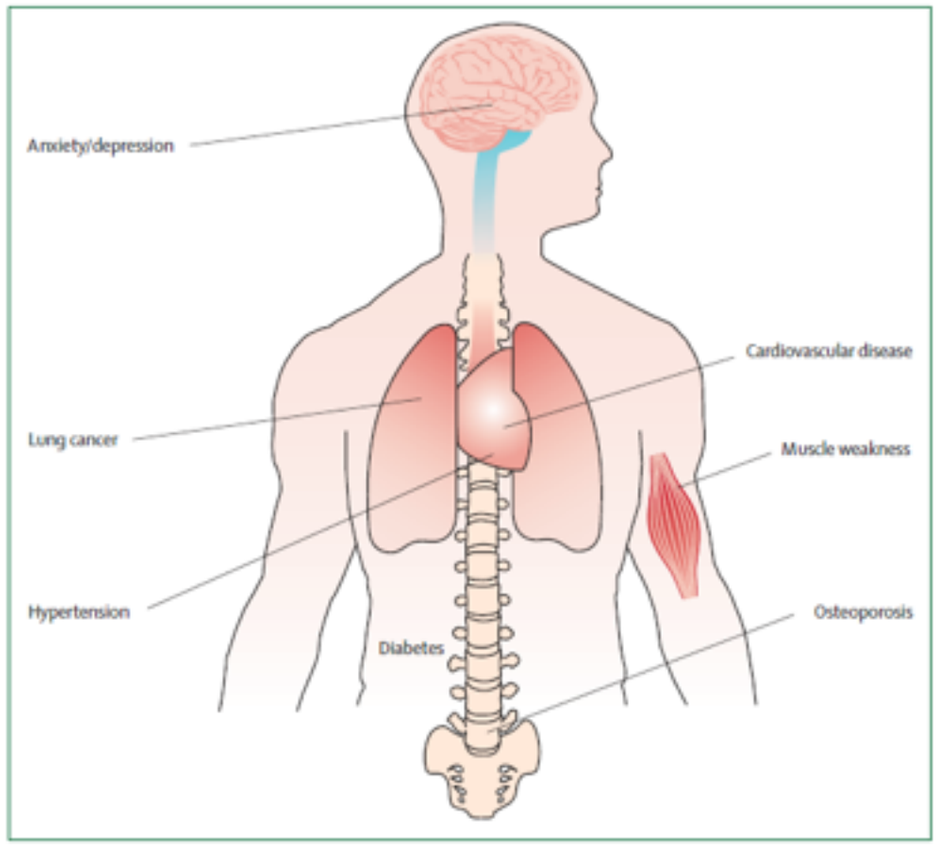


Figure 4: Comorbidities of chronic obstructive pulmonary disease

Chronic obstructive pulmonary disease. The Lancet, Volume 379, Issue 9823, Pages 1341 - 1351, 7 April 2012

Challenges on Disease Progression Modeling

- Multiple covariates
- Progression heterogeneity
  - Disease subtypes, e.g. asthma vs. non-asthma
- Missing data
  - Doctors only document the relevant clinical context
- Incomplete records
  - Use censored records to capture lengthy progression path
- Irregular visits
  - Continuous-time model is needed
- Limited supervision
  - No ground truth regarding the current stage of progression

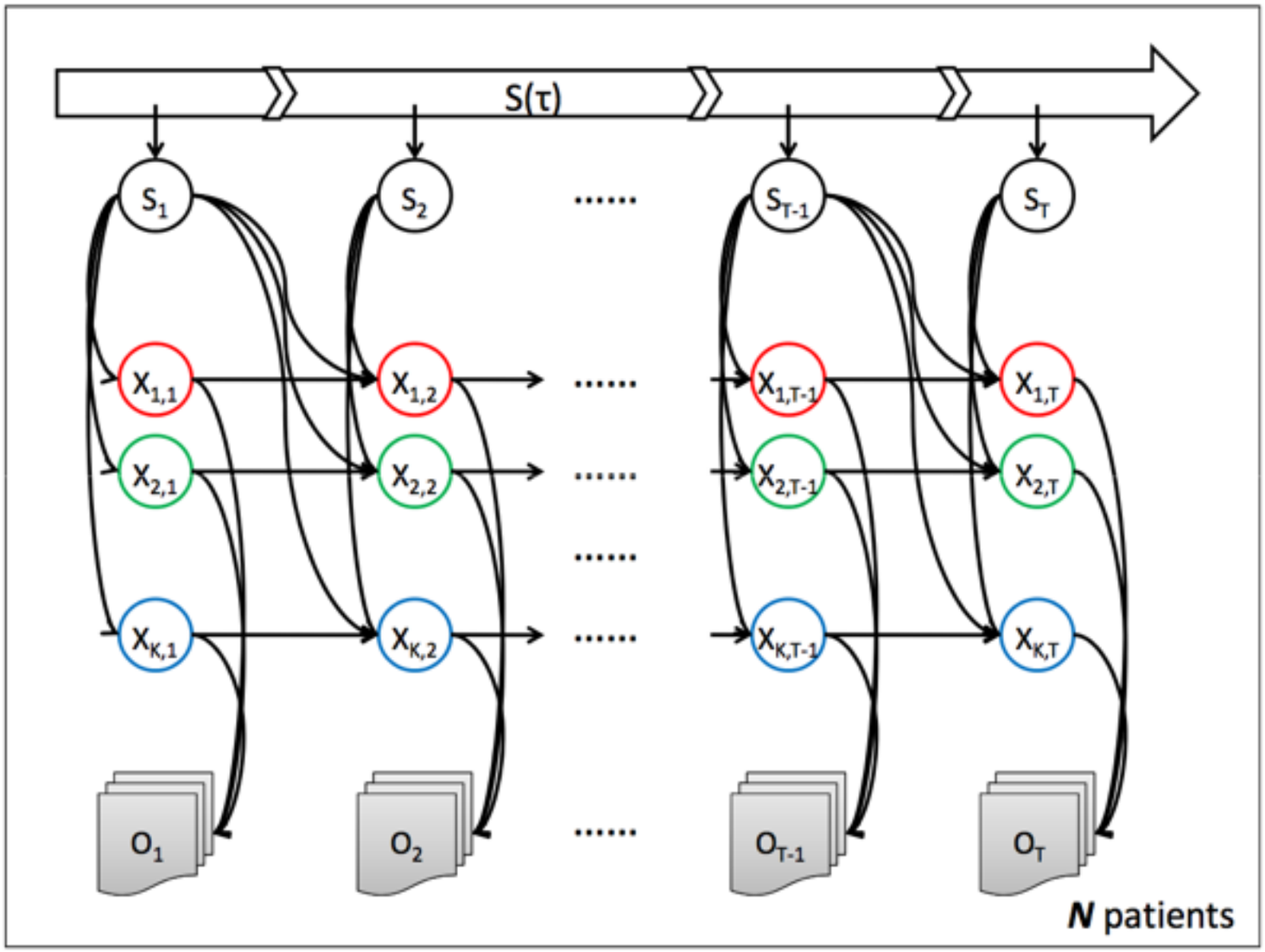
Our Proposed Model

Markov Jump Process

Progression States

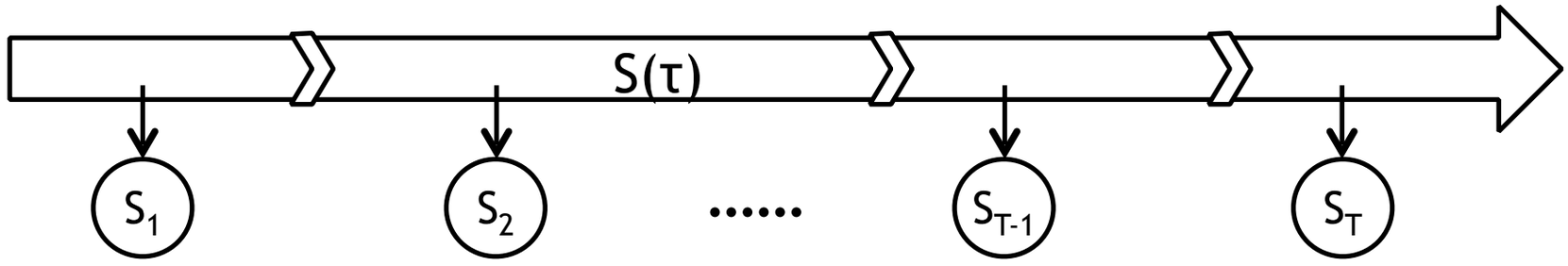
K Comorbidity Groups in Markov Chain

Observations



Xiang Wang, David Sontag, Fei Wang. Unsupervised Learning of Disease Progression Models. KDD 2014.

Markov Jump Process



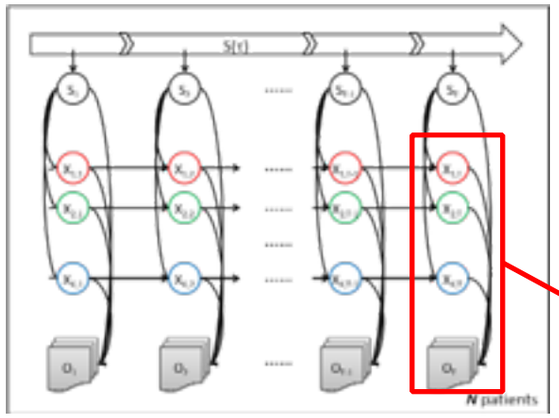
- A continuous-time Markov process with irregular discrete-time observations
- The transition probability is defined by a intensity matrix and the time interval:

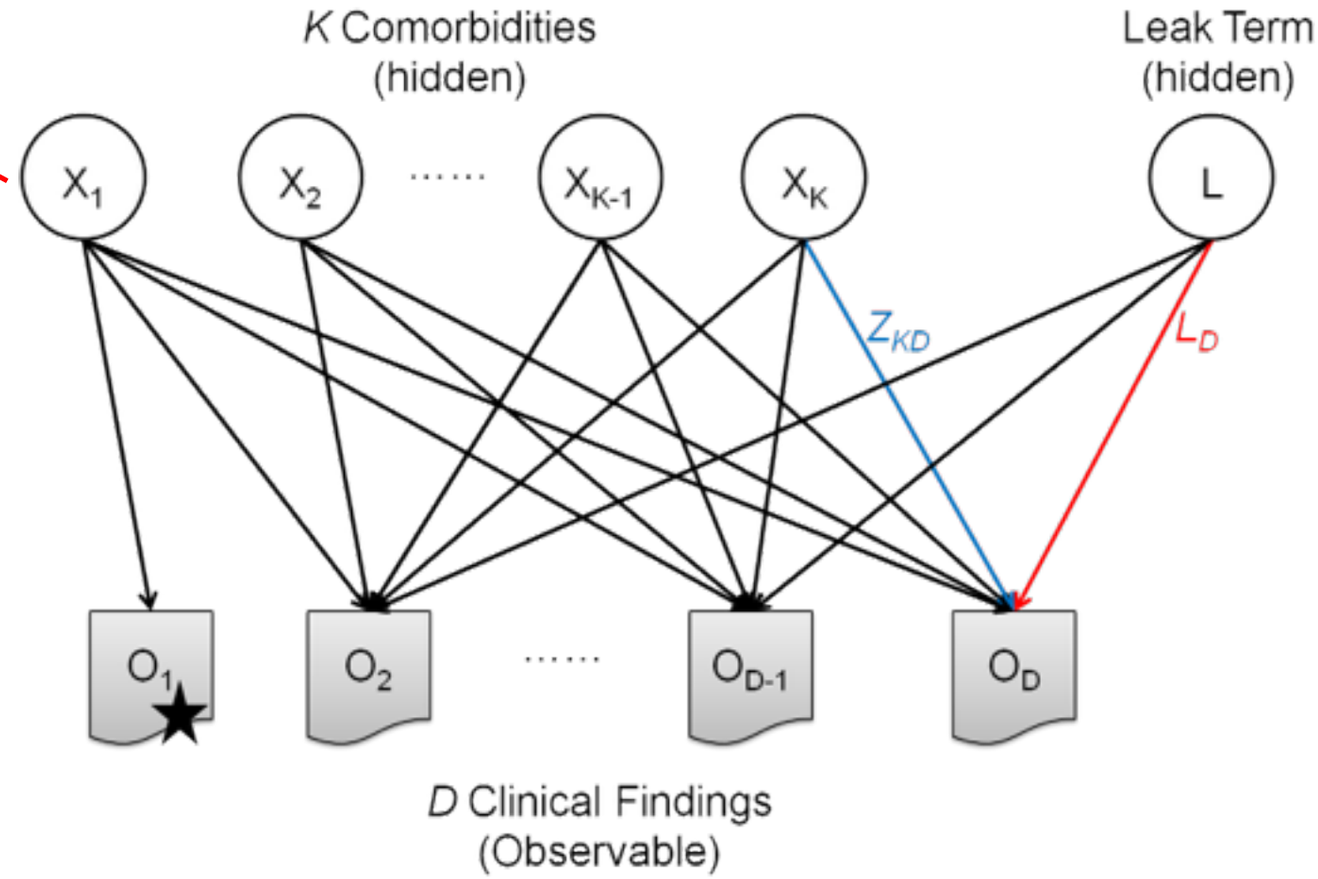
$$A_{ij}(\Delta) \triangleq P(S_t = j | S_{t-1} = i, \tau_t - \tau_{t-1} = \Delta; Q)$$
$$= \operatorname{expm}(\Delta Q)_{ij},$$

http://en.wikipedia.org/wiki/Markov_process

The Noisy-Or Network

Also known as QMR-DT network:
Quick Medical Reference, Decision Theoretic





M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann and G.F. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base I: the probabilistic model and inference algorithms, Methods of Information in Medicine, vol. 30, pp. 241 -- 255. 1991.

Anchored Noisy-Or Network

Use anchors findings to enable injection of domain expertise

An anchor is a medical finding that signifies the presence of a specific comorbidity

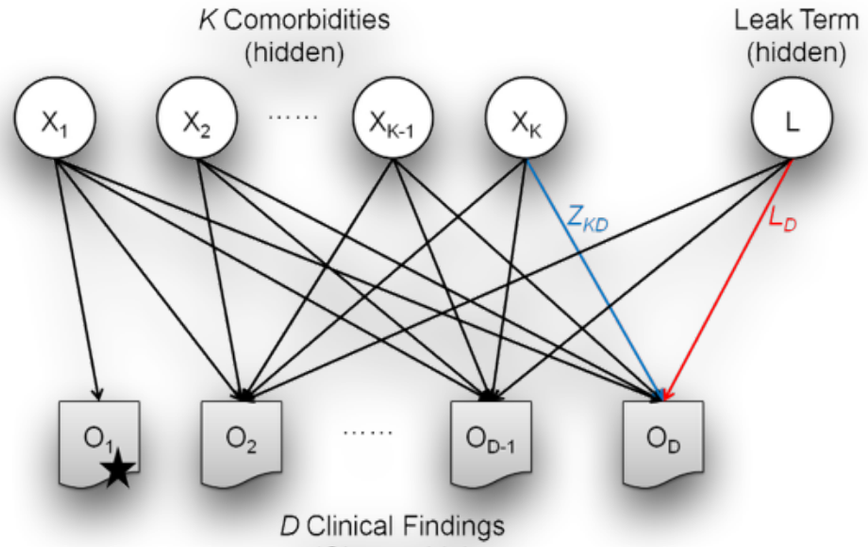


Table 2: COPD comorbidities, associated medical conditions, and anchor ICD-9 diagnosis codes used.

Comorbidity	Representative Conditions (Anchor ICD-9 Codes)
COPD	Chronic Bronchitis (491), Emphysema (492, 518), Chronic Airway Obstruction (496)
Asthma	Asthma (493)
Cardiovascular	Hypertension (401), Congestive Heart Failure (428), Arrhythmia (427), Ischemic Heart Disease (414)
Lung Infection	Pneumonia (481, 485, 486)
Lung Cancer	Malignant Neoplasm of Upper/Lower Lobe, Bronchus or Lung (162)
Diabetes	Diabetes with Different Types and Complications (250)
Musculoskeletal	Spinal Disorders (724), Soft Tissue Disorders (729), Osteoporosis (733)
Kidney	Acute Kidney Failure (584), Chronic Kidney Disease (585), Renal Failure (586)
Psychological	Anxiety (300), Depression (296, 311)
Obesity	Morbid Obesity (278)

Formalization & Implementation

Variables

$$\begin{aligned} \text{Disease States (hidden):} \quad & S_{n,t} \in \{1, \ldots, M\} \\ \text{Comorbidities (hidden):} \quad & X_{k,n,t} \in \{0, 1\} \\ \text{Clinical Findings (observable):} \quad & O_{d,n,t} \in \{0, 1\} \end{aligned}$$

Initial state probability

$$\pi = Pr(S_0)$$

State transition probability

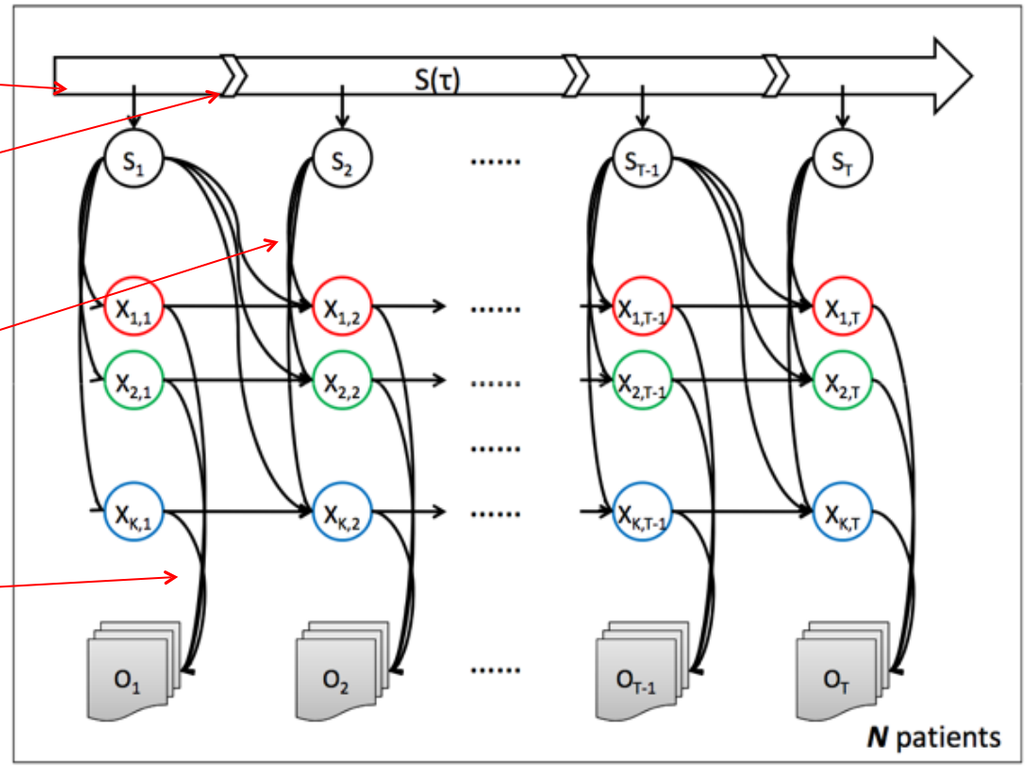
$$A(\Delta) = Pr(S_t | S_{t-1}, \tau_t - \tau_{t-1} = \Delta, Q) = \text{expm}(\Delta Q)$$

Comorbidity Onset Probability

$$B = Pr(X_t | X_{t-1}, S_t, S_{t-1}), \;\; \rho = Pr(X_0 | S_0)$$

Noisy-or Network

$$Pr(O_{d,n,t} = 1 | S, X) = 1 - (1 - L_d) \prod_k (1 - X_{k,n,t} Z_{k,d})$$



COPD Cohort Information

- Identify patients with COPD
  - At least one COPD-related diagnosis code
  - At least one COPD-related drug
  - OP_DATE is earliest occurrence of the COPD-related diagnosis

- Removed patients with too few records
- Removed ICD-9 codes that only occurred to a small number of patients
- Combined visits into 3-month time window
- 3,705 patients, 264 ICD-9 codes
- 34,976 visits, 189,815 observations

Comorbidity Groups

Asthma

Asthma*
Allergic Rhinitis
Cough
Acute Bronchitis
Acute Upper Respiratory Infections

Cardiovascular

Benign Essential Hypertension*
Unspecified Essential Hypertension*
Atrial Fibrillation*
Congestive Heart Failure*
Hyperlipidemia

Diabetes

Type II Diabetes without Complication*
Type II Diabetes without Complication, Uncontrolled*
Hyperlipidemia
Pure Hypercholesterolemia
Type II Diabetes with Renal Manifestations*

Kidney

Chronic Kidney Disease, Moderate* .
Anemia
Chronic Kidney Disease, Unspecified*
Urinary Tract Infection
Chronic Kidney Disease, Severe

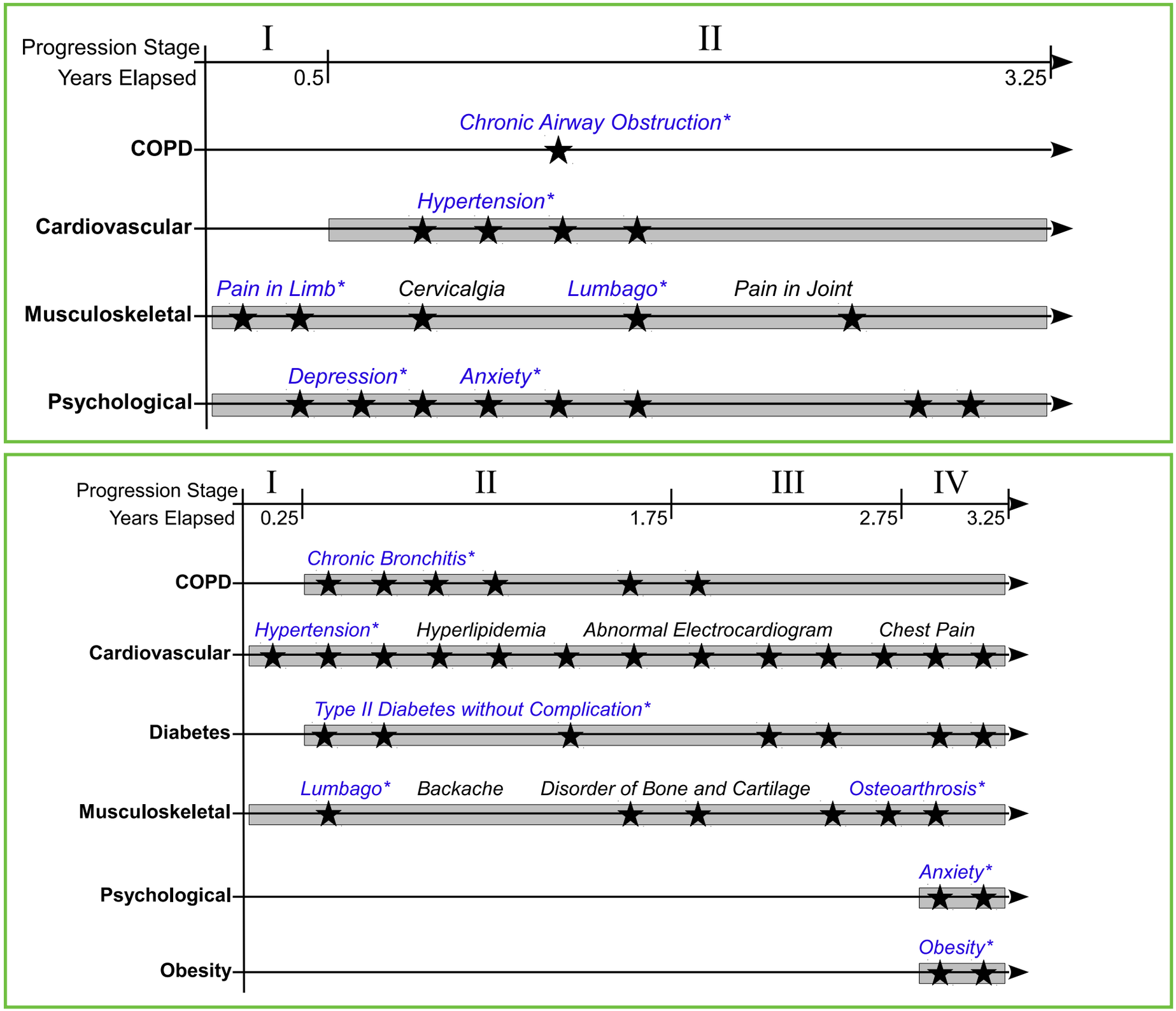
Lung Cancer

Cancer of Bronchus and Lung, Unspecified*
Other Diseases of Lung
Cancer of Other Parts of Bronchus or Lung*
Cancer of Upper Lobe, Bronchus or Lung*
Swelling, Mass, or Lump in Chest

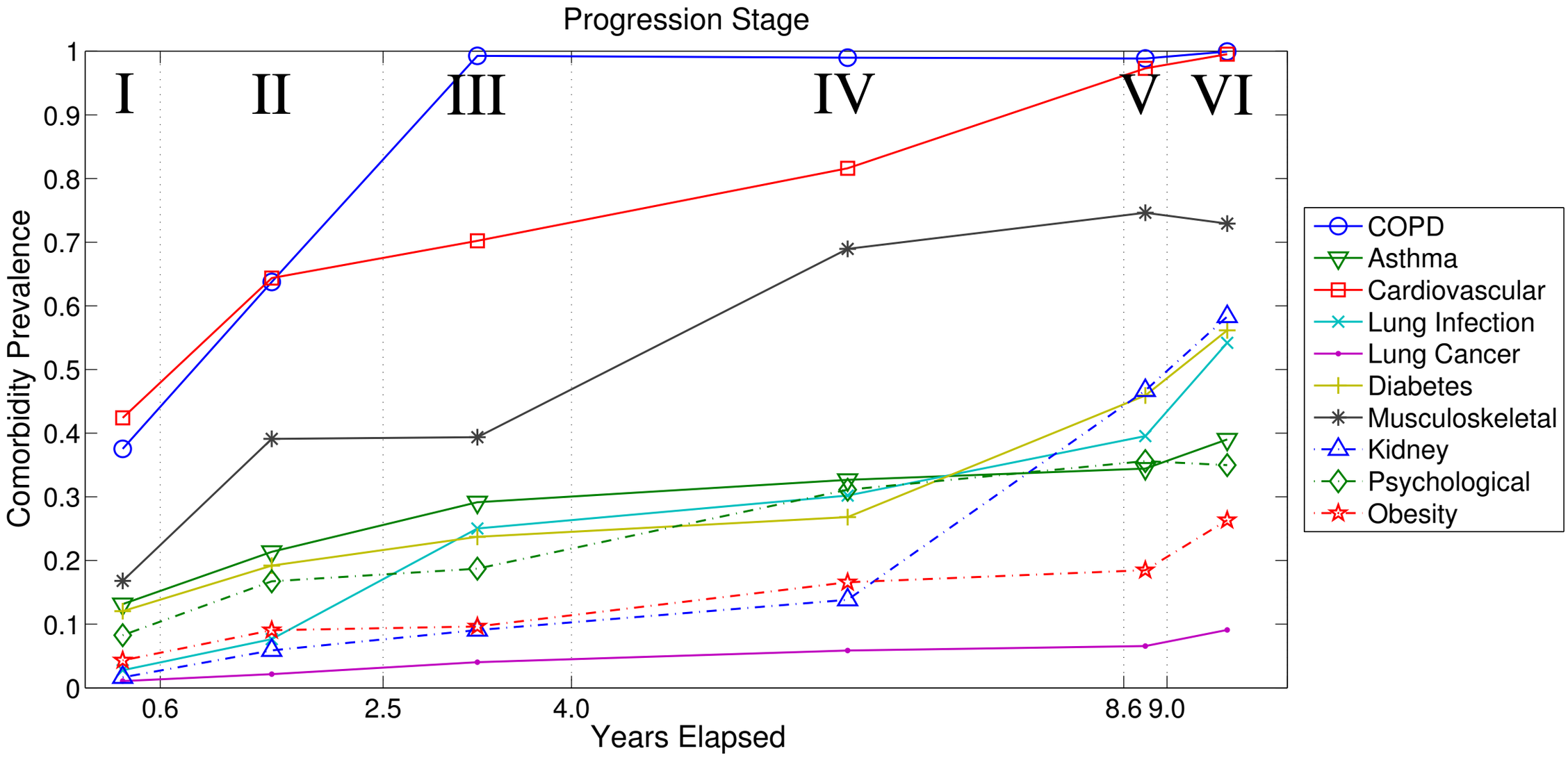
Obesity

Obesity, Unspecified*
Morbid Obesity*
Pure Hypercholesterolemia
Edema
Sleep Apnea

Inference of Individual Progression Path



Progression Rate



Roadmap

- Background
- Electronic Health Records
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Future Directions

- Integration of heterogeneous data
  - Medical: EHR, Drug
  - Chemistry: Drug
  - Biological: Genotype
  - SocialPsycoBehavioral: Social media, Wearable device, Mobile …
  - Environmental

Activities

- SDM 2015 The 4th Workshop on Data Mining for Medicine and Healthcare.

- ICHI 2015: IEEE International Conference on Health Informatics. Oct. 2015. Dallas, Texas.

Acknowledgement

- IBM:
  - Shahram Ebadollahi
  - Jianying Hu
  - Xiang Wang
  - Ping Zhang
  - Robert Sorrentino
  - Nan Cao
  - Zhaonan Sun
  - Xinxin Zhu
  - Daby Sow
  - Harry Stavropolous
  - Kenney Ng
  - Adam Perer
- Georgia Tech: Jimeng Sun
- NYU: David Sontag
- Stanford: Nigam Shah
- ASU: Hanghang Tong
- UMich: Jieping Ye
- UPitts: Yu-Ru Lin
- Rutgers: Hui Xiong
- UC Davis: Ian Davidson
- Temple: Zoran Obradovic
- SUNY Buffalo: Marianthi Markatou
- Columbia: Robert Moskovitch
- UTHealth: Hua Xu
- University of Maribor: Gregor Stiglic

Thank You

email: fei_wang@uconn.edu

Slides available through my website
http://www.engr.uconn.edu/~fwang/tutorials/AAAI15_tutorial.pdf